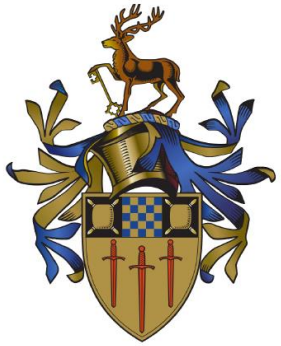
Towards Practicality of Sketch-Based Visual Understanding

PhD candidate: Ayan Kumar Bhunia

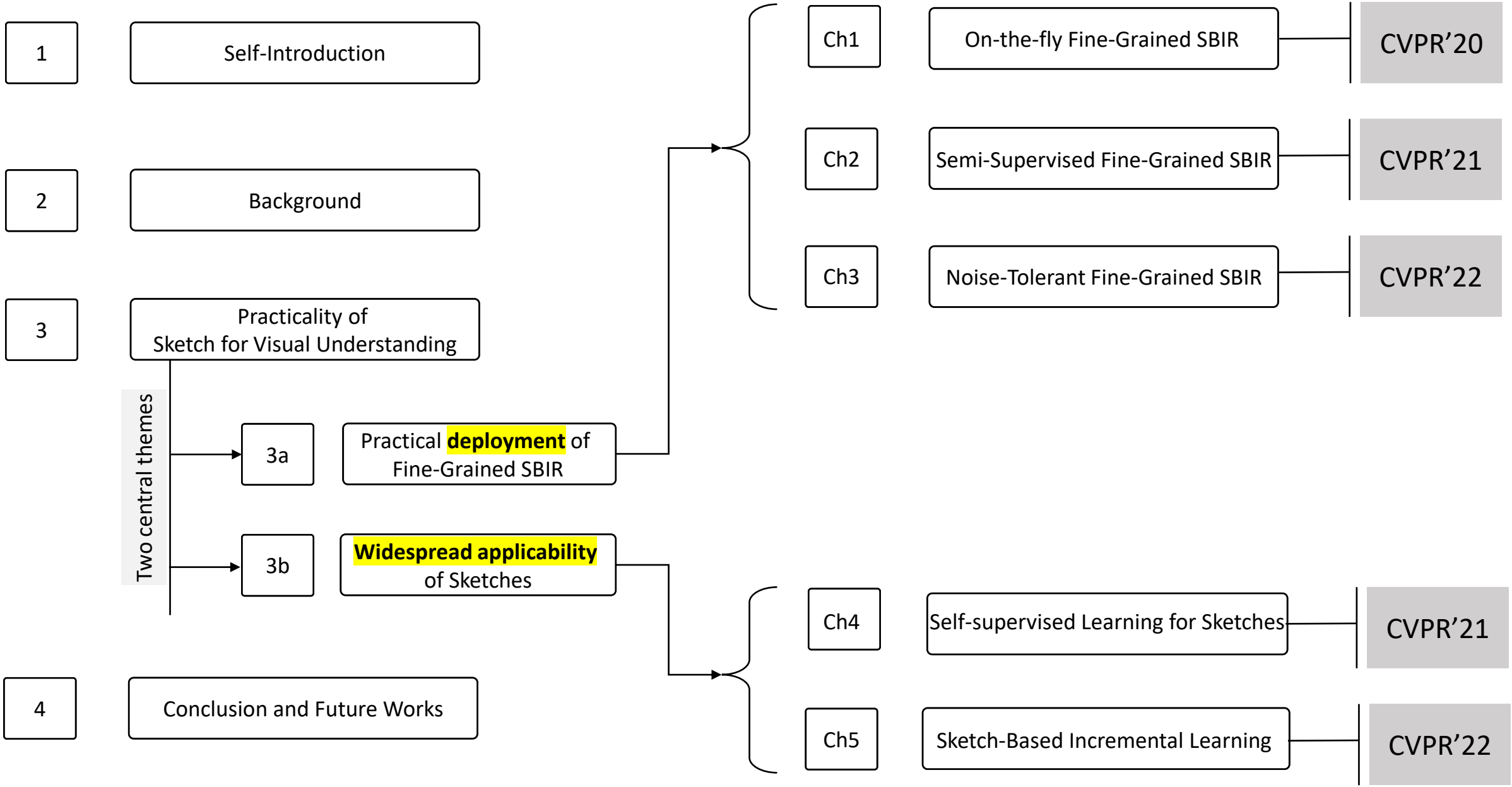
Supervisor: Prof. Yi-Zhe Song



Centre for Vision, Speech and Signal Processing

Faculty of Engineering and Physical Processing

University of Surrey


1
Self-Introduction
2
Background
3
Practicality of
Sketch for Visual Understanding
Two central themes
3a
Practical deployment of
Fine-Grained SBIR
3b
Widespread applicability
of Sketches
4
Conclusion and Future Works
Ch1
On-the-fly Fine-Grained SBIR
CVPR'20
Ch2
Semi-Supervised Fine-Grained SBIR
CVPR'21
Ch3
Noise-Tolerant Fine-Grained SBIR
CVPR'22
Ch4
Self-supervised Learning for Sketches
CVPR'21
Ch5
Sketch-Based Incremental Learning
CVPR'22

Self-Introduction

Kolkata, West Bengal, India

WEST BENGAL UNIVERSITY OF TECHNOLOGY
Utech
In Pursuit Of Knowledge And Excellence

United Kingdom

SketchX
UNIVERSITY OF SURREY

Publications contributed towards this thesis

i) **Ayan Kumar Bhunia**, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, Yi-Zhe Song."*Sketch Less for More: On-the-Fly Fine-Grained Sketch Based Image Retrieval*". IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**) 2020.

ii) **Ayan Kumar Bhunia**, Pinaki Nath Chowdhury, Aneeshan Sain, Yongxin Yang, Tao Xiang, Yi-Zhe Song. "*More Photos are All You Need: Semi-Supervised Learning for Fine-Grained Sketch Based Image Retrieval*". IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**) 2021.

iii) **Ayan Kumar Bhunia**, Pinaki Nath Chowdhury, Yongxin Yang, Timothy Hospedales, Tao Xiang, Yi-Zhe Song. "*Vectorization and Rasterization: Self-Supervised Learning for Sketch and Handwriting*". IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**) 2021.

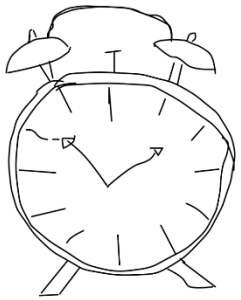
iv) **Ayan Kumar Bhunia**, Subhadeep Koley, Abdullah Faiz Ur Rahman Khilji, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, Yi- Zhe Song. "*Sketching without Worrying: Noise Tolerant Sketch-Based Image Retrieval*". IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**) 2022.

v) **Ayan Kumar Bhunia**, Viswanatha Reddy Gajjala, Subhadeep Koley, Rohit Kundu, Aneeshan Sain, Tao Xiang, Yi-Zhe Song. "*Doodle It Yourself: Class Incremental Learning by Drawing a Few Sketches*". IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**) 2022.

Background

What is Sketch?

- Dense color pixels (photo) *versus* sparse black and white line (sketch).
- Free-hand sketch is highly *abstract* and *subjective*.
- Sketch *carries personal style*, *subjective abstraction*, *human creativity*



Why Sketch for Visual Understanding?

- Sketch has been used to *conceptualise and depict visual objects*.
- Language of *sketch is universal* – to some *significant* extent. Compared to text, sketches are incredibly *intuitive* to humans.
- Sketch can model ***fine-grained details*** of a visual concept easily.

Background

Challenges for Sketches

- Time taken to draw a sketch
- Training data (sketch-photo pair) is limited
- Sketching is difficult
- Beyond supervised sketch representation learning
- Learning from very few sketch exemplars

❑ Solution

- **On-the-fly FG-SBIR**
- **Semi-supervised FG-SBIR**
- **Noise-Tolerant FG-SBIR**
- **Self-supervised learning on Sketches**
- **Sketch-based Few-shot Class Incremental Learning**

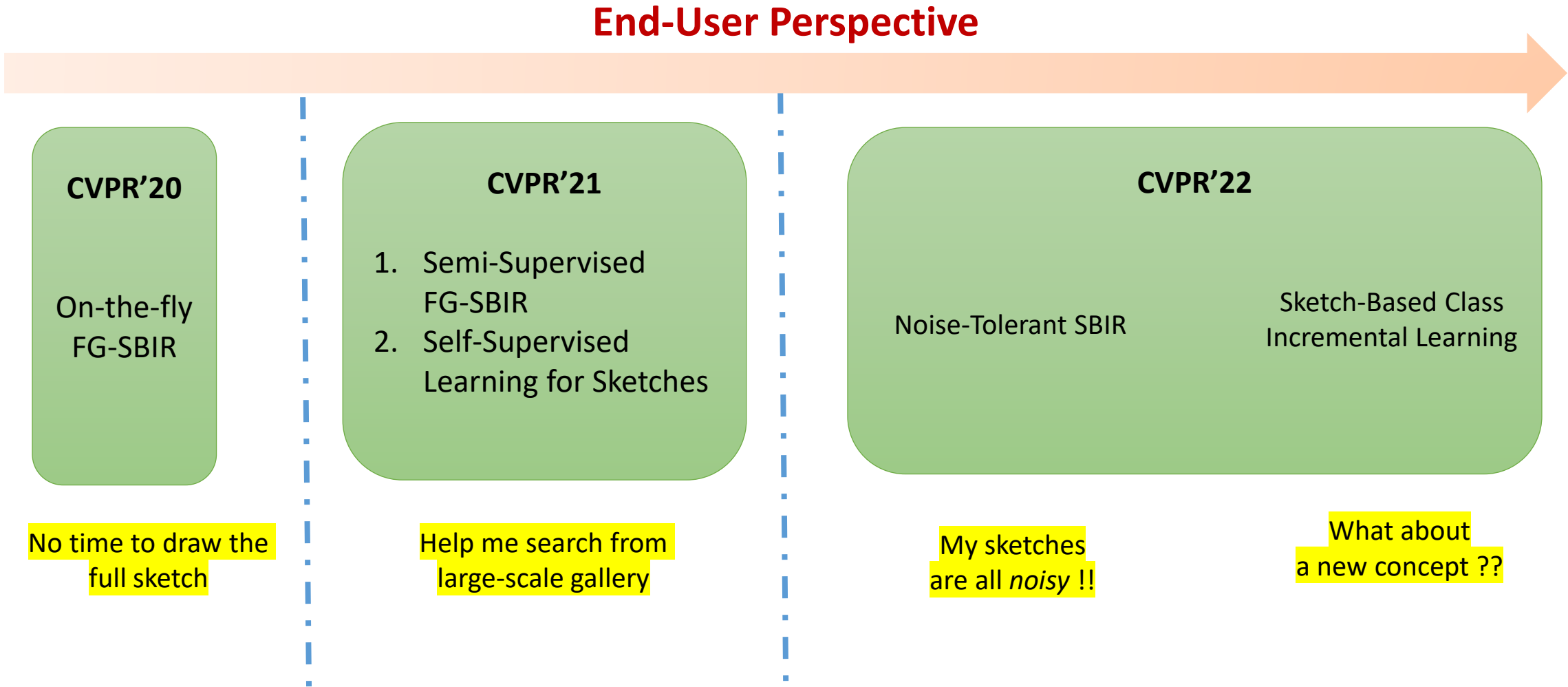
Background

Thesis Outline

- **Theme 1:** Practical deployment of FG-SBIR
  - **On-the-fly FG-SBIR [CVPR'20]**
    - Drawing a sketch *takes time*.
    - Many struggle to draw a complete/faithful sketch
  - **Semi-Supervised FG-SBIR [CVPR'21]**
    - Lack of sketch-photo pairs for FG-SBIR
  - **Noise-Tolerant FG-SBIR [CVPR'22]**
    - Irrelevant (noisy) strokes drawn by the user are detrimental.
- **Theme 2:** Widespread applicability of sketches for real-world applications
  - **Self-Supervised Learning for Sketches (Sketch2Vec) [CVPR'21]**
    - Sketch-specific *pre-text task* to alleviate the data annotation bottleneck
  - **Sketch-Based Class Incremental Learning (SBIL) [CVPR'22]**
    - *Update a 10-class classifier to (10+3)-class classifier* using *a few sketch exemplars* from novel classes.

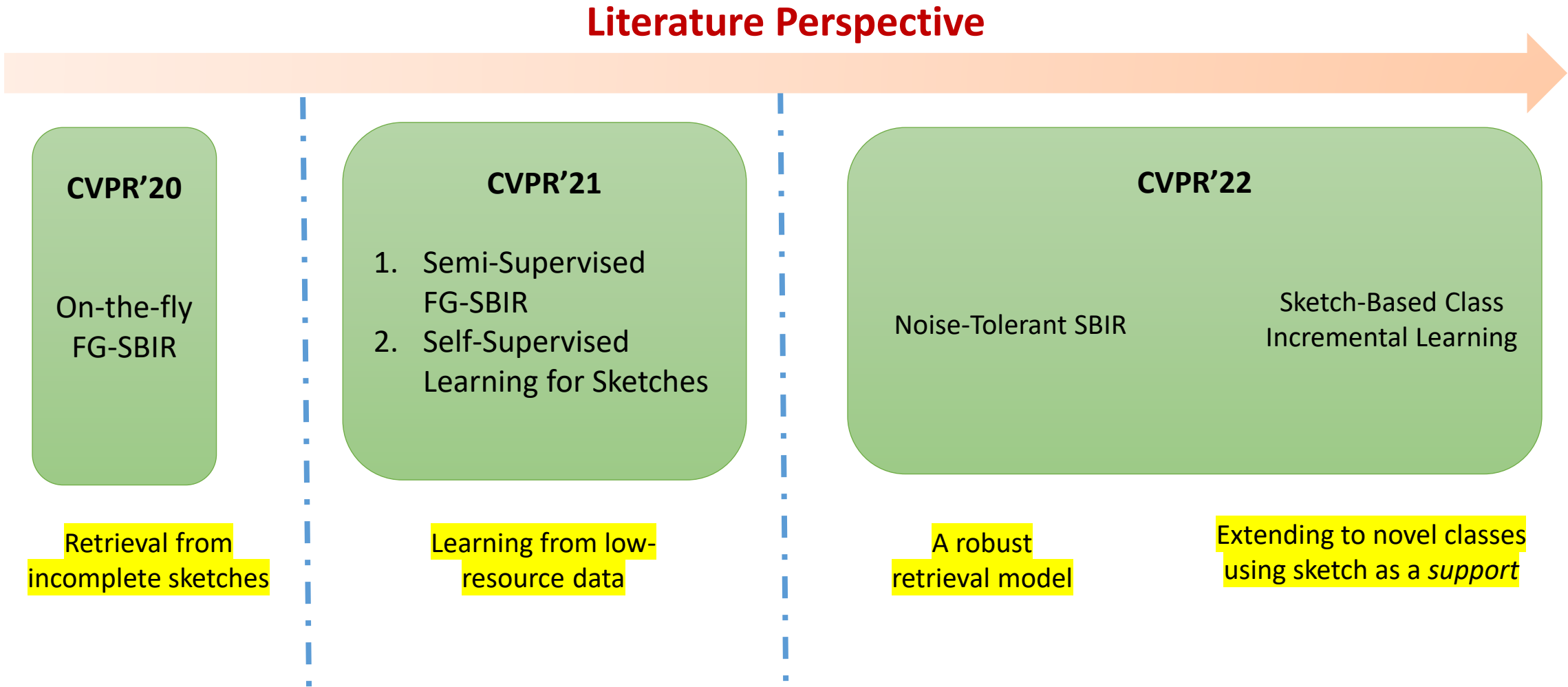
Background

How are they connected?



Background

How are they connected?



Background

How are they connected?

Sketch as a Modality for Vision

1. On-the-fly FG-SBIR
2. Semi-supervised FG-SBIR
3. Noise-Tolerant SBIR

4. Self-supervised Learning on sketches.

5. Sketch-Based Incremental Learning

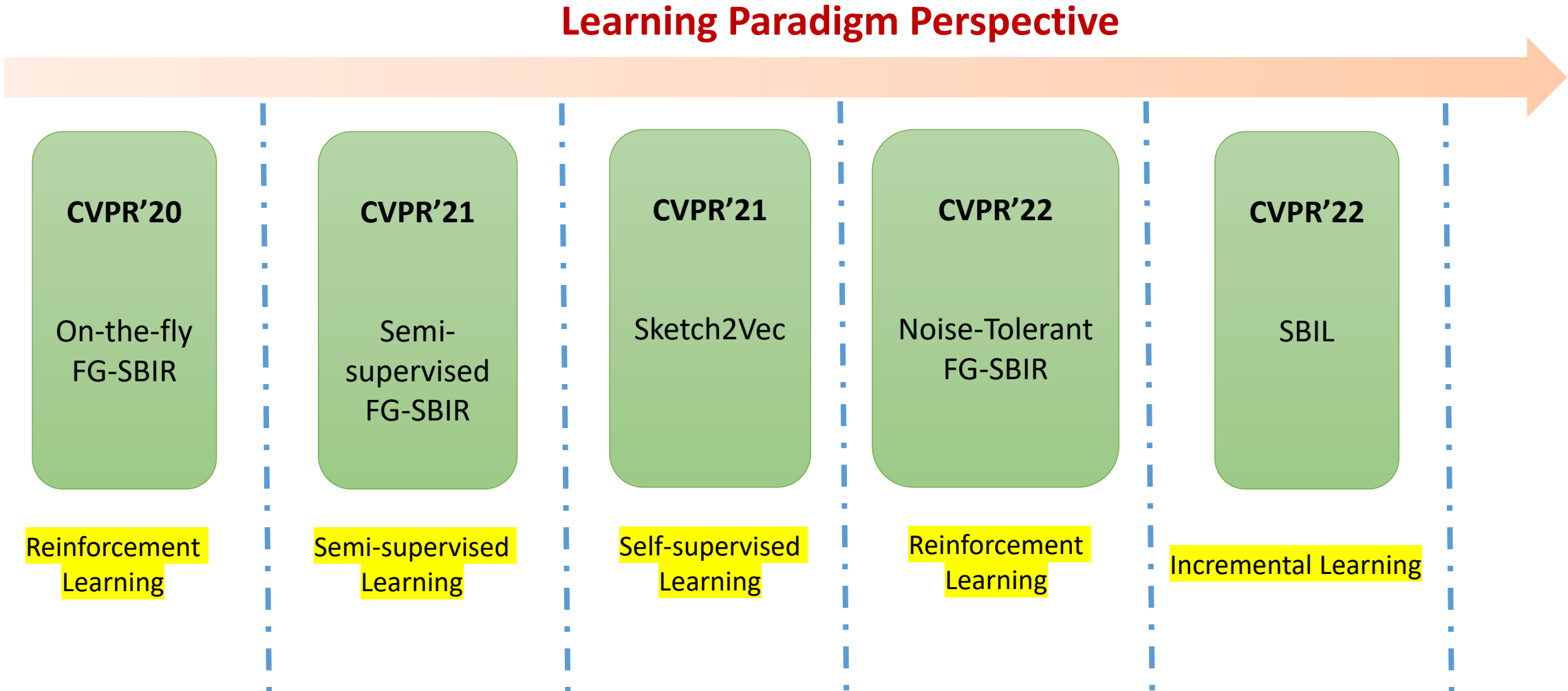
Sketch and Photo

Sketch Only

Sketch as a support

Background

How are they connected?



Background

How to leverage the fine-grained potential of sketch?

❑ **Fine-Grained Sketch-Based Image Retrieval**

- ❖ FG-SBIR aims at retrieving a particular photo instance given a user's query sketch.
- ❖ Focus is on *fine-grained SBIR* over *category-level SBIR*

Sketch | Gallery Images

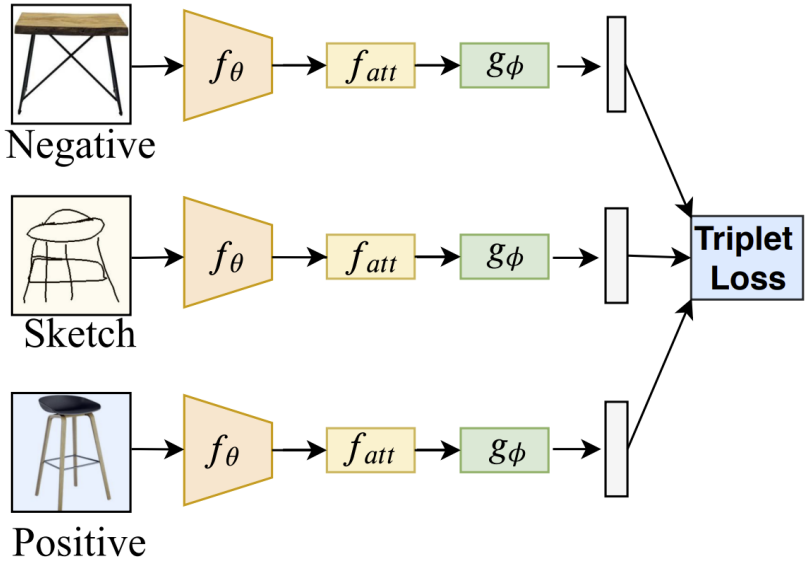
[R1] Deep Learning for Free-Hand Sketch: A Survey, IEEE TPAMI, 2022.

❑ Applications of Sketches in Computer Vision and Graphics Community:

- Sketch-to-RGB Image Generation
- Sketch-Based Image Editing
- Sketch-Based Image Retrieval
  - Category-level Retrieval
  - Fine-grained Retrieval ✓
- Sketch-Based 3D shape Modelling
- Sketch-Based 3D Shape retrieval
- Sketch-Based Object Localization
- Sketch for AR/VR
- Sketch and Visual Correspondence
- Sketch-Based Video Synthesis
- Sketch-Based Garment Design
- Sketch-Based Segmentation
- Sketch-based Manga Restoration and Inpainting
- Sketch for Medical Image Analysis
- Sketch for Representation Learning ✓
- Sketch for Incremental Learning ✓
- etc.

Background on FG-SBIR

Baseline FG-SBIR



$$L_{triplet} = \max\{0, \mu + \beta^+ - \beta^-\}$$

where, $\beta^+ = distance(sketch, positive\ photo)$
$\beta^+ = distance(sketch, negative\ photo)$

Dataset: QMUL-ShoeV2, ChairV2

Paired Annotations

On-the-Fly FG-SBIR

Sketch

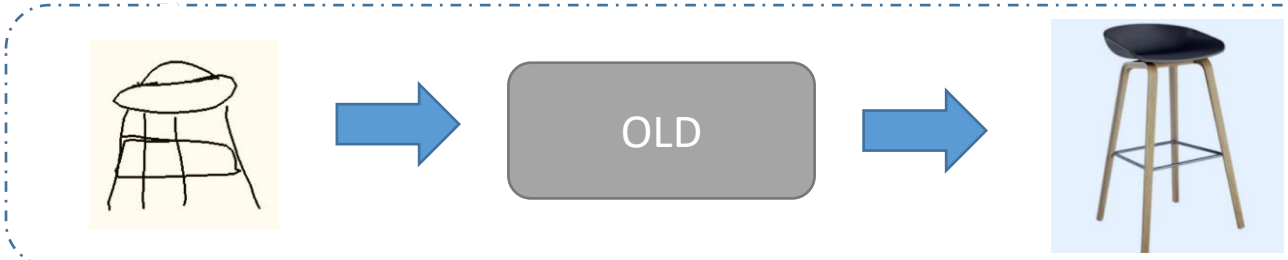
Gallery Images

Problem – "I can't sketch"

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

On-the-Fly FG-SBIR

Old Setup: sketch first, *then* retrieve



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

Less is more!

On-the-Fly FG-SBIR

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!

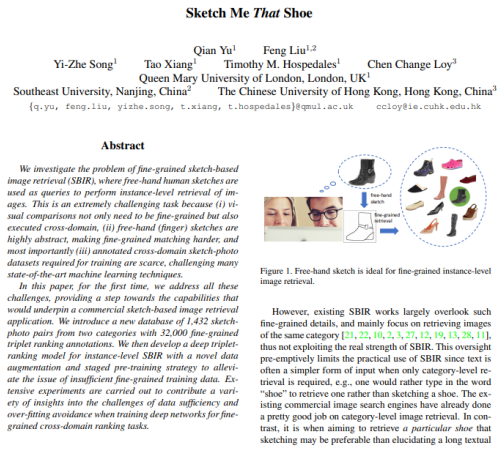
On-the-Fly FG-SBIR

Why Challenging?

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

Why not Triplet-Loss for on-the-fly FG-SBIR?

- Noisy gradient for partial or initial sketches
- No early retrieval mechanism.
- Does not consider the temporal nature of sketch.



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

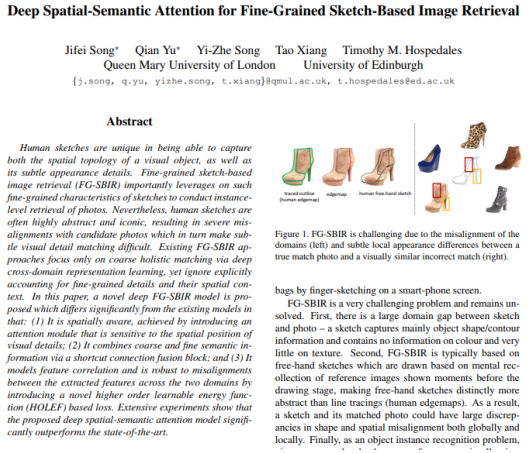
We investigate the problem of fine-grained sketch-based image retrieval (SBIR), where free-hand human sketches are used as queries to perform instance-level retrieval of images. This is an extremely challenging task because (i) visual comparisons not only need to be fine-grained but also executed cross-domain, (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

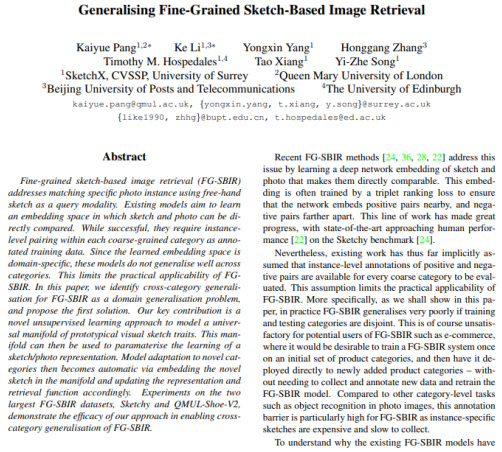
Human sketches are unique in being able to capture both the spatial topology of a visual object, as well as its subtle appearance details. Fine-grained sketch-based image retrieval (FG-SBIR) importantly leverages on such fine-grained characteristics of sketches to conduct instance-level retrieval of 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 parameterise 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

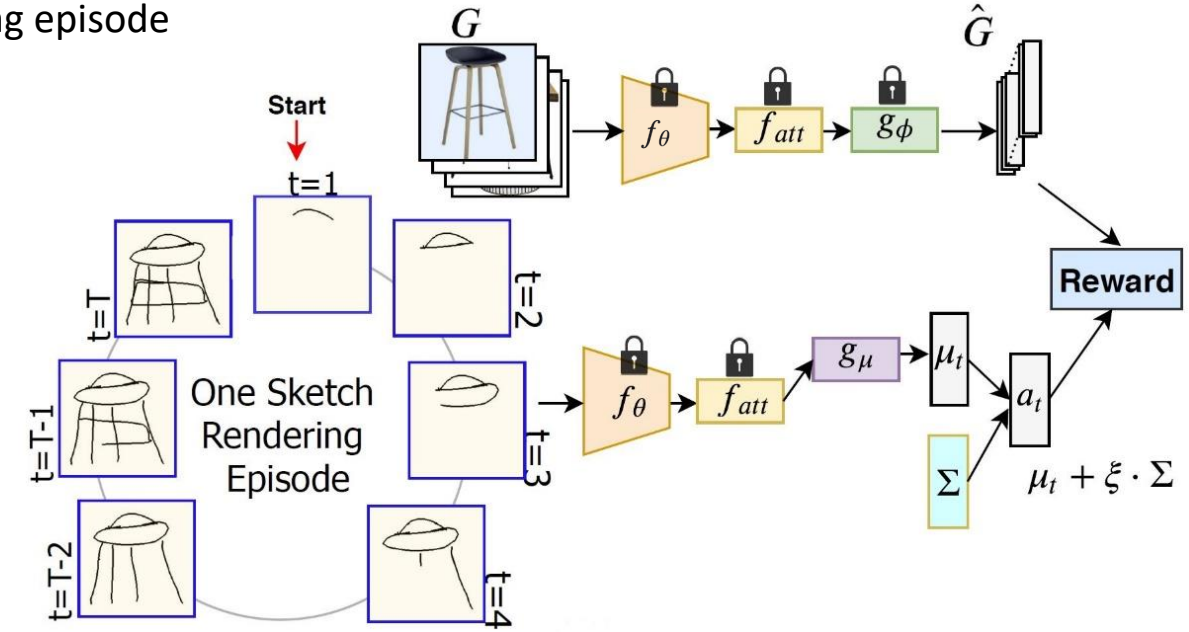
[C]

[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

On-the-Fly FG-SBIR

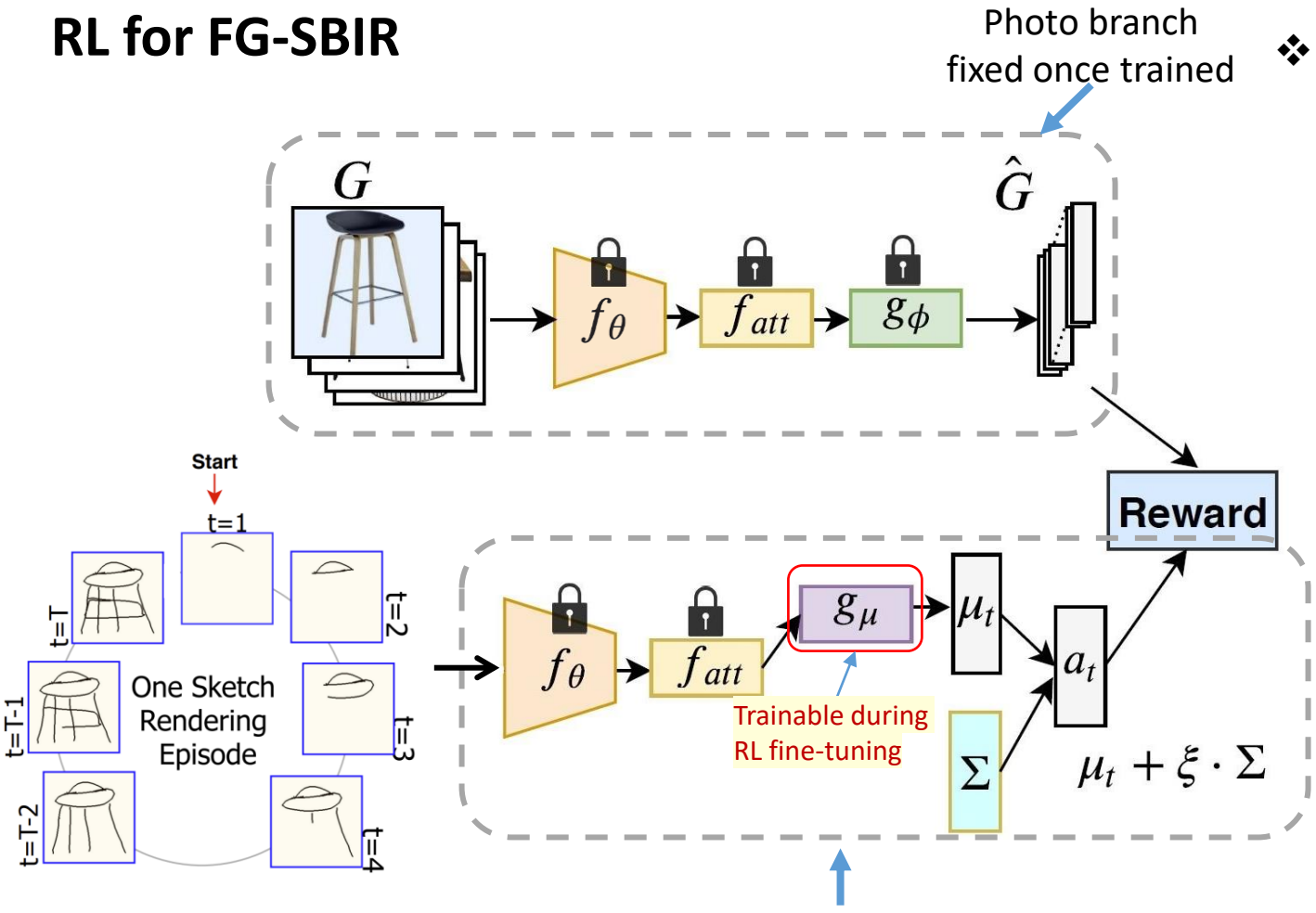
Contributions

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



On-the-Fly FG-SBIR

RL for FG-SBIR



❖ RL Design

- ❑ Sketch branch acts as the RL agent
- ❑ Based on Stochastic Continuous Gaussian Policy
- ❑ Action generation with multivariate normal distribution
- ❑ a_t is the feature representation at t-th sketch rendering step

$$a_t = \mu_t + \varepsilon * \Sigma$$

where, $\varepsilon = N(0, I)$

On-the-Fly FG-SBIR

Reward Design

Total Reward

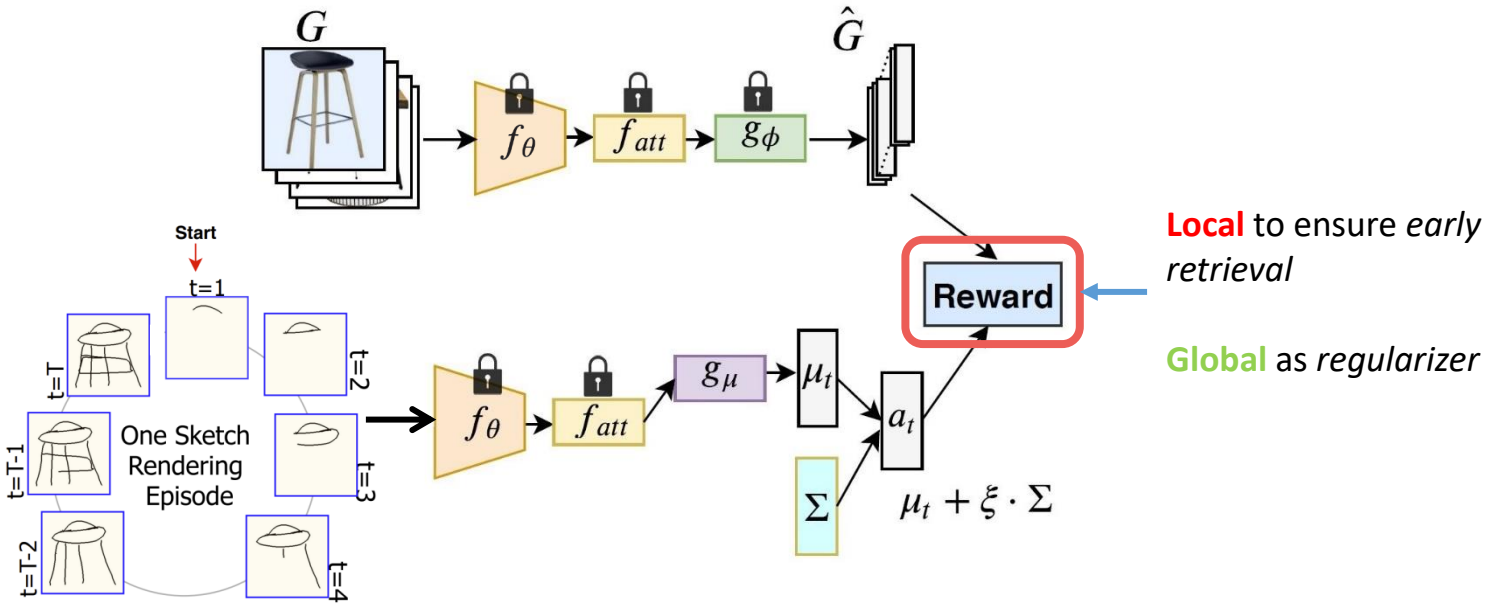
$$R_t = \gamma_1 R_t^{Local} + \gamma_2 R_t^{Global}$$

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

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*

- **Training Paradigm**
  - Local and Global reward.
  - Local reward aims at early retrieval.
    - maximise the inverse rank.
  - Global reward as a *regularizer*
    - resistance against noisy strokes
  - Proximal Policy Optimization instead of basic Policy Gradient.

On-the-Fly FG-SBIR

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

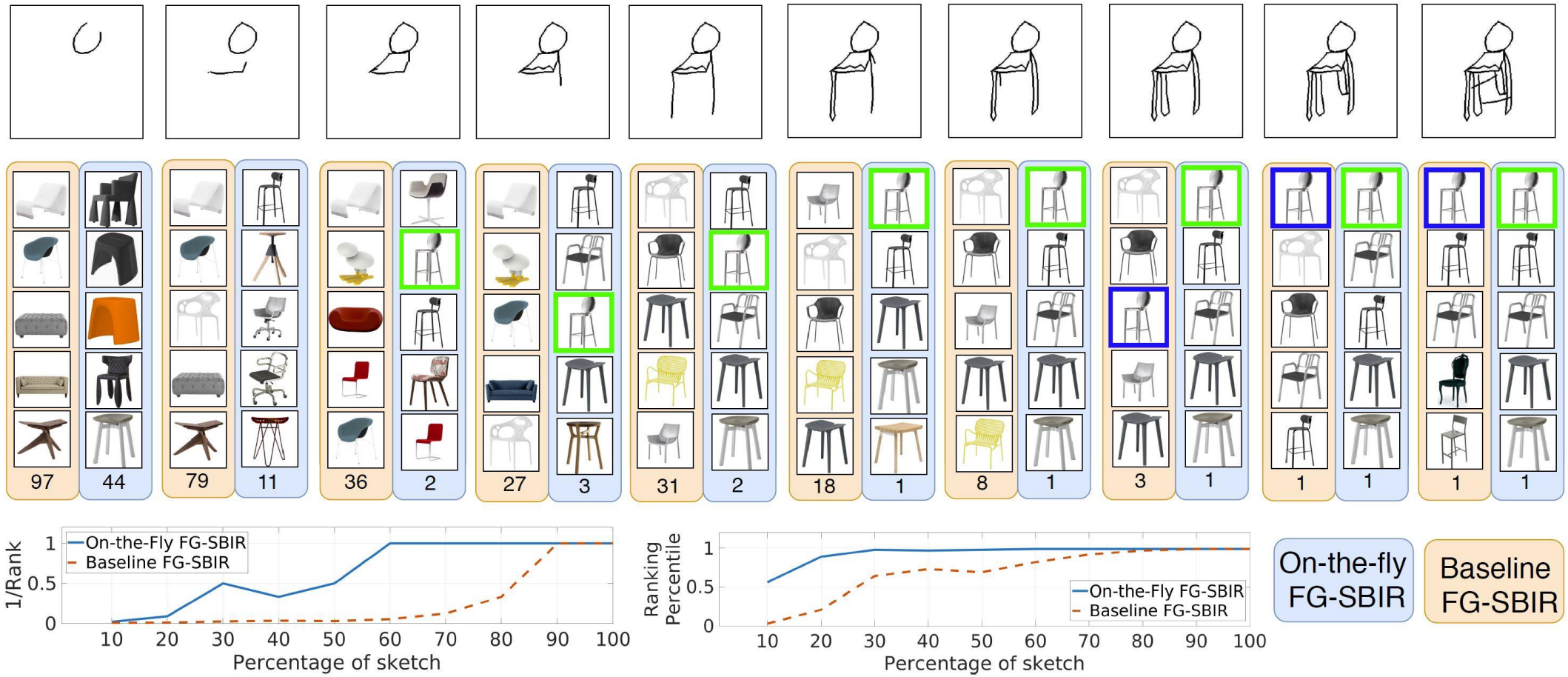
- **Evaluation Metric:**
  - top-q accuracy (A@q)
  - area under ranking percentile *vs* percentage of sketch (m@A)
  - area under 1/rank *vs* percentage of sketch (m@B)

- **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

On-the-Fly FG-SBIR

Results



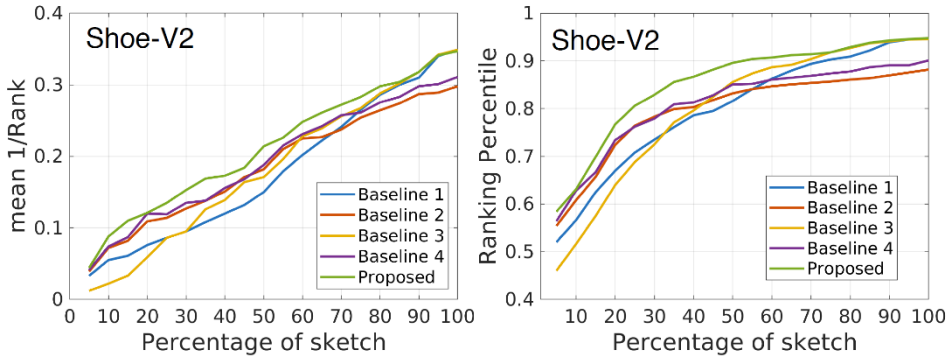
On-the-Fly FG-SBIR

Results

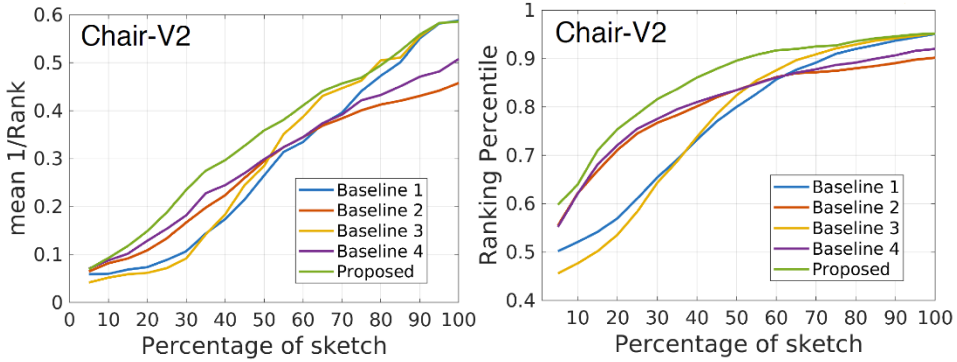
Quantitative Results on Different Baselines (A@q, m@A, and m@B)

	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)

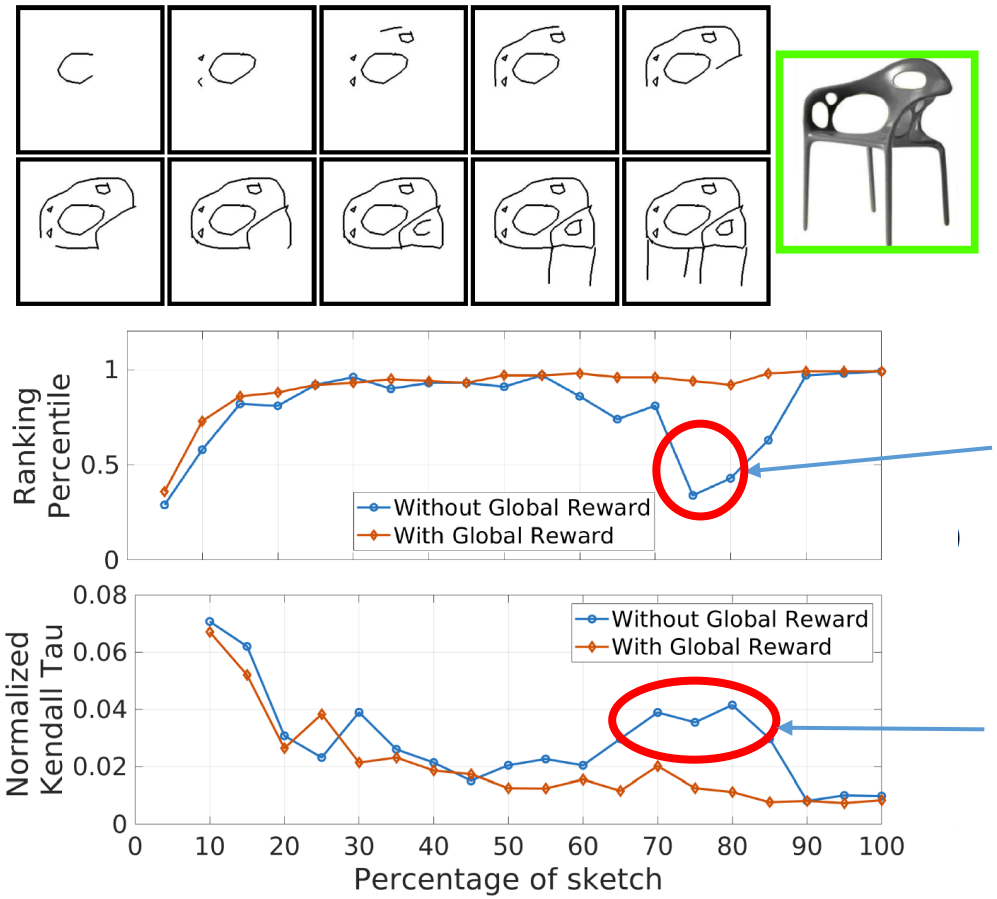


Percentage-wise Results for Chair-V2 (m@A, and m@B)



On-the-Fly FG-SBIR

Ablation Study



Progressive order of sketching

Drop in ranking percentile

Corresponding explosive increase of *Kendall-Tau distance*

On-the-Fly FG-SBIR

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**

On-the-Fly FG-SBIR

TL;DR:

- On-the-fly with early retrieval objective for better ease of sketch.
- Retrieve from partial sketches: *no need* to draw a complete sketch
- Through RL design, representation learning over a complete sketching episode.
- Local (early retrieval) + Global (consistency) reward for better performance.

Impact and Future Directions:

Specific -Sketch Community

- On-the-fly as a new benchmark paradigm [A, B, C]
- On-the-fly for other sketch-based tasks
  - e.g. Image generation/editing.

Broad Computer Vision Community

- On-the-fly for Text-based Image Retrieval
- Human in the loop.

Next: Semi-Supervised FG-SBIR?

[A] Bi-lstm sequence modeling for on-the-fly fine-grained sketch-based image retrieval, IEEE Trans on AI 2022.
[B] Deep reinforced attention regression for partial sketch-based image retrieval, ICDM 2021.
[C] Multi-granularity association learning framework for on-the-fly fine-grained sketch-based image retrieval, Knowledge Based System, 2022.

Semi-Supervised FG-SBIR

Sketch

Gallery Images

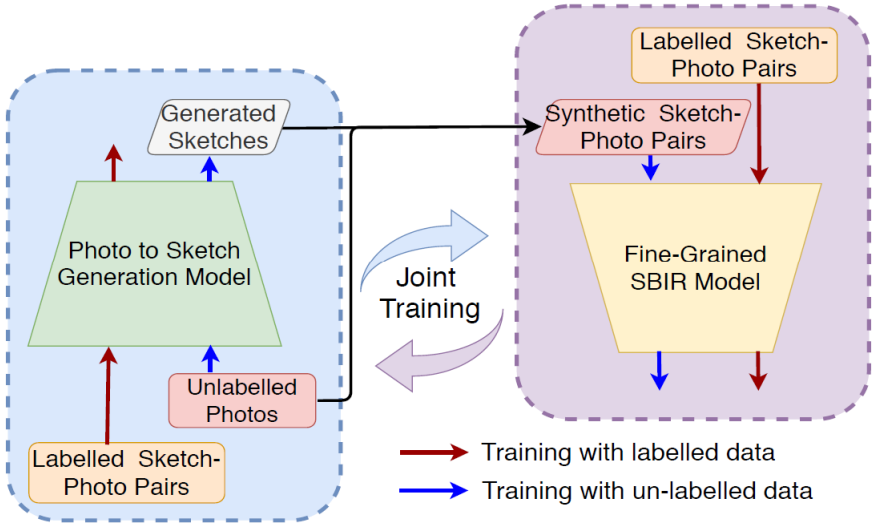
Problem – "Lack of photo-sketch pairs"

- ❑ photos can be *easily* scaled
- ❑ collecting **photo-sketch pairs is costly.**
- ❑ but sketches need to be *individually produced*.
- ❑ ***Our Solution***: make use of unlabelled photos - without having paired sketches.

Semi-Supervised FG-SBIR

Semi-Supervised Design

"generate paired sketches for unlabelled photos"

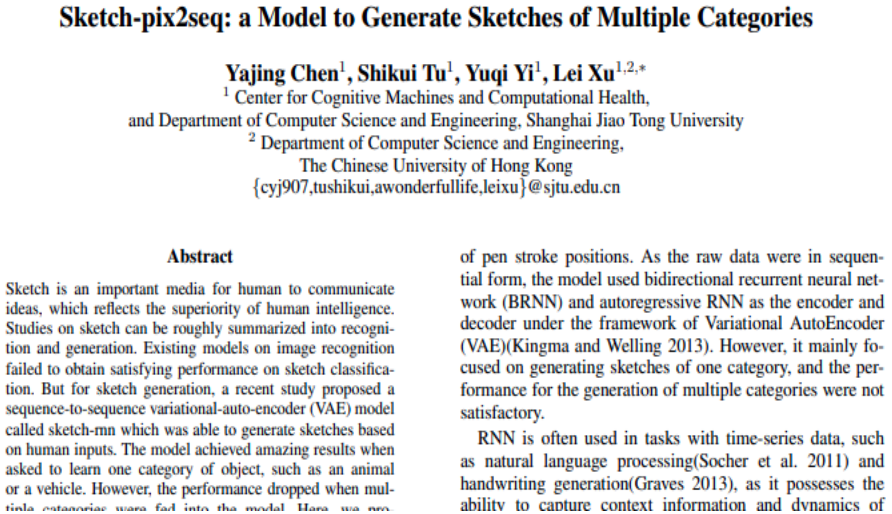


Note: *Unlabelled Photos* means – 'photos *without* associated paired sketches'

Semi-Supervised FG-SBIR

Why Challenging?

- Photo-to-sketch generation models could sometimes *generate unfaithful sketches*.
- Downstream Retrieval model have no way to know *which pseudo-sketch and photo pair* are worth training with.



Sketch-pix2seq: a Model to Generate Sketches of Multiple Categories

Yajing Chen[1], Shikui Tu[1], Yuqi Yi[1], Lei Xu[1,2,*]
[1] Center for Cognitive Machines and Computational Health,
and Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2] Department of Computer Science and Engineering,
The Chinese University of Hong Kong
{cyj907,tushikui,awonderfullife,leixu}@sjtu.edu.cn

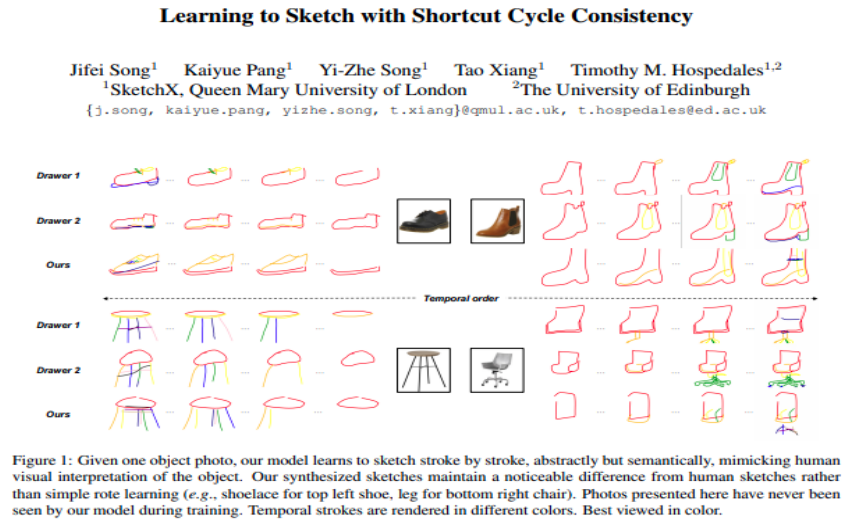
Abstract

Sketch is an important media for human to communicate ideas, which reflects the superiority of human intelligence. Studies on sketch can be roughly summarized into recognition and generation. Existing models on image recognition failed to obtain satisfying performance on sketch classification. But for sketch generation, a recent study proposed a sequence-to-sequence variational-auto-encoder (VAE) model called sketch-rnn which was able to generate sketches based on human inputs. The model achieved amazing results when asked to learn one category of object, such as an animal or a vehicle. However, the performance dropped when multiple categories were fed into the model. Here, we pro-

of pen stroke positions. As the raw data were in sequential form, the model used bidirectional recurrent neural network (BRNN) and autoregressive RNN as the encoder and decoder under the framework of Variational AutoEncoder (VAE)(Kingma and Welling 2013). However, it mainly focused on generating sketches of one category, and the performance for the generation of multiple categories were not satisfactory.

RNN is often used in tasks with time-series data, such as natural language processing(Socher et al. 2011) and handwriting generation(Graves 2013), as it possesses the ability to capture context information and dynamics of

A. Sketch-pix2seq: a Model to Generate Sketches of Multiple Categories, Chen et al., arXiv:1709.04121



Learning to Sketch with Shortcut Cycle Consistency

Jifei Song[1] Kaiyue Pang[1] Yi-Zhe Song[1] Tao Xiang[1] Timothy M. Hospedales[1,2]
[1]SketchX, Queen Mary University of London [2]The University of Edinburgh
{j.song, kaiyue.pang, yizhe.song, t.xiang}@qmul.ac.uk, t.hospedales@ed.ac.uk

Figure 1: Given one object photo, our model learns to sketch stroke by stroke, abstractly but semantically, mimicking human visual interpretation of the object. Our synthesized sketches maintain a noticeable difference from human sketches rather than simple rote learning (*e.g.*, shoelace for top left shoe, leg for bottom right chair). Photos presented here have never been seen by our model during training. Temporal strokes are rendered in different colors. Best viewed in color.

B. Learning to Sketch with Shortcut Cycle Consistency, Song et al., CVPR 2018

Semi-Supervised FG-SBIR

Contributions

- **Consistency loss** by using a pre-trained retrieval model (from labelled data) as a *"weak teacher".*
- **Discriminator guided instance weighting** to *quantify* the quality of synthetic photo-sketch pairs.
- **Joint Learning Framework** based on RL by *coupling* sketch generation with FG-SBIR.
- **First semi-supervised approach** to solve data scarcity problem in FG-SBIR.

Semi-Supervised FG-SBIR

Overall Pipeline

❑ Train a sequential Photo-to-Sketch Generator

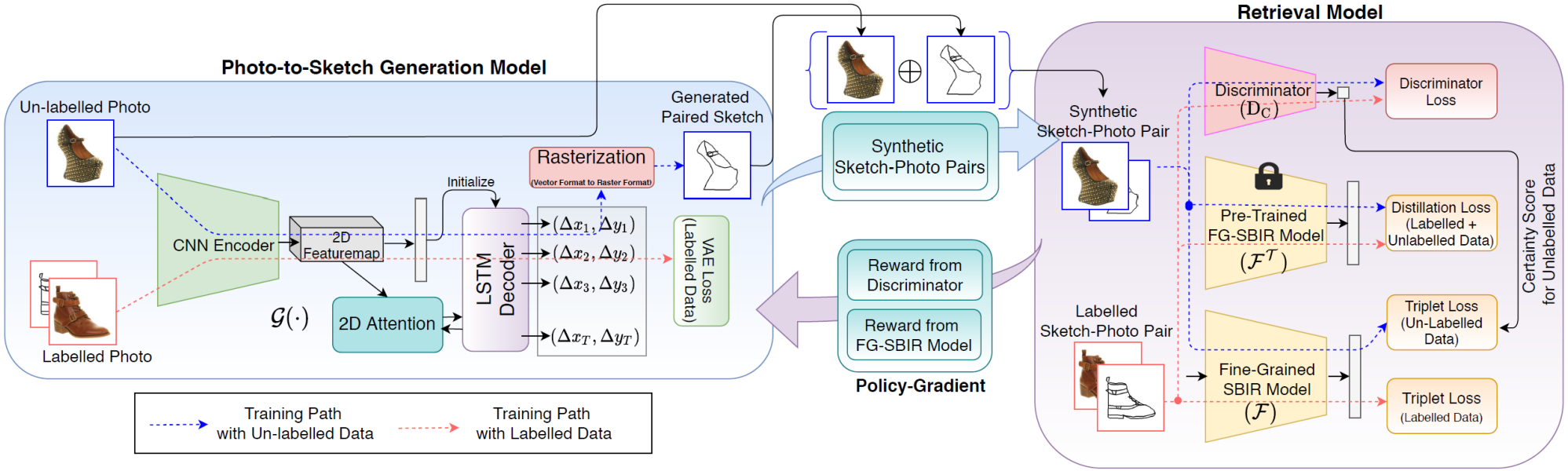
- Sequential sketch coordinate decoding – to model the hierarchical abstract nature of sketches

❑ Train the Retrieval model using *synthetic sketch-photo pairs* along with the *real sketch-photo pairs*

- Discriminator score to quantify the quality (certainty score) of synthetic pairs.
- Distillation loss to to provide tolerance against noisy training samples.
- Joint training: Generation and Retrieval improving each other.

Semi-Supervised FG-SBIR

A glimpse to our detailed framework



- **Why sequential sketch generation?**
  - to model the *hierarchical abstract nature* of sketch.

- Why **rasterization?**
  - FG-SBIR model requires *rasterized sketch-images* to obtain the sketch embedding.
  - Performance collapses on using sketch-coordinate instead.

Semi-Supervised FG-SBIR

Understanding each component of the Puzzle

Photo to Sketch Generation Model

$$\text{offsets } \Delta z = (\Delta x, \Delta y),\ \text{pen states } (p_1, p_2, p_3)$$

$$L_{\mathcal{G}}^{vae} = -\frac{1}{T}\Big[\sum_{i=1}^{T} \log p(\Delta z_i|\lambda_i) + \hat{p}\log(p_i)\Big] + \omega_{kl} L_{\mathcal{G}}^{kl}$$

Baseline FG-SBIR Model

$$L_{\mathcal{F}}^{trip} = \max\{0, \mu + \beta^{+} - \beta^{-}\};\ \text{where } \mu > 0$$

$$\beta^{+} = ||\mathcal{F}(a) - \mathcal{F}(p)||_2;\ \ \beta^{-} = ||\mathcal{F}(a) - \mathcal{F}(n)||_2;$$

Discriminator's Confidence to quantify the quality of Synthetic data

$$\mathrm{L_{D_C}} = -\mathbb{E}_{(\mathrm{p_L};\, \mathrm{s_{p,L}})\sim\mathcal{D}_{\mathrm{L}}}\left[\log \mathrm{D_C}\left(\mathrm{p_L}, \mathrm{s_{p,L}}\right)\right] - \mathbb{E}_{(\mathrm{p_U};\, \mathrm{s_{p,U}})\sim\mathcal{D}'_{\mathrm{U}}}\left[\log\left(1 - \mathrm{D_C}\left(\mathrm{p_U}, \mathrm{s_{p,U}}\right)\right)\right]$$

Consistency Loss using relative distillation from Weak Teacher

$$L_{\mathcal{F}}^{KD} = ||d(\mathcal{F}^T(p), \mathcal{F}^T(s_p)) - d(\mathcal{F}(p), \mathcal{F}(s_p))||_2$$

Semi-Supervised FG-SBIR

Joint Learning Framework: Two Conjugate Problems

Optimising FG-SBIR Model

$$\mathrm{L}_{\mathcal{F}}^{all} = L_{\mathcal{F}}^{trip}(\mathcal{D}_L^i) + \omega_j \cdot L_{\mathcal{F}}^{trip}(\mathcal{D'}_U^j) + \lambda_{kd} \cdot L_{\mathcal{F}}^{KD}(\mathcal{D}_L^i, \mathcal{D'}_U^j)$$

Triplet Loss
On Labelled data

Discriminator-based
Certainty Score for
Unlabelled data

Triplet Loss
On Un-labelled data

Distillation Loss
as a regularizer

Semi-Supervised FG-SBIR

Joint Learning Framework: Two Conjugate Problems

Optimising FG-SBIR Model

$$\mathrm{L}_{\mathcal{F}}^{all} = L_{\mathcal{F}}^{trip}(\mathcal{D}_L^i) + \omega_j \cdot L_{\mathcal{F}}^{trip}(\mathcal{D'}_U^j) + \lambda_{kd} \cdot L_{\mathcal{F}}^{KD}(\mathcal{D}_L^i, \mathcal{D'}_U^j)$$

Joint Training

Optimising Photo2Sketch Model

$$\nabla_{\theta_{\mathcal{G}}} L(\theta_{\mathcal{G}}) = \underbrace{\nabla_{\theta_{\mathcal{G}}} L_{\mathcal{G}}^{vae}(\theta_{\mathcal{G}})}_{\text{over only labelled data}}$$

$$-\lambda_{\mathcal{G}} \sum_{i=1}^{T} \underbrace{\mathbb{E}_{\substack{p_i \sim p(q_i) \\ \Delta z_i \sim p(\Delta z_i | \lambda_i)}} \nabla_{\theta_{\mathcal{G}}} \Big(\log p(\Delta z_i \mid \lambda_i) + \log p(p_i) \Big) \cdot R_{\mathcal{G}}}_{\text{over both labelled and unlabelled data (via policy gradient)}}$$

$$\mathrm{R}_{\mathcal{G}} = -\lambda_{\mathrm{r1}} \cdot \mathrm{L}_{\mathcal{F}}^{\mathrm{trip}}(\mathcal{D}^{\mathrm{i}}) + \lambda_{\mathrm{r2}} \cdot \mathrm{D}_{\mathrm{C}}(\mathcal{D}^{\mathrm{i}})$$

Reward

Semi-Supervised FG-SBIR

Experiments

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

- **Evaluation Metric:**
  - FG-SBIR: top-q accuracy (A@q)

  - Sketch Generation:
    - ✓ *Recognition* – using ResNet-50 sketch-classifier
    - ✓ *Retrieval* – top-q accuracy via a pre-trained FG-SBIR model
    - ✓ *Generation* – calculate FID score using pre-trained sketch-classifier.

Semi-Supervised FG-SBIR

Competitors:

- Sketch Generation:
  1) ***Pix2Pix*** [A] adapted to perform cross-modal translation in image space.
  2) ***PhotoSketch*** [B] allows one-to-many photo-conditioned sketch image generation.
  3) ***Pix2Seq*** [C] comprises of convolutional encoder and LSTM decoder, without 2D attention.
  4) ***L2S*** [D] uses two-way cross domain translation with self-domain reconstruction

- Fine-Grained SBIR:
  1) ***SN-Triplet*** [E] triplet ranking loss with Sketch-a-Net as its baseline feature extractor
  2) ***SN-HOLEF*** [F] extension over SN-Triplet using spatial attention along with higher order ranking loss
  3) ***SN-RL*** [G] reinforcement learning based fine-tuning for on-the-fly retrieval.
  4) **Edgemap-Pretrain** [H] use edge-maps of unlabelled photos to pre-train the retrieval model.

[A] Isola et al., Distilling the knowledge in a neural network.
[B] Li et al., Photo-Sketching: Inferring Contour Drawings from Images.
[C] Chen et al., Sketchpix2seq: a model to generate sketches of multiple categories.
[D] Song et al., Learning to sketch with shortcut cycle consistency.
[E] Qian et al., Sketch me that shoe.
[F] Song et al., Deep spatial-semantic attention for finegrained sketch-based image retrieval.
[G] Bhunia et al, Sketch less for more: On-the-fly fine-grained sketch based image retrieval.
[H] Muhammad et al., Learning deep sketch abstraction.

Semi-Supervised FG-SBIR

Results

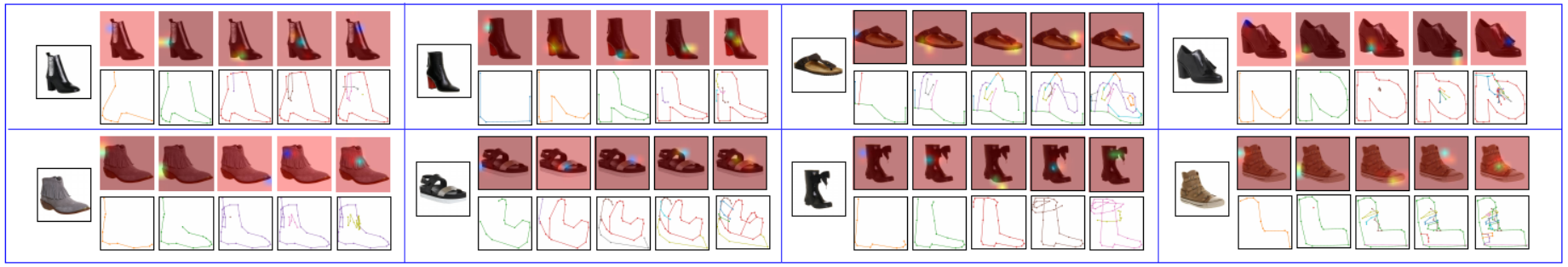
Quantitative results of fine-grained SBIR

Methods	Chair-V2		Shoe-V2	
	Acc.@1	Acc.@10	Acc.@1	Acc.@10
SN-Triplet [52]	47.4%	84.3%	28.7%	71.6%
SN-HOLEF [42]	50.7%	86.3%	31.2%	74.6%
SN-RL [5]	51.2%	86.9%	30.8%	74.2%
Edgemap-Pretrain [32]	53.9%	87.7%	33.8%	80.9%
Edge2Sketch-Pretrain [33]	54.3%	88.2%	34.2%	81.2%
Jigsaw-Pretrain [30]	56.1%	88.7%	36.5%	85.9%
Ours-F (only labelled data)	53.3%	87.5%	33.4%	80.7%
Vanilla-SSL-F	49.6%	85.6%	30.6%	74.3%
Ours-F-Pix2Pix	53.2%	87.5%	33.2%	80.1%
Ours-F-L2S	57.6%	89.4%	36.6%	84.7%
Ours-F-Full	60.2%	90.8%	39.1%	87.5%

Quantitative results of photo-to-sketch generation

Chair-V2	Recognition (↑)		Retrieval(↑)		FID Score(↓)
	Acc.@1	Acc.@10	Acc.@1	Acc.@10	
Pix2Pix [21]	4.5%	12.1%	2.4%	16.2%	33.4
PhotoSketch [25]	7.1%	14.3%	4.2%	17.9%	25.7
Pix2Seq [7]	5.4%	52.1%	4.0%	31.8%	14.5
L2S [41]	12.3%	53.8%	8.3%	36.7%	12.7
Ours-G (only labelled data)	15.2%	56.9%	13.4%	40.7%	10.1
Ours-G-Full	16.4%	58.6%	14.9%	42.6%	8.9
Shoe-V2	**Recognition(↑)**		**Retrieval(↑)**		**FID Score (↓)**
	Acc.@1	Acc.@10	Acc.@1	Acc.@10	
Pix2Pix [21]	6.2%	14.5%	1.8%	8.4%	31.7%
PhotoSketch [25]	8.9%	17.3%	3.4%	10.2%	24.3%
Pix2Seq [7]	51.3%	86.6%	5.1%	25.8%	11.3
L2S [41]	53.7%	89.7%	6.2%	28.6%	10.7
Ours-G (only labelled data)	56.3%	91.9%	9.7%	33.6%	9.5
Ours-G-Full	58.1%	93.4%	12.3%	35.4%	8.3

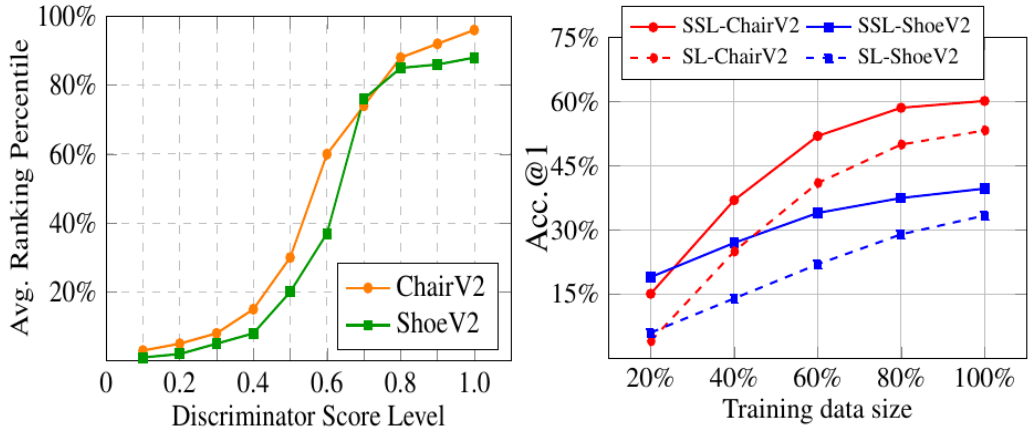
Qualitative results on our photo-to-sketch generation process, where sketch is shown with attention-maps at progressive instances.



Semi-Supervised FG-SBIR

Ablation Study

(Left) Consistency of discriminator's certainty score. (Right) Varying training data size for FG-SBIR – Semi-Supervised Learning (SSL) vs Supervised-Learning (SL).



Ablative study on Shoe-V2: Instance Weighting (IW), Teacher Regularisation (TR), Attention (AT), Joint-Training (JT)

IW	TR	AT	JT	Fine-Grained SBIR		Sketch Generation	
				Acc.@1	Acc.@10	Recognition Acc.@1	Retrieval Acc.@1
✓	✓	✓	✓	39.1%	87.5%	58.1%	12.3%
✗	✓	✓	✓	36.8%	85.4%	57.3%	11.2%
✓	✗	✓	✓	37.3%	86.1%	57.8%	12.1%
✓	✓	✗	✓	37.6%	86.1%	51.3%	5.1%
✓	✓	✓	✗	37.9%	86.6%	56.3%	9.7%
✗	✗	✗	✗	31.1%	75.4%	51.3%	5.1%

Semi-Supervised FG-SBIR

TL;DR:

- Synergise Photo-to-Sketch and FG-SBIR
- A novel discriminator guided instance weighting
- Distillation loss to provide tolerance against noisy synthetic samples.

Impact and Future Directions:

Specific -Sketch Community

- Photo-to-sketch generation as conjugate for Sketch-to-image generation.

Broad Computer Vision Community

- Cross-modal instance-level retrieval
- Useful for other instance-matching problem like Image captioning, text/tag based retrieval.
- Discriminator guided instance-specific – a generic design choice.

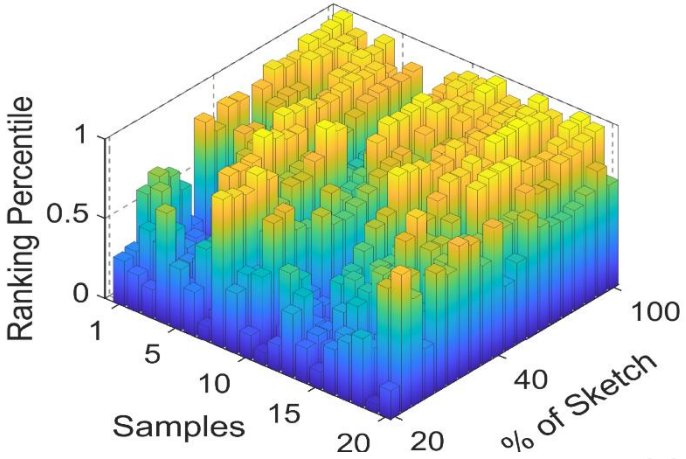
Next: Noise Tolerant FG-SBIR?

Noise-Tolerant FG-SBIR

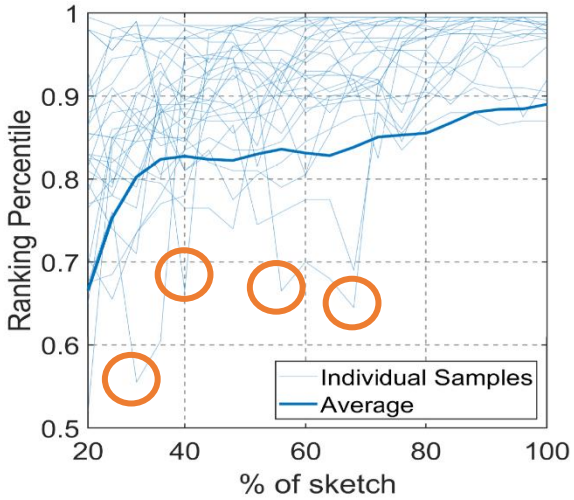
Motivation

Popular yet *less* used.

Main Challenge?



Statistics over the entire ShoeV2[1] dataset



Unwanted sudden drops

- It's not about how badly a sketch is drawn..
- It's about those ***extra irrelevant*** strokes !

Select an optimal subset of strokes

[1] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In CVPR, 2016.

Noise-Tolerant FG-SBIR

How to solve ??

- Consider all possible stroke subsets for training ?
  - Theoretically ***Maybe***, Practically **INFEASIBLE !!** – $O(2^N)$ complexity.
- Dropping strokes randomly?
  - Noisy gradient during training – too coarse.
- **Proposed Solution:**
  - *Detect* noisy strokes
  - *Eliminate* them to form meaningful subset of strokes.

How to implement??

- Quantify *stroke-relevance* based on its worth of retrieval.
- Stroke subset selector – binary selection using stroke level data from vector sketches.

Noise-Tolerant FG-SBIR

Side Benefits:

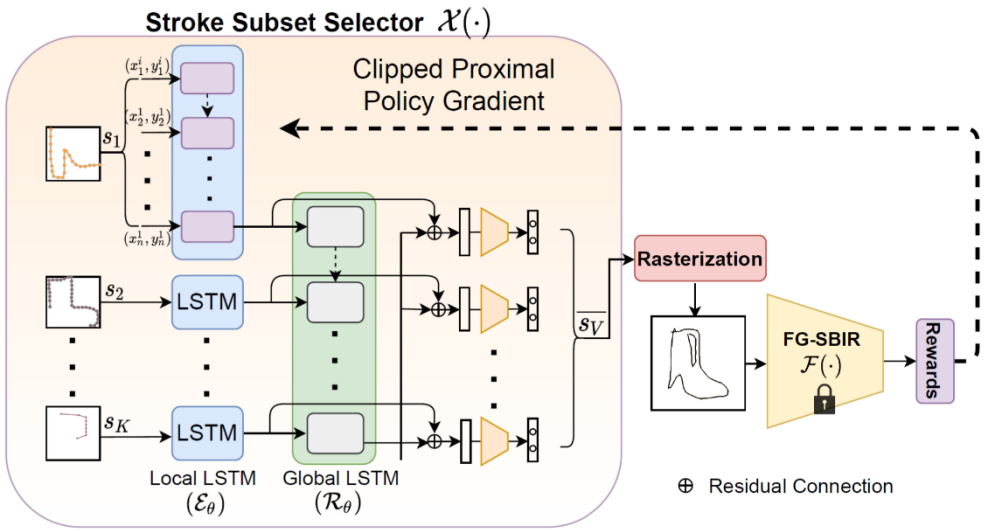
- ❑ Stroke importance quantifier.
- ❑ Can speed up existing works on interactive "on-the-fly" retrieval[1].
- ❑ Sketch data augmenter ***over*** random stroke dropping.

[1] Ayan Kumar Bhunia, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In CVPR, 2020.

Noise-Tolerant FG-SBIR

Methodology

- Stroke-subset selector as a pre-processing module in vector space
- Pretrained FG-SBIR (as a critic) that uses rasterized version of predicted subset
- Proximal Policy Optimization (PP0-AC) with clipped surrogate objective for optimization



Reward:

$$R = \omega_1 \cdot \frac{1}{rank} + \omega_2 \cdot (-\mathcal{L}_{Triplet})$$

Basic Policy-Gradient

$$L^{PG}(\theta) = -\frac{1}{K}\sum_{i=1}^{K} \log p_\theta(a_i|s_i) \cdot R$$

PPO-Actor

$$L^{A}(\theta) = -\frac{1}{K}\sum_{i=1}^{K} \min(L^{CPI}, L^{CLIP})$$

PPO Actor-Critic

$$L^{AC}(\theta) = -\frac{1}{K}\sum_{i=1}^{K} (L^{A} - c_1(V_\theta(S) - R)^2 + c_2 E_n)$$

[1] Ayan Kumar Bhunia, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In CVPR, 2020.

Noise-Tolerant FG-SBIR

Experiments

- **Datasets used:** QMUL-Shoe-V2[1] and QMUL-Chair-V2[1].

- **Competitors:**
  - State-of-the-art contemporary methods – Triplet-SN[1], Triplet-Attn-HOLEF[2], Triplet-RL[3], Mixed-Jigsaw[4], Semi-Sup[5], Cross-Hier[6].
  - **Baselines**
    - Siamese – Random stroke dropping to create multiple instances of same sketches.
    - Augment – Adds noisy strokes inside training to learn invariance against it.
    - StyleMeUp+Augment – Variant of StyleMeUp using noisy strokes in its inner loop.
    - Contrastive+Augment – Imposes additional contrastive loss.

- **Evaluation metric:**
  - Acc@k – Percentage of sketches having true-matched photos in the *top-k* list.
  - For On-the-fly FGSBIR[3] – ranking-percentile and $\frac{1}{rank}$ vs. *percentage of sketch* drawn.

[1] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In CVPR, 2016.
[2] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Deep spatial-semantic attention for finegrained sketch-based image retrieval. In CVPR, 2017.
[3] Ayan Kumar Bhunia, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In CVPR, 2020.
[4] Kaiyue Pang, Ke Li, Yongxin Yang, Honggang Zhang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Generalising fine-grained sketch-based image retrieval. In CVPR, 2019.
[5] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. More photos are all you need: Semi-supervised learning for FGSBIR. In CVPR, 2021.
[6] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Cross-modal hierarchical modelling for fine-grained sketch based image retrieval. In BMVC, 2020.

Noise-Tolerant FG-SBIR

Performance

Table 1: Results under Standard FG-SBIR setup.

		Chair-V2		Shoe-V2	
		Acc@1	Acc@5	Acc@1	Acc@5
SOTA	Triplet-SN [57]	47.4%	71.4%	28.7%	63.5%
	Triplet-Attn-HOLEF [47]	50.7%	73.6%	31.2%	66.6%
	Triplet-RL [8]	51.2%	73.8%	30.8%	65.1%
	Mixed-Jigsaw [34]	56.1%	75.3%	36.5%	68.9%
	Semi-Sup [4]	60.2%	78.1%	39.1%	69.9%
	StyleMeUp [41]	62.8%	79.6%	36.4%	68.1%
	Cross-Hier [40]	62.4%	79.1%	36.2%	67.8%
BL	(B)aseline-Siamese	53.3%	74.3%	33.4%	67.8%
	Augment	54.1%	74.6%	33.9%	68.2%
	StyleMeUp+Augment	56.1%	76.9%	36.9%	69.9%
	Contrastive+Augment	58.8%	77.1%	37.6%	70.1%
Limits	Upper-Limit	78.6%	90.3%	66.3%	88.3%
	Linear-Limit	59.4%	77.3%	42.5%	73.2%
	Proposed	64.8%	79.1%	43.7%	74.9%

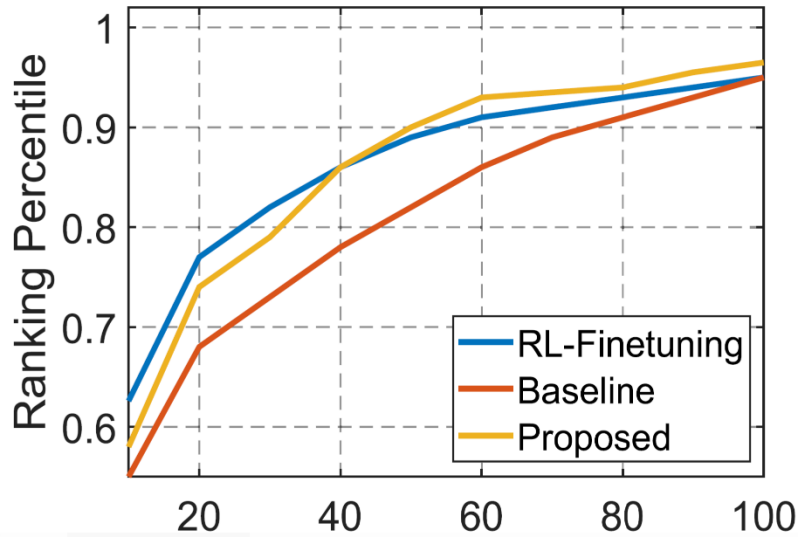
Quantitative Results

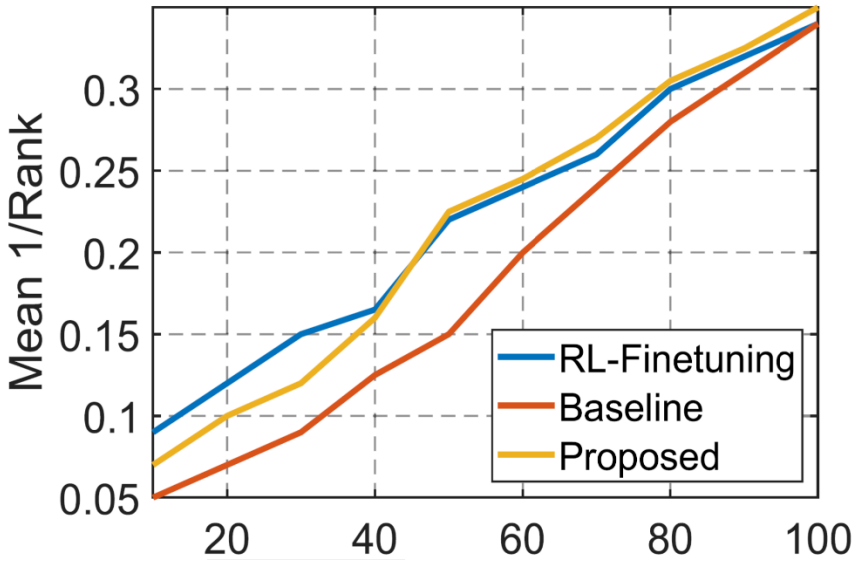
[1] Ayan Kumar Bhunia, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In CVPR, 2020.

Noise-Tolerant FG-SBIR

Performance

Comparative results under **On-The-Fly** setup (Shoe-V2)

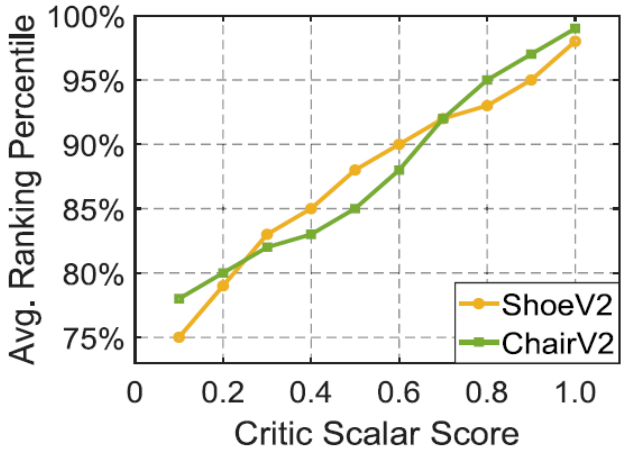


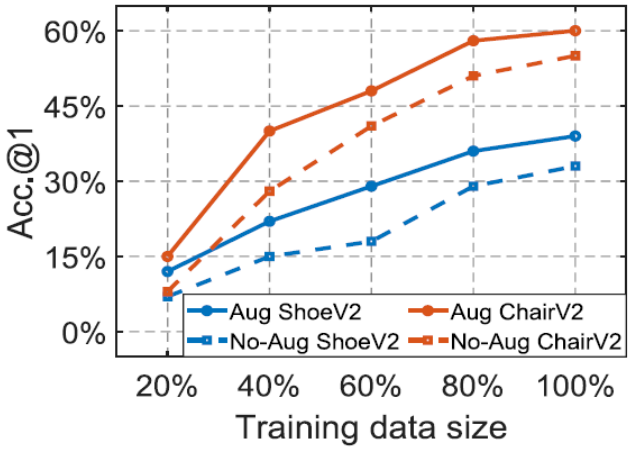


Higher area under the plots indicates better early retrieval performance

Noise-Tolerant FG-SBIR

Performance



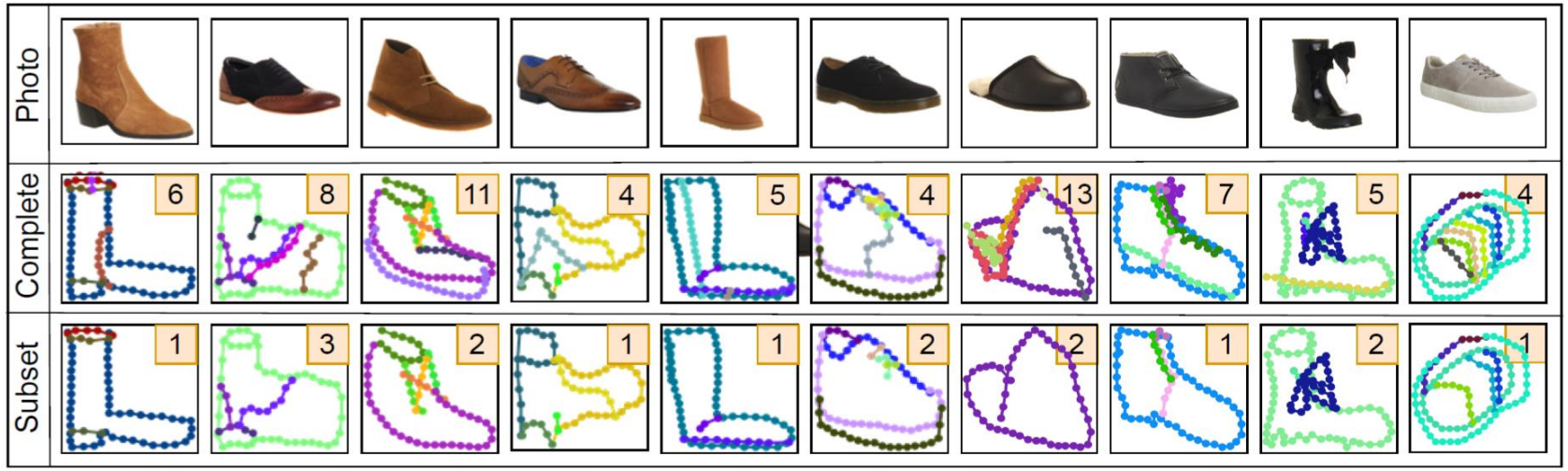


Retrieval ability of partial sketch: correlation between critic network V(S) predicted score and ranking percentile.

Performance at varying training data size with **stroke-subset selector based data augmentation**

Noise-Tolerant FG-SBIR

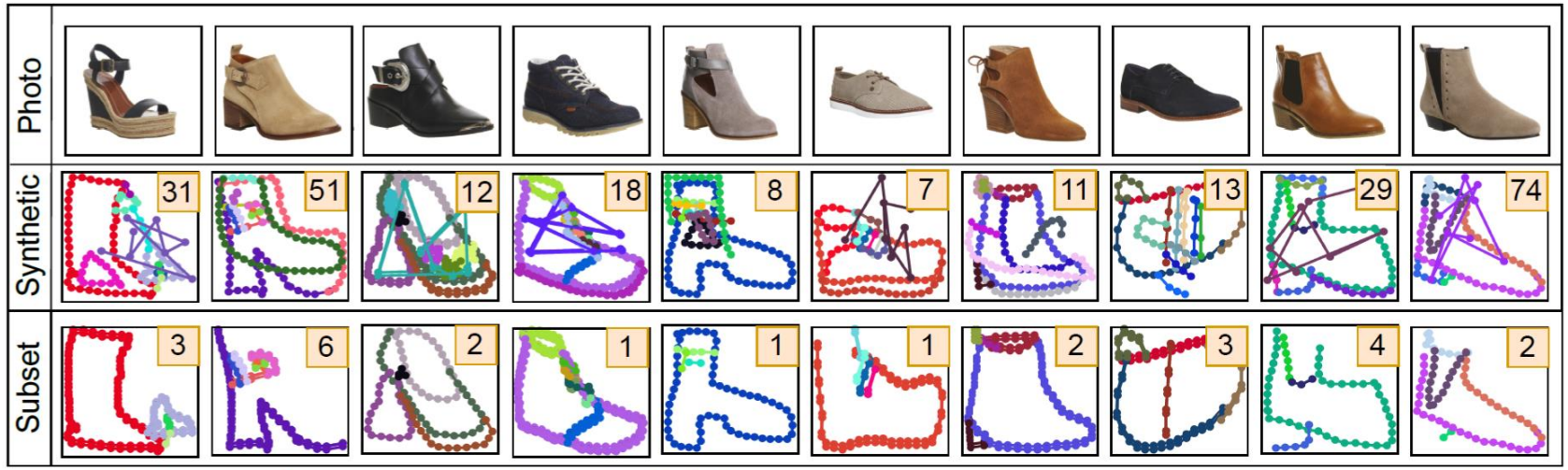
Qualitative Results



Examples showing selected subset performing better (rank in box) than complete sketch from ShoeV2.

Noise-Tolerant FG-SBIR

Qualitative Results



Examples showing ability to perform (rank in box) under *synthetic* noisy sketch input on ShoeV2.

Noise-Tolerant FG-SBIR

TL;DR:

- A stroke sub-set selector as a pre-processing module in vector space.
- Stroke subset selector is trained using reinforcement learning.
- Multiple side-benefits [sketch augmentation, retrievability, stroke-wise importance, etc.]

Impact and Future Directions:

Specific -Sketch Community

- Hierarchical sketch-vector encoder as a generic feature extractor for various task
- Removing inconsistent strokes for sketch2RGB

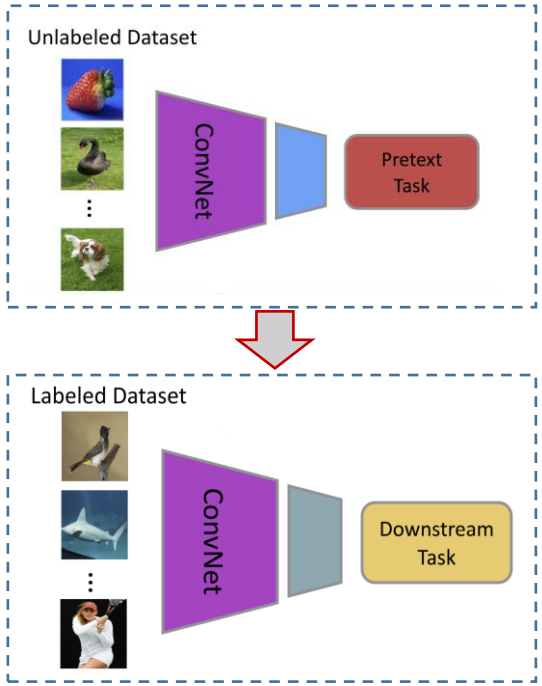
Broad Computer Vision Community

- Video/text summarisation
- Pre-trained model as critic where we do not have any hard labels.

Self-Supervised Learning on Sketches

Background on Self-Supervised Learning

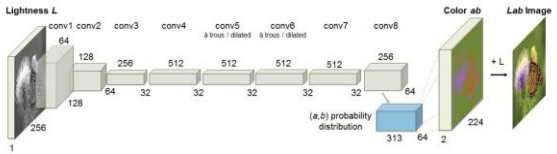
Self-Supervised *pretext* task training



Self-Supervised *Downstream* task training

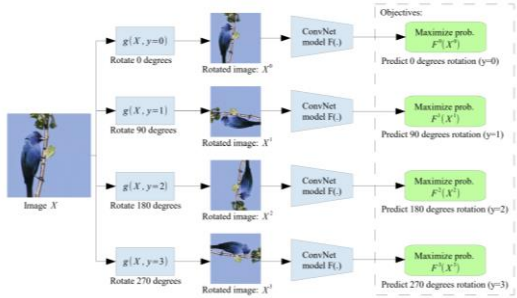
- **Main criteria of *pretext* task :**
  (a) free of cost labels
  (b) able to encode high-level semantic understanding of the data

- **Existing popular *pretext* task:**



(a) Image colorization

(b) Jigsaw solving



(c) Image rotation prediction

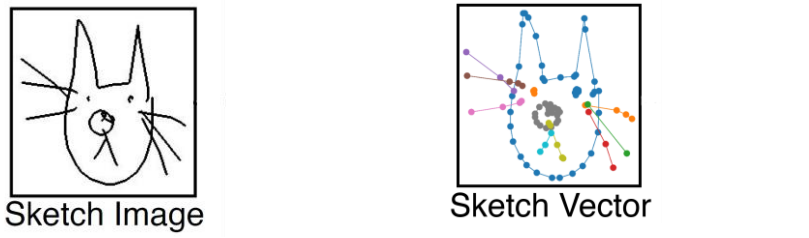
(d) Image super-resolution
(e) Image in-painting
(f) Many more……

A. Image Credit: Jing et al. Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey

Self-Supervised Learning on Sketches

Why sketch needs specific attention for *pretext* task design:

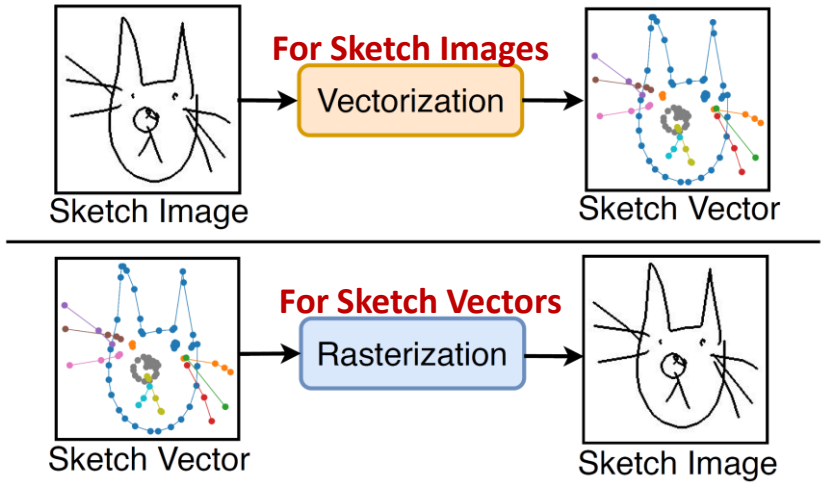
- Sketch is different from photos.
- Many existing pretext tasks do not fit for sketches.
- Sketch has *dual modality* of representation.



Rasterized pixel space **Temporal sequence of points**

Our Objective:

- A common pretext task.
- Exploits the ***dual-nature*** of sketch.
- Simple and easy to implement.



Self-Supervised Learning on Sketches

Self-supervised learning literature:

(a) **Generative Modeling**: VAE [A], etc.

(b) **Contrastive learning**: SimCLR [B], BYOL[C], MoCo [D], etc.

(c) **Clustering based approach**: Deep Cluster [E], etc.

(d) **Defining various pretext tasks**: Colorization [F], Super-resolution [G], jigsaw-solving [H], frame-order prediction [I], etc.

Specific to multi-modal data:

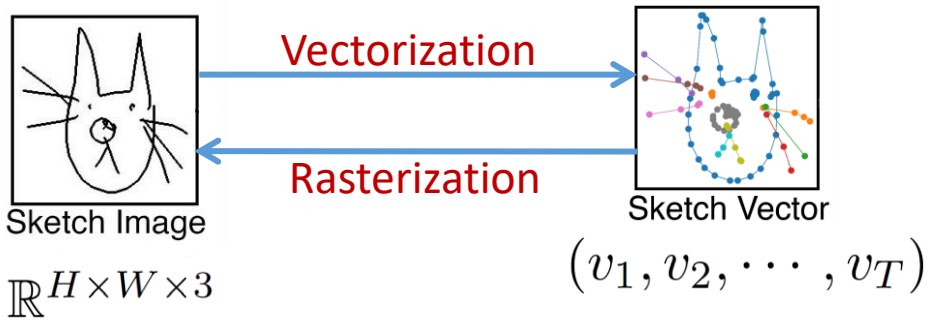
(a) Learning visual-audio correspondence [J]

(b) RGB-flow depth correspondence [K]

A. Kingma et al., Auto-encoding variational bayes.
B. Chen et al., A simple framework for contrastive learning of visual representations.
C. Grill et al., Bootstrap your own latent.
D. He et al,, Momentum contrast for unsupervised visual representation learning.
E. Caron et al., Deep clustering for unsupervised learning of visual features.
F. Zhang et al., Colorful image colorization.
G. Ledig et al, Photorealistic single image super-resolution using a generative adversarial network.
H. Nooroozi et al., Unsupervised learning of visual representations by solving jigsaw puzzles.
I. Mishra et al., Unsupervised learning using temporal order verification.
J. Korbar et al., Cooperative learning of audio and video models from self-supervised synchronization.
K. Tian et al, Contrastive multiview coding.

Self-Supervised Learning on Sketches

Methodology



Each temporal point

$$(x_t, y_t, q_t^1, q_t^2, q_t^3)$$

a) **Training data**: paired sketch-image and sketch-vector.

b) No costly annotation.

c) Encoder-Decoder architecture.

d) Only encoder is used for feature extraction on downstream tasks.

Self-Supervised Learning on Sketches

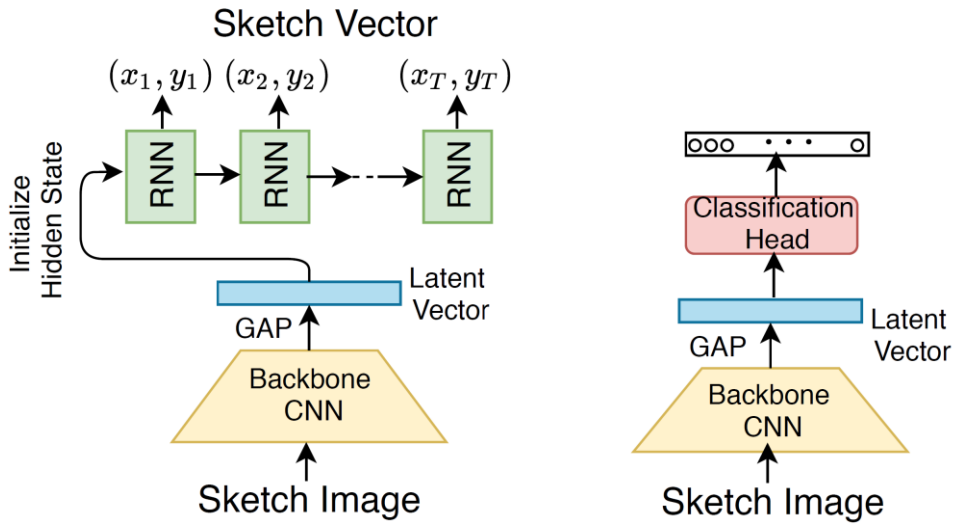
Vectorization

(a) Image Encoder: e.g., ResNet

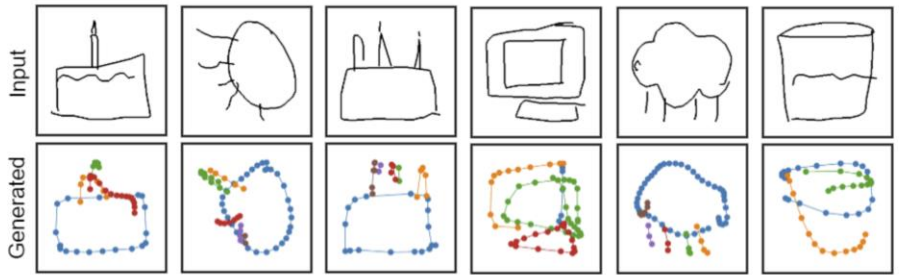
(b) Sequential Decoder: e.g., RNN

Training Objective:

$$L_{I \to V} = \frac{1}{T} \sum_{t=1}^{T} \|\hat{x}_t - x_t\|_2 + \|\hat{y}_t - y_t\|_2$$
$$- \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{3} \hat{q}_t^i \log \left(\frac{\exp(q_t^i)}{\sum_{j=1}^{3} \exp(q_t^j)} \right)$$



Self-supervised Training **Downstream Task**



Vectorization: raster sketch image to vector sketch

Self-Supervised Learning on Sketches

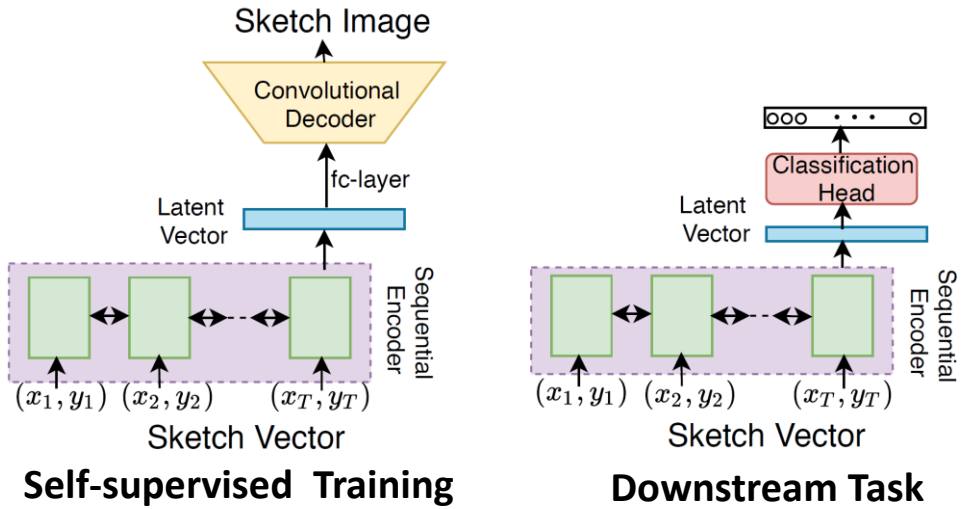
Rasterization

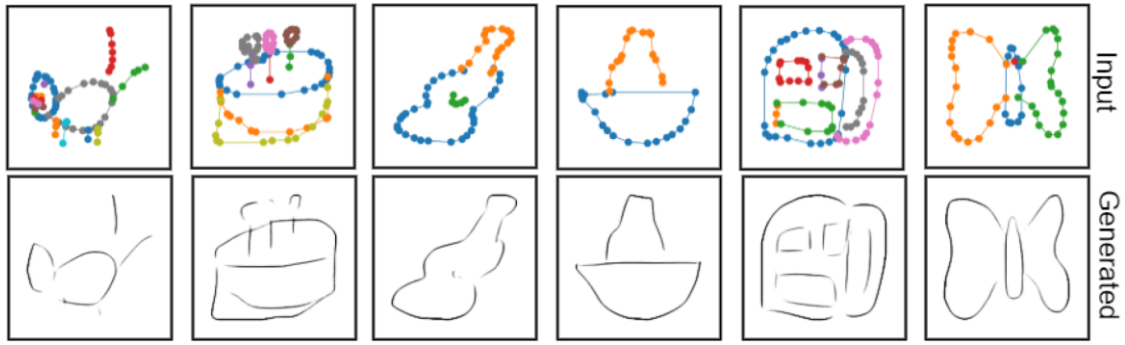
(a) Sequential Encoder: e.g., RNN or Transformer

(b) Convolutional Decoder: with Deconvolutional Layers

Training Objective:

$$L_{V \to I} = -\mathbb{E}_{(I,V)\sim(\mathcal{I},\mathcal{V})} \left\| I - D_I(E_V(V)) \right\|_2$$





Rasterization: vector sketch to raster sketch image.

Self-Supervised Learning on Sketches

Experiments

- **Application of learned representation:**
  (a) Sketch recognition and retrieval
  (b) Handwriting recognition

- **Datasets:**
  - Sketch Datasets: QuickDraw [C] and TU-Berlin [B]
  - Handwriting Dataset: IAM

- **Implementation Details**:
  - ResNet50 [A] as sketch image encoder
  - Transformer as sketch vector encoder

- **Evaluation protocol and metric:**
  - Linear evaluation and semi-supervised evaluation (using 1% and 10% data).
  - Top-1 and top-5 for classification.
  - Acc@top1 and mAP@top10 for retrieval.

A. Deep residual learning for image recognition, CVPR'16
B. How Do Humans Sketch Objects?, Siggraph'12
C. A Neural Representation of Sketch Drawings, ICLR'18.

Self-Supervised Learning on Sketches

Competitors:

- **By different Pre-text task:**
  - Context prediction [A]
  - Auto-Encoding [B]
  - Jigsaw solving [C]
  - Rotation prediction [D]
- **Contrastive Methods:**
  - SimCLR [E]
  - BYOL [F]
  - MoCo [G]
- Sketch-BERT [H]
- Contrastive Predictive Coding [I]
- Contrastive Multi-view Coding (CMC) [J]

A. Doersch et al., Unsupervised visual representation learning by context prediction.
B. Kingma et al., Auto-encoding variational bayes.
C. Noorozi et al., Unsupervised learning of visual representations by solving jigsaw puzzles
D. Gidaris et al., Unsupervised representation learning by predicting image rotations
E. Chen et al., A simple framework for contrastive learning of visual representations.
F. Grill et al., Bootstrap your own latent.
G. He et al., Momentum contrast for unsupervised visual representation learning.
H. Lin et al., Sketch-bert: Learning sketch bidirectional encoder representation from transformers by self-supervised learning of sketch gestalt.
I. Oord et al., Representation learning with contrastive predictive coding
J. Tian et al., Contrastive multiview coding.

Self-Supervised Learning on Sketches

Results on Sketch Datasets:

Linear Evaluation

Table 1. Linear model evaluation of fixed pre-trained features. ResNet50 for image space and Transformer for vector space inputs.

	Recognition								Retrieval							
	Image Space				Vector Space				Image Space				Vector Space			
	QuickDraw		TU-Berlin		QuickDraw		TU-Berlin		QuickDraw		TU-Berlin		QuickDraw		TU-Berlin	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	A@T1	mAP@t10	A@T1	mAP@t10	A@T1	mAP@T10	A@T1	mAP@T10
Supervised	76.1%	91.3%	78.6%	90.1%	73.5%	90.1%	62.9%	80.7%	62.3%	69.4%	69.1%	74.7%	58.5%	77.1%	50.2%	67.4%
Random	15.5%	26.2%	18.4%	29.3%	12.7%	23.6%	9.6%	19.4%	10.6%	21.3%	13.4%	26.7%	9.8%	21.5%	9.2%	17.6%
Context [15]	44.6%	69.2%	43.3%	67.5%	-	-	-	-	30.7%	34.9%	28.4%	32.7%	-	-	-	-
Auto-Encoder [31]	26.4%	48.1%	22.6%	47.5%	-	-	-	-	16.4%	24.4%	15.3%	20.4%	-	-	-	-
Jigsaw [43]	46.9%	71.5%	45.7%	69.8%	-	-	-	-	31.6%	38.9%	30.6%	35.4%	-	-	-	-
Rotation [18]	53.5%	78.7%	51.2%	77.1%	-	-	-	-	37.5%	45.1%	36.4%	41.8%	-	-	-	-
Deep Cluster [10]	39.4%	62.7%	38.7%	60.2%	-	-	-	-	29.2%	36.8%	27.3%	31.9%	-	-	-	-
MoCo [23]	65.7%	85.1%	64.3%	82.8%	-	-	-	-	42.5%	46.8%	42.5%	46.9%	-	-	-	-
SimCLR [11]	65.5%	85.1%	64.3%	82.9%	-	-	-	-	43.3%	50.7%	41.5%	46.7%	-	-	-	-
BYOL [21]	66.8%	85.8%	65.7%	83.7%	-	-	-	-	45.4%	52.5%	43.8%	49.1%	-	-	-	-
Sketch-BERT [37]	-	-	-	-	65.6%	85.3%	52.9%	78.1%	-	-	-	-	48.9%	68.1%	40.7%	58.8%
CMC [59]	63.6%	83.9%	61.7%	81.3%	61.2%	81.5%	51.4%	77.5%	40.6%	45.8%	38.5%	43.3%	45.2%	66.7%	40.3%	58.2%
CPC [44]	54.3%	79.0%	52.9%	77.9%	59.3%	81.3%	50.5%	76.6%	37.9%	43.1%	36.4%	40.9%	43.1%	63.6%	39.3%	57.9%
Ours-(L)	71.9%	89.7%	70.6%	85.9%	67.2%	86.5%	55.6%	79.4%	52.3%	59.5%	47.7%	59.1%	49.5%	68.9%	42.1%	59.6%

Semi-Supervised Evaluation

Table 2. Semi-supervised fine-tuning using 1% and 10% labelled training data on QuickDraw.

	Recognition								Retrieval							
	Image Space				Vector Space				Image Space				Vector Space			
	1% Training		10% Training		1% Training		10% Training		1% Training		10% Training		1% Training		10% Training	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	A@T1	mAP@t10	A@T1	mAP@t10	A@T1	mAP@T10	A@T1	mAP@T10
Supervised	25.1%	47.3%	55.4%	79.0%	17.3%	37.5%	43.9%	65.9%	13.4%	34.3%	43.9%	63.7%	9.1%	29.0%	41.0%	60.8%
Context [15]	33.9%	55.8%	56.8%	80.5%	-	-	-	-	24.4%	30.5%	42.6%	48.4%	-	-	-	-
Auto-Encoder [31]	21.5%	40.7%	45.1%	70.6%	-	-	-	-	15.2%	21.4%	32.7%	37.6%	-	-	-	-
Jigsaw [43]	36.5%	57.4%	57.4%	80.3%	-	-	-	-	27.7%	35.2%	44.7%	51.2%	-	-	-	-
Rotation [18]	38.8%	59.1%	59.6%	80.7%	-	-	-	-	28.4%	35.2%	44.7%	51.8%	-	-	-	-
Deep Cluster [10]	32.2%	54.5%	54.7%	79.2%	-	-	-	-	24.4%	31.2%	43.6%	47.7%	-	-	-	-
MoCo [23]	46.0%	70.5%	62.2%	83.7%	-	-	-	-	35.9%	43.1%	52.7%	57.4%	-	-	-	-
SimCLR [11]	46.1%	70.5%	62.1%	83.6%	-	-	-	-	35.1%	42.7%	52.3%	57.4%	-	-	-	-
BYOL [21]	47.3%	72.0%	62.7%	84.1%	-	-	-	-	36.5%	43.0%	52.8%	59.8%	-	-	-	-
Sketch-BERT [37]	-	-	-	-	45.1%	69.8%	62.4%	81.7%	-	-	-	-	36.5%	60.0%	52.9%	72.9%
CMC [59]	44.6%	68.2%	61.7%	82.7%	44.6%	68.4%	61.7%	81.6%	34.7%	41.9%	51.1%	57.4%	35.4%	58.1%	52.6%	72.8%
CPC [44]	40.6%	65.7%	60.7%	81.9%	43.5%	67.7%	61.6%	81.7%	33.4%	40.1%	50.5%	57.7%	34.1%	56.6%	52.3%	72.8%
Ours	51.2%	76.4%	65.6%	85.2%	46.8%	70.9%	63.2%	83.9%	38.6%	45.6%	60.4%	81.4%	37.1%	61.5%	53.2%	74.3%

Self-Supervised Learning on Sketches

Results on Handwriting Datasets:

Linear Evaluation

Table 4. Handwriting recognition using feature extracted from fixed pre-trained encoder.

	Offline		Online	
	Lexicon	No Lexicon	Lexicon	No Lexicon
Supervised [56]	87.1%	81.5%	88.4%	82.8%
Random	10.4%	6.3%	7.4%	4.9%
CPC [44]	72.2%	63.7%	71.5%	62.8%
Ours	75.4%	68.6%	73.1%	66.9%

Semi-Supervised Evaluation

Table 5. Handwriting recognition under semi-supervised setup.

	Offline		Online	
	1% Training	10% Training	1% Training	10% Training
Supervised [56]	19.7%	40.6%	20.5%	42.4%
CPC [44]	29.1%	55.4%	27.8%	54.2%
Ours	38.5%	59.2%	36.8%	56.7%

Self-Supervised Learning on Sketches

Ablative study on architecture design:

Table 6. Ablative study (Top-1 accuracy) on architectural design using QuickDraw. (V)ectorization and (R)asterization indicate representation learning on image and vector space, respectively.

Ablation Experiment	Image Space	Vector Space
(a) Absolute coordinate in the decoding (V):	71.9%	–
(b) Offset coordinate in the decoding (V)	69.5%	–
(c) Absolute coordinate in the encoding (R):	–	67.2%
(d) Offset coordinate in the encoding (R)	–	67.1%
(e) LSTM decoder (V) :	70.7%	–
(f) GRU decoder (V) :	71.9%	–
(g) Transformer decoder (V) :	68.6%	–
(h) LSTM encoder (R) :	–	66.7%
(i) GRU encoder (R) :	–	66.1%
(j) Transformer encoder (R) :	–	67.2%
(k) Two-way Translation (V+R) :	70.3%	66.1%
(l) Attentional Decoder (V) :	68.0%	–

Design question?

(i) Absolute coordinate vs offset coordinate
(ii) LSTM vs GRU vs Transformer decoder
(iii) LSTM vs GRU vs Transformer encoder
(iv) If attentional decoder helpful

Self-Supervised Learning on Sketches

Generalizability of representation learning:

(a) Cross-Category Generalization:

Table 7. Cross-category recognition accuracy on QuickDraw.

	Image Space		Vector Space	
	Top-1	Top-5	Top-1	Top-5
MoCo [23]	53.4%	77.6%	–	–
SimCLR [11]	53.6%	77.6%	–	–
CPC [44]	46.8%	71.3%	48.1%	73.3%
Ours	65.1%	85.6%	58.4%	81.2%

(b) Cross-Dataset Generalization:

Table 8. Cross-dataset (QuickDraw $\mapsto$ Tu-Berlin) recognition accuracy: Model pre-trained on QuickDraw is used to extract fixed latent feature on TU-Berlin, followed by linear model evaluation.

	Image Space		Vector Space	
	Top-1	Top-5	Top-1	Top-5
MoCo [23]	47.5%	62.1%	–	–
SimCLR [11]	47.2%	62.0%	–	–
CPC [44]	41.4%	60.8%	27.7%	50.9%
Ours	58.9%	80.5%	36.9%	61.7%

(c) Cross-Task Generalization:

Table 9. Cross-task (Sketch $\leftrightarrow$ Handwriting) generalisation results on extracted fixed latent feature . Lexicon: (L), No-Lexicon: (NL)

	Sketch (QuickDraw)				Handwriting (IAM)			
	Image		Vector		Image		Vector	
	Top-1	Top-5	Top-1	Top-5	L	NL	L	NL
Random	14.6%	25.7%	11.8%	22.9%	9.8%	6.1%	7.1%	4.5%
CPC [44]	19.7%	37.8%	17.6%	36.9%	19.5%	12.5%	15.7%	9.7%
Ours	37.6%	58.4%	33.7%	55.8%	33.8%	28.4%	31.6%	26.3%

TSNE Visualization

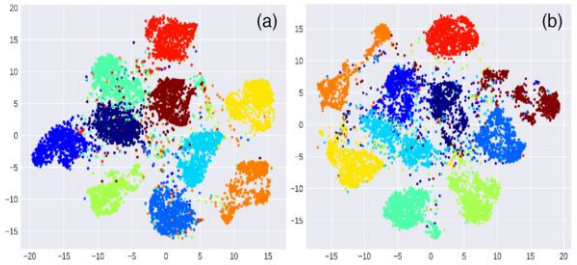


Figure 6. T-SNE Plots on features extracted by our self-supervised method (a) vectorization (sketch images) (b) rasterization (sketch vectors) for 10 QuickDraw classes.

Self-Supervised Learning on Sketches

TL;DR:

- First self-supervised pre-text task for sketch and handwriting.
- Applicable to both raster and vector format data.
- Dual raster/vector representation nature unique to sketch

Impact and Future Directions:

Specific -Sketch Community

- Sketch and handwriting could help each other for unsupervised feature learning
- A common latent space for raster and vector modality, and some mutual information-based modelling could be further explored.
- Recently adapted in zero-shot SBIR [A] under test-time adaptation pipeline

Broad Computer Vision Community

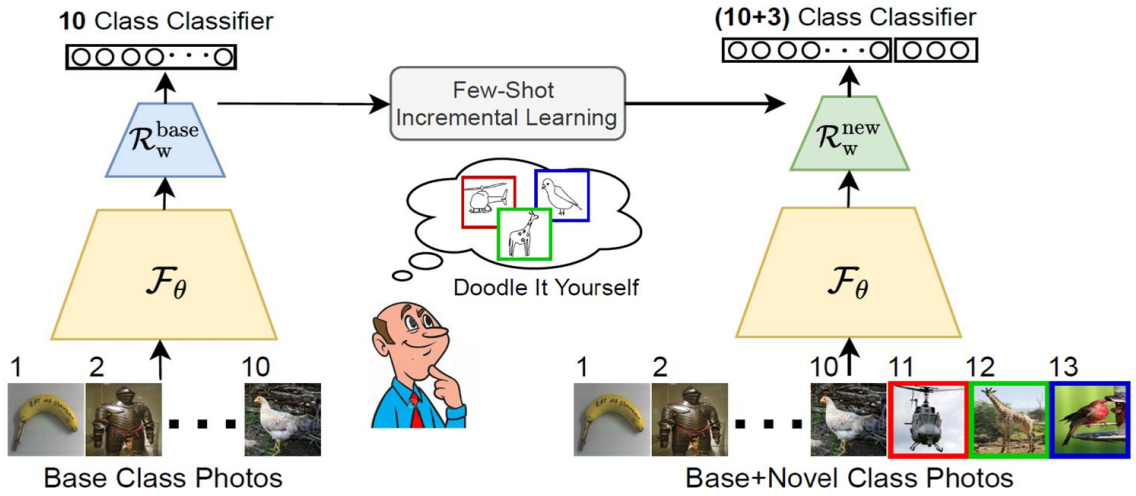
- Sketch2Vec self-supervised task could be extended for vector graphics data.
- Relevant in context of multi-modal self supervised learning.

[A] Sketch3T: Test-Time Training for Zero-Shot SBIR, CVPR 2022.

Sketch-Based Incremental Learning

What is Class Incremental Learning?

- A 10-class photo classifier to (10+13) class-classifier



Issues:

- Modality of support-samples?
- How to achieve them?

Intuition:

- Humans can learn from various modalities – not just photos.
- Photos are not always available owing to privacy/ethical constraints.

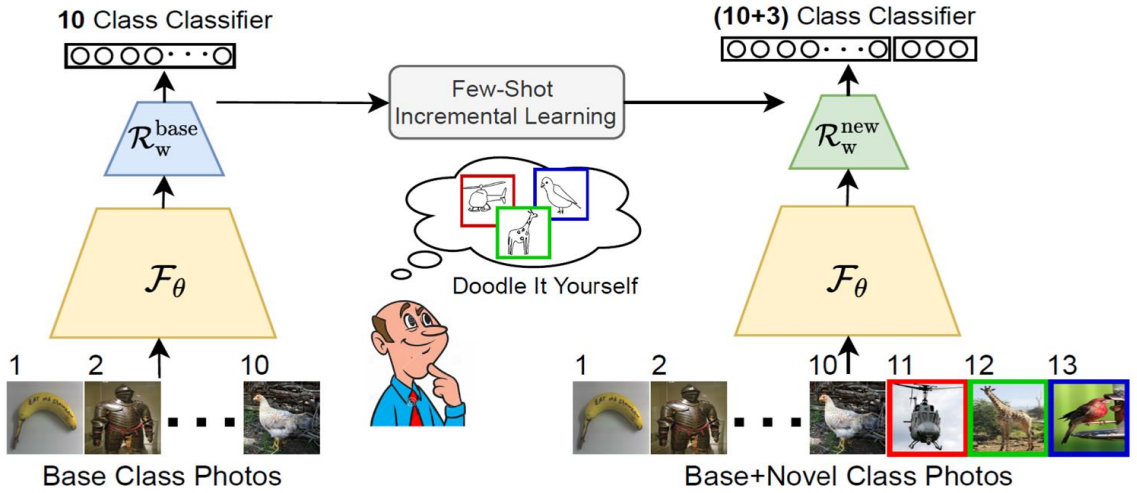
Proposed Solution:

- Doodle some sketches of new classes.
- Flexible cross-modal learning.
- No more source photos needed.

Sketch-Based Incremental Learning

Why Sketch for Incremental Learning?

- Faithful visual representative of images – extensively used in Sketch based Image Retrieval.
- Abstract yet *detailed*.
- One of the most *expressive mediums* for *humans* to describe an image.



Sketch-Based Incremental Learning

Challenges for Sketch-Based Few-Shot Class Incremental Learning

1. How to make the model work **cross-modal**?
   - Gradient-consensus based strategy to make a domain invariant feature extractor.

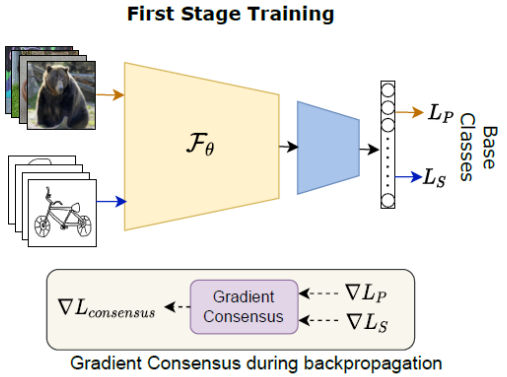
2. How to **preserve** old class information?
   - Knowledge distillation loss to *retain* old knowledge while acquiring new classes' data.

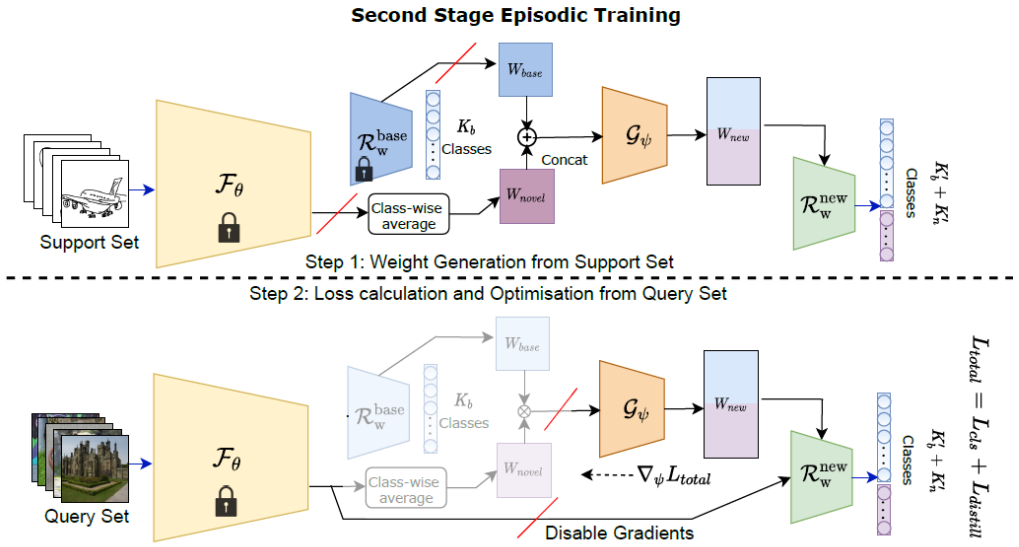
3. How to **leverage information** from old classes to learn new ones ?
   - Generate highly discriminative decision boundaries, via message passing between old and novel classes.

Sketch-Based Incremental Learning

Framework: Two Stage Training

- Base classes – sufficiently labelled photos and sketches
- Novel classes – limited number of sketches only.
- Two stage training
  - Cross modal pre-training for base classes
    - Gradient consensus for modality invariant feature extractor
  - Few shot classifier weight generation
    - Episodic Few-shot Pseudo Incremental Learning pipeline



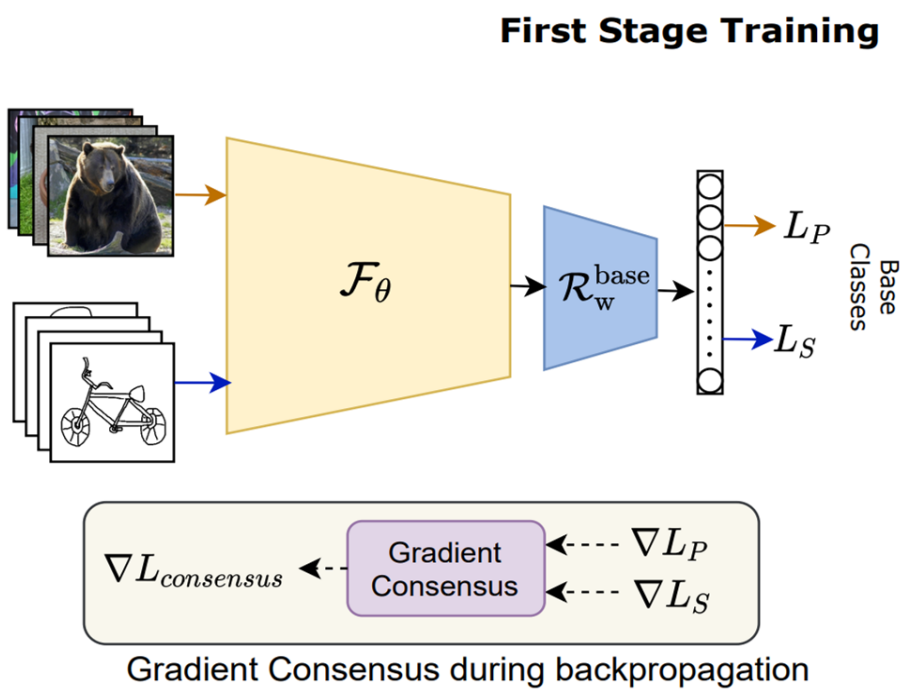


Sketch-Based Incremental Learning

Stage I: Cross modal pre-training for base classes

- Conflicting gradients from different domains – sketch and photo.
- Gradient consensus (δ): Agreement in the gradient space between two domains.

$$\mathcal{L}_{total} = \frac{1}{b}\sum_{(p,y)\sim\mathcal{D}^P_{base}} \mathcal{L}_P(p,y) + \frac{1}{b}\sum_{(s,y)\sim\mathcal{D}^S_{base}} \mathcal{L}_S(s,y)$$



$$\delta(\nabla L^n_P, \nabla L^n_S) = \begin{cases} 1, & \text{sig}(\nabla L^n_P) = \text{sig}(\nabla L^n_S) \\ 0, & \text{otherwise} \end{cases}$$

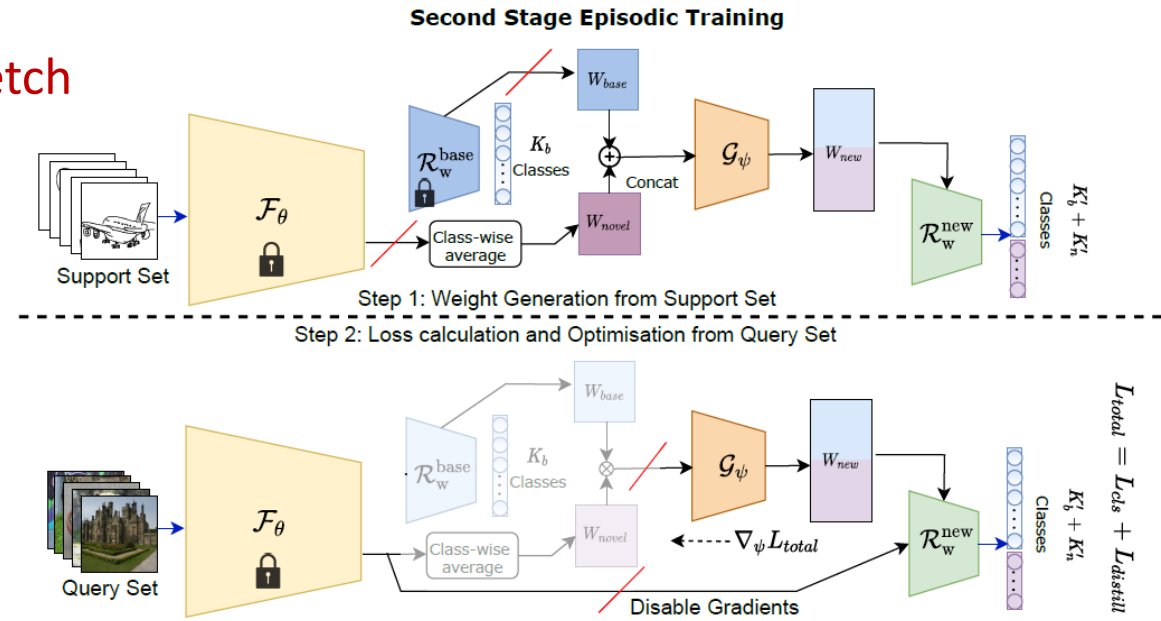
$$\nabla L^n_{consensus} = \begin{cases} \nabla L^n_P + \nabla L^n_S, & \text{if} \quad \delta^n = 1 \\ 0, & \text{if} \quad \delta^n = 0 \end{cases}$$

Sketch-Based Incremental Learning

Stage II: Few shot classifier weight generation

- Objectives:
  - Learn the knowledge of novel classes from fewer sketch exemplars, while classifying photos of novel classes through cross-modal generalization
  - Avoid catastrophic forgetting of the base classes

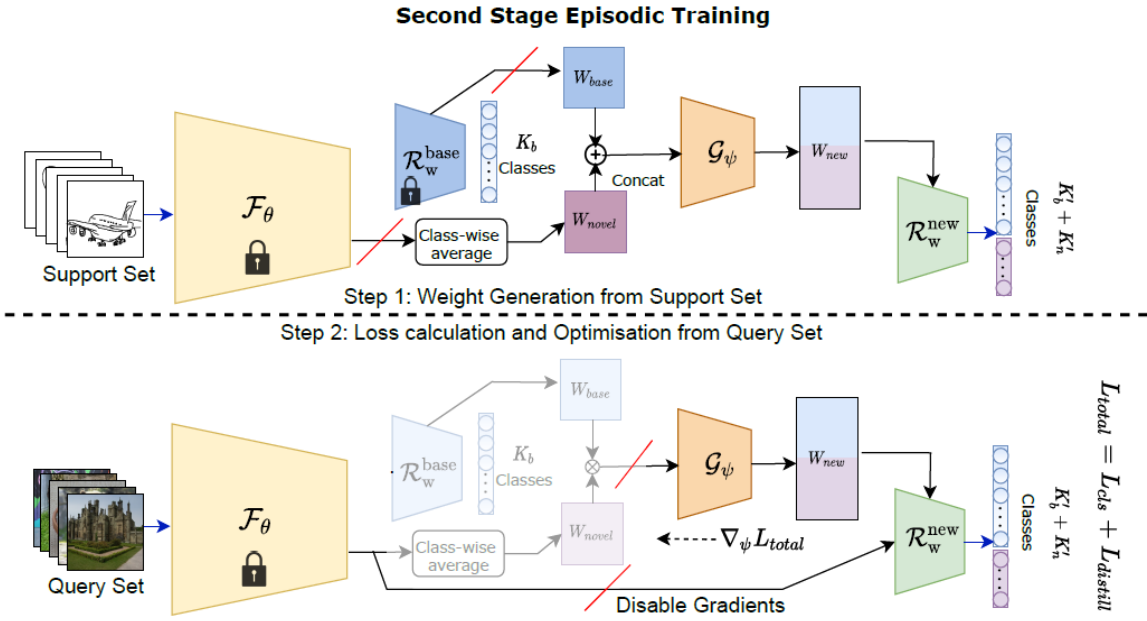
- Two Steps:
  - Weight generation using support set
    - $W_{new} \leftarrow Input$: $W_{base} + sketch - exmeplars$
  - Loss calculation on query set
    - W_{new} is used to classify query set photos,
    - utilized to optimize the weight generation module



Sketch-Based Incremental Learning

Stage II: Few shot classifier weight generation

- Graph Attention Network for *message passing between base and novel classes*.
- Episodically construct pseudo incremental task based only on the base classes
- ***Some base-class vectors are dropped*** *and* ***treated as pseudo-novel classes*** *for episodic training*



$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{distil}$$

$$\mathcal{L}_{cls} = \frac{1}{|\mathcal{Q}|} \sum_{(p,y)\sim\mathcal{Q}} \mathcal{H}(\mathcal{R}_w^{new}(\mathcal{F}_\theta(p)), y)$$

$$\mathcal{L}_{distil} = \frac{1}{|\mathcal{Q}|} \sum_{(p,y)\sim\mathcal{Q}} \mathcal{H}(\mathcal{R}_w^{new}(\mathcal{F}_\theta(p)), \mathcal{R}_w^{base}(\mathcal{F}_\theta(p)))$$

Sketch-Based Incremental Learning

Experiments

- **Dataset used:** Sketchy Extended[1]

- **Competitors:**
  - B1 – Uses a combination of old base and new-novel classes to retrain the complete model.
  - B2 – Only fine-tunes the model using the novel classes.
  - B3 – Removes the GAT module from our proposed framework.
  - B4 – Trains feature extractor (F_θ) along with the weight generator (G_ψ).
  - B5 – Real images are used as the support set during inference, serving as an upper bound.
  - Naively adopt existing FSCIL methods ([2], [3], [4]) under our sketch-based FSCIL setup.

- **Evaluation:**
  - Acc@base – base categories only, judging *catastrophic forgetting*.
  - Acc@novel – novel categories only, judging a model's ability to learn new classes as well as its *generalizing potential on cross-domain* data.
  - Acc@both – both base and novel categories, judging how the base classes' knowledge affects novel classes and vice-versa.

[1] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In CVPR, 2017.
[2] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. In NeurIPS, 2017.
[3] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class incremental learning. In CVPR, 2020.
[4] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In CVPR, 2018.

Sketch-Based Incremental Learning

Quantitative Results:

Methods		5-Shot Learning			1-Shot Learning		
		Acc@both	Acc@base	Acc@novel	Acc@both	Acc@base	Acc@novel
Baselines	$B1$	36.29 %	73.94%	38.92%	31.52%	73.98%	34.68%
	$B2$	25.86%	32.85%	70.58%	28.81%	40.91%	50.24%
	$B3$	58.92%	73.81%	72.34%	53.35%	73.75%	59.93%
	$B4$	54.5%	71.68%	71.81%	51.41%	71.68%	51.44%
	$B5^*$	71.52%	75.72%	85.46%	63.47%	75.83%	73.90%
SOTA FSCIL	[17]	50.45%	74.35%	65.81%	44.71%	73.98%	64.21%
	[50]	45.25%	74.10%	63.46%	41.97%	74.60%	61.85%
	[54]	51.54%	73.21%	66.82%	45.81%	73.58%	63.95%
Ours	DIY-FSCIL	60.54%	74.38%	75.84%	54.97 %	74.06%	64.10%

Average classification accuracy of DIY-FSCIL framework using our self-designed baselines and adopted SOTA FSCIL [1]. B5* is an upper bound.

[1] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In CVPR, 2018.
[2] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. In NeurIPS, 2017.
[3] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class incremental learning. In CVPR, 2020.

Sketch-Based Incremental Learning

Ablative studies

GAT – Graph Attention Network
GC – Gradient Consensus
KD – Knowledge Distillation Loss
CMT – Cross-Modal Training

GAT	GC	KD	CMT		Metrics		
					Acc@both	Acc@base	Acc@novel
✔	✔	✔	✔	5−shot	60.54%	74.38%	75.84%
				1−shot	54.97%	74.06%	64.10%
✗	✔	✔	✔	5−shot	58.92%	73.81%	72.34%
				1−shot	53.35%	73.75%	59.93%
✗	✗	✔	✔	5−shot	58.47%	73.96%	71.67%
				1−shot	53.22%	73.67%	59.46%
✗	✗	✗	✔	5−shot	57.47%	70.96%	69.67%
				1−shot	51.22%	71.67%	57.46%
✗	✗	✗	✗	5−shot	35.19%	62.98%	40.52%
				1−shot	27.67%	61.72%	32.83%

		Metrics		
		Acc@both	Acc@base	Acc@novel
5− way	1−shot	54.97%	74.06%	64.10%
	5−shot	60.54%	74.38%	75.84%
	10−shot	61.61%	74.14%	76.95%
	15−shot	62.08%	73.95%	77.48%
	20−shot	62.35%	74.83%	78.35%
10− way	1−shot	43.62%	73.24%	47.31%
	5−shot	51.82%	73.37%	59.97%
	10−shot	53.75%	73.54%	61.21%
	15−shot	55.46%	73.38%	62.74%
	20−shot	57.58%	73.23%	64.37%

Performance with varying n-way/k-shot evaluation

Sketch-Based Incremental Learning

Ablative studies

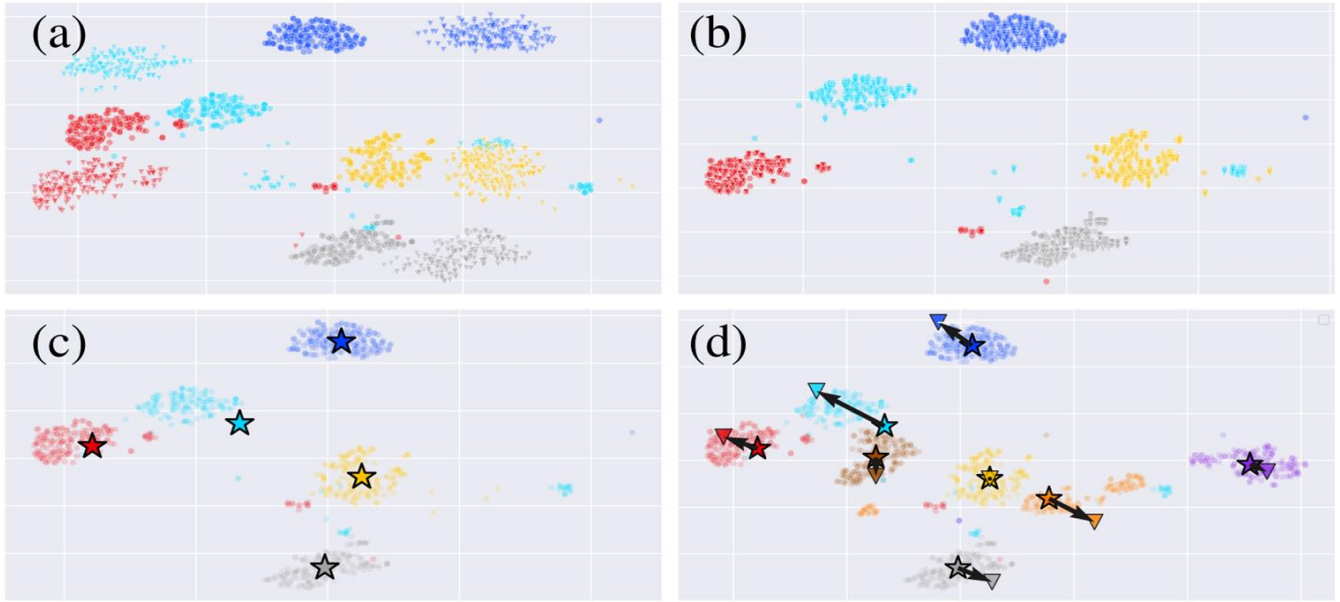
t-SNE plots showing:

a. Photos and sketches on shared latent space for a naïve baseline.

b. Same for our framework.

c. Base classes.

d. Shifting classifier weights to produce better decision boundaries using DIY-FSCIL.



Comparative study between *sketch* vs *text* for support set

	One-shot learning		
	Acc@both	Acc@base	Acc@novel
Text (Word2Vec)	22.85%	73.98%	26.15%
Text (GloVe)	22.80%	74.04%	26.85%
Sketch (Ours)	54.97%	74.06%	64.10%

Sketch-Based Incremental Learning

TL;DR:

- Sketch a new modality for few-shot class incremental learning
- Addresses the ethical/privacy concerns while collecting photos.
- Gradient consensus + Graph attention networks + Knowledge distillation loss.

Impact and Future Directions:

Specific -Sketch Community

- Few-shot sketch-based dynamic hierarchical classification.

Broad Computer Vision Community

- Evaluation paradigm to check the generalizability of few shot frameworks.
- Three dimensions of computer vising literature into a single pipeline – few-shot learning, domain generalization, and incremental learning.

Sketch-Based Incremental Learning

Summary

- On-the-fly FG-SBIR with RL based optimization
- Photo2Sketch as a conjugate task for semi-supervised FG-SBIR
- Stroke-subset selector for Noise-Tolerant FG-SBIR
- Dual modality of sketches for self-supervised learning
- Sketch for few-shot class-incremental learning

Sketch-Based Incremental Learning

Future Sketch Research Directions: I

❖ Sketch + Pretrained Large-Scale Models

1. Sketch + Pretrained StyleGAN [A] for high quality fine-grained content generation
2. Sketch + Pretrained StyleGAN for fine-grained retrieval
3. Sketch + Pretrained CLIP [B] for zero-shot SBIR
4. Sketch + Pretrained CLIP + Prompt Learning for zero/few-shot any hand-drawn drawing recognition

[A] Analyzing and improving the image quality of stylegan. In CVPR, 2021.
[B] Learning Transferable Visual Models From Natural Language Supervision, In ICML, 2021.

Sketch-Based Incremental Learning

Future Sketch Research Directions: II

❖ Analysing Sketch Further

1. Adversarial attacks for sparse sketches (for both vector and raster domain)
2. Explainablity for Sketches (for both vector and raster domain)
3. 2D Sketch to 3D lifting using NERF
4. Sketch is Salient: Sketch as a label for saliency detection

Sketch-Based Incremental Learning

Future Sketch Research Directions: III

❖ Exploiting Fine-grained Potential Further

1. Sketch for dynamic hierarchical classification [e. g. make 10^{th} class as 10A, 10B and 10-rest]
2. Weakly-supervised Sketch-based 3D shape retrieval via pivot learning
3. Sketch for zero-shot fine-grained object detection
4. Sketch for fine-grained caricature generation.

Thanks for your attention!