

Fine-grained Visual Analysis: From Classification to Retrieval

Yi-Zhe Song

SketchX Lab, CVSSP, University of Surrey, UK

<http://sketchx.ai>

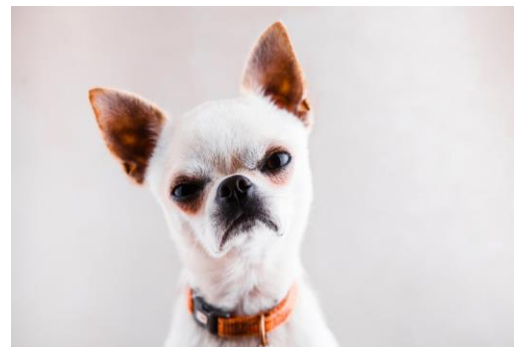
Why fine-grained?



Dog



Dog



Dog

I am not just a "dog" 😞 😞 😞

Why fine-grained?



Husky



Chihuahua



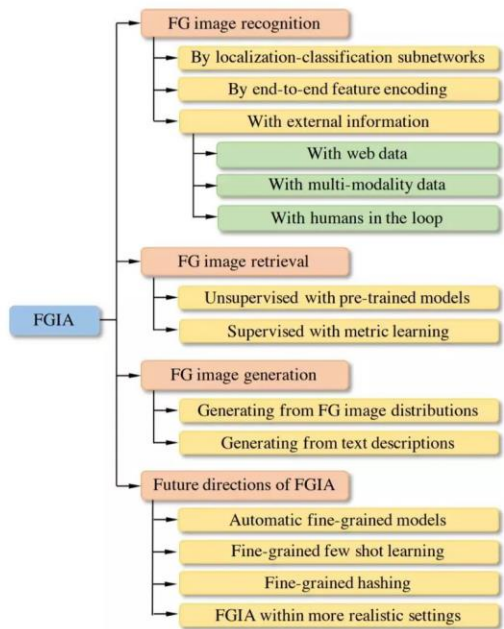
Bulldog

Better 😊

At the very heart of *human and computer vision*!!

What is fine-grained?

- Surveys + Seminars exist
 - a good survey [1]
 - First Edition of 见微知著 (2019年12月11日)
- Classification + Retrieval most studied
 - Classification being the favourite child
 - Images → video, 3D, text
 - Recent branching to generation, transfer learning, hashing...

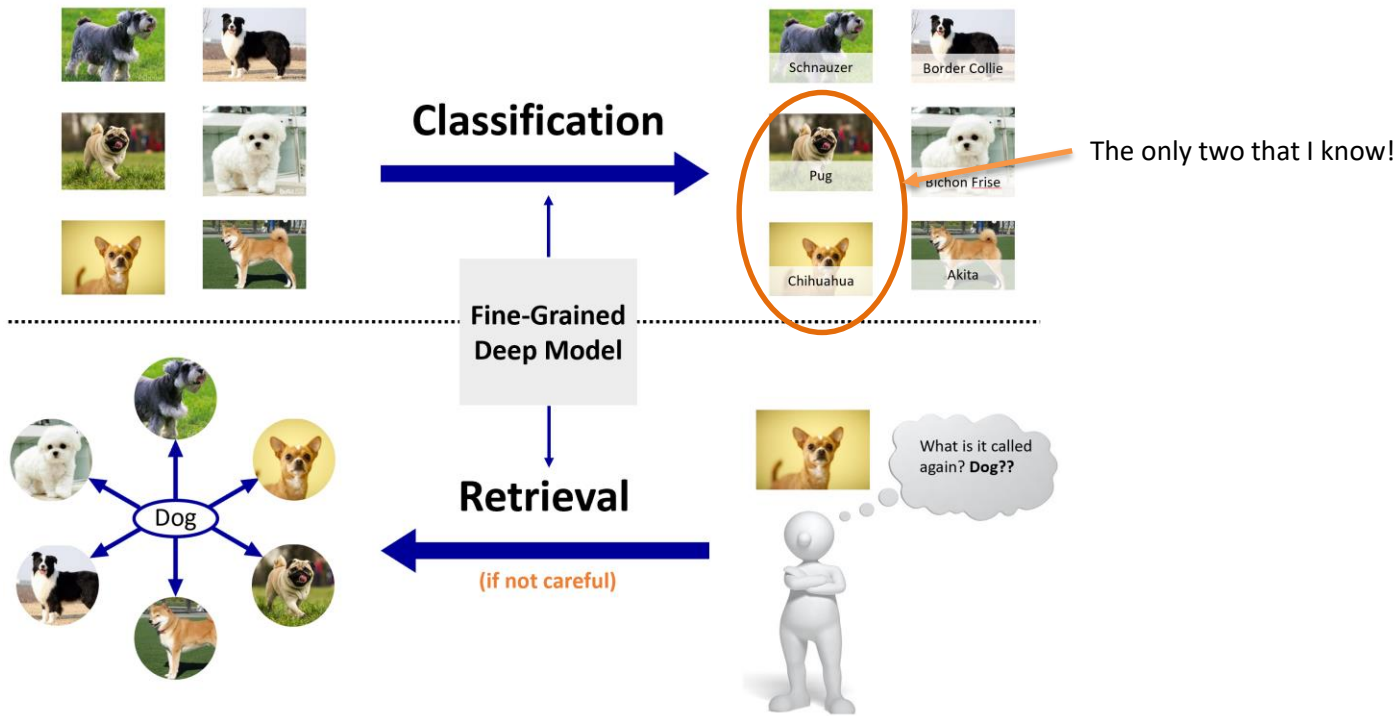


[1]

Classification vs. Retrieval

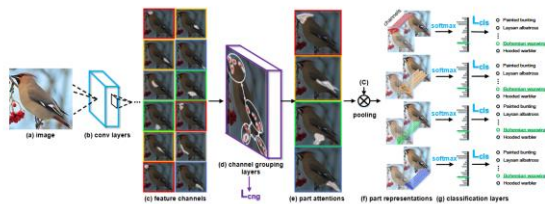
- “The Curse of the Labels”

- **Classification** → hard to obtain expert labels
- **Retrieval** → one can not retrieve without knowing the label

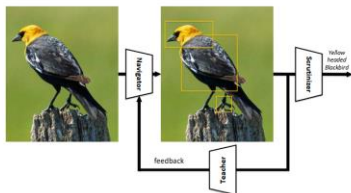


Problem with Classification

- Dataset! Dataset! Dataset! → Label! Label! Label!
- Obsession with parts
 - **Explicit** to start with
 - Now **implicit** as well → part is not everything

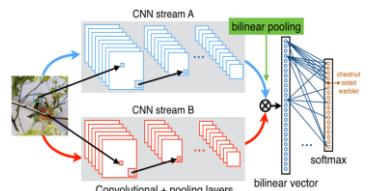


MA-CNN (ICCV17)

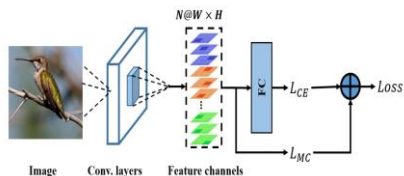


NTS-Net (ECCV18)

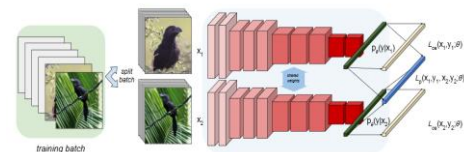
Explicit Models



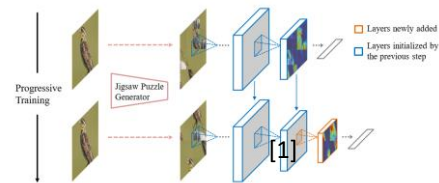
B-CNN (ICCV15)



MC-Loss (TIP20)



Pairwise confusion (ECCV18)



PMG (ECCV20)

Implicit Models

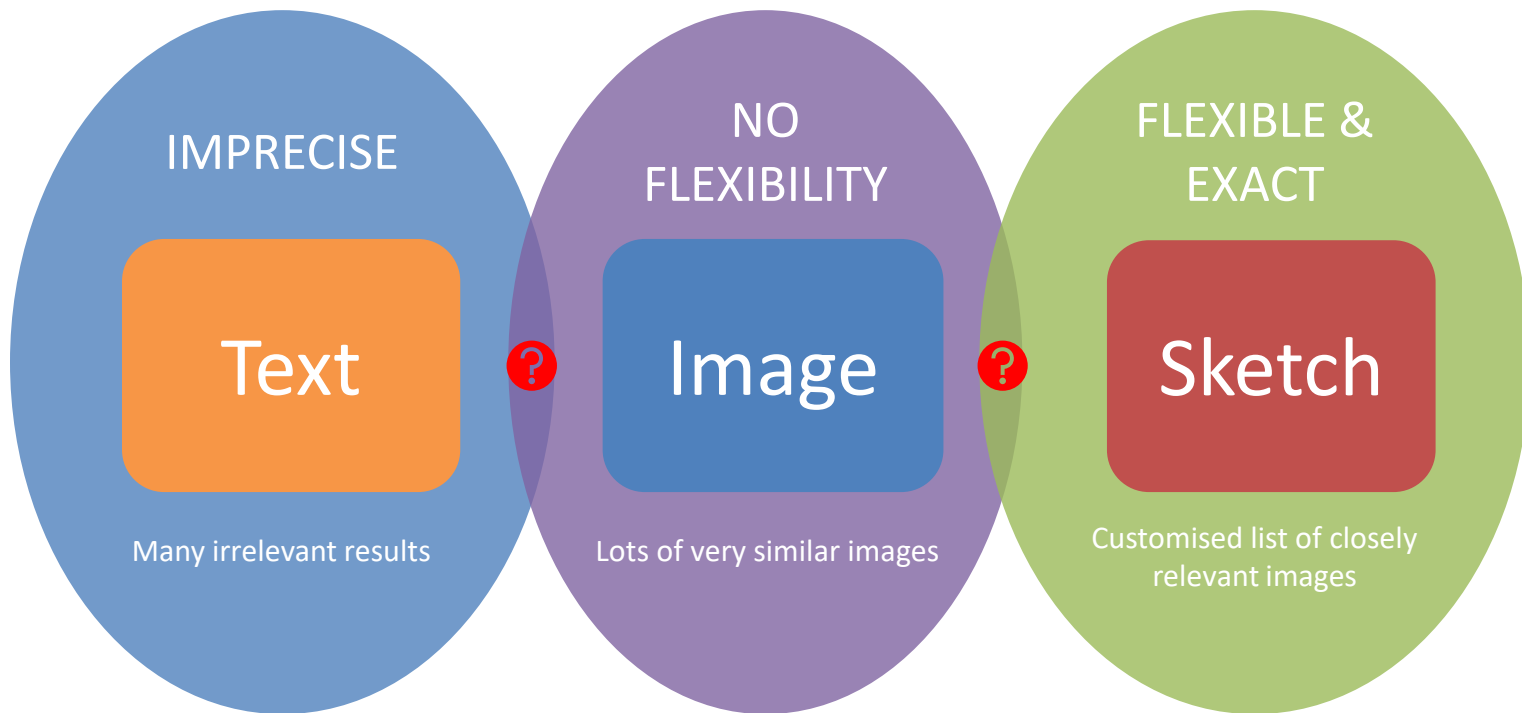
Problem with Retrieval

- Ill-posed to start with → where do we get the labels?
 - Retrieval **dictates expert knowledge** to start with!
- Best input modality?
 - Yes, there is image (but is it the only choice?)
 - Human subjectivity → text best for that (?)
- There is just not enough work!

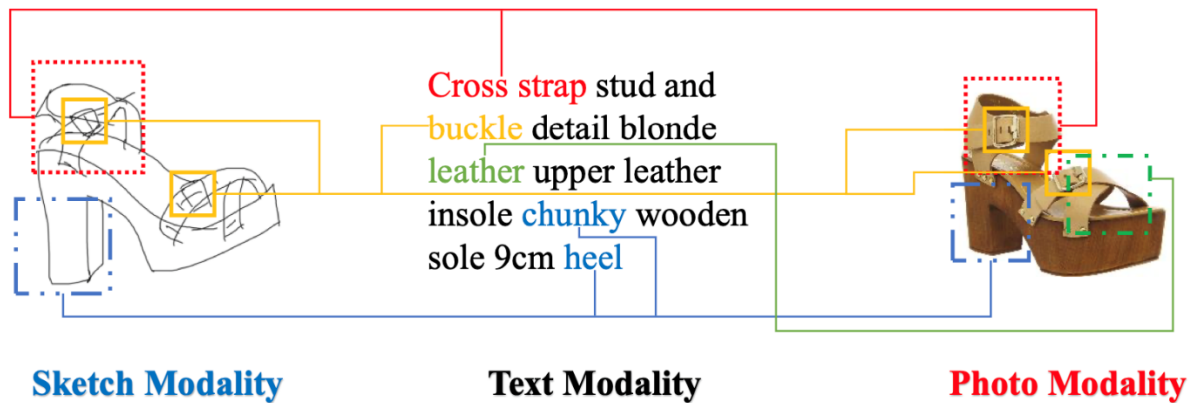
All about Retrieval

- Is the old “fine-grained” enough? → more than just names (labels)!
 - Pose, instance-level details
 - “a Labrador standing on two feet, looking at the camera with a smile”
 - Latent sub-classes
 - Labrador → English Labrador and American Labrador
- Flexibility to meet **human subjectivity**
 - as flexible as text?
- What would be the **best input modality**?
- **More practical** with real application scenarios?

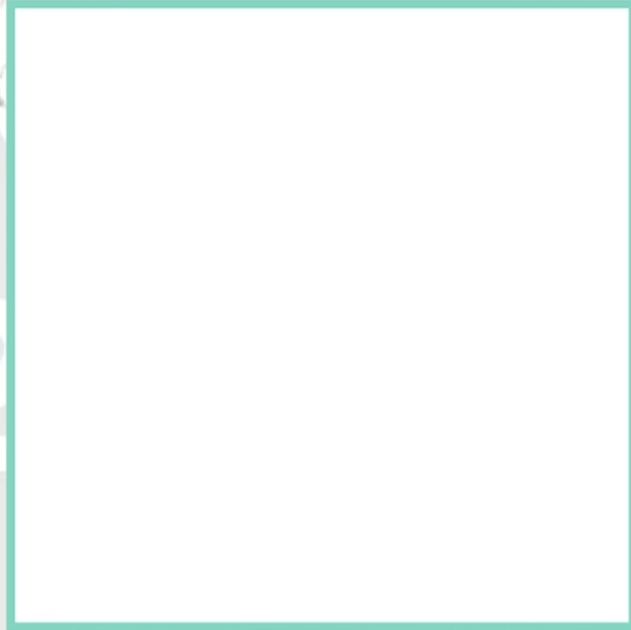
Sketch for Retrieval



Sketch for Retrieval

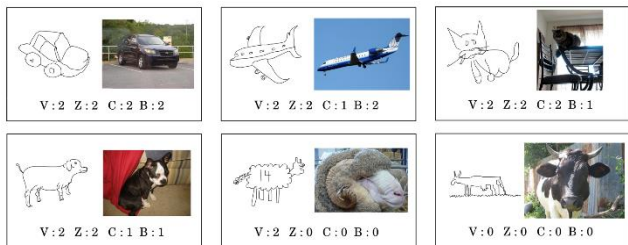


- Specific challenges
 - Cross-modal
 - Human subjectivity
 - Learning under small data



FG-SBIR: Fine-Grained Sketch-Based Image Retrieval

FG-SBIR 1.0 – pose correspondence (BMVC'15)



FG-SBIR 2.0 – instance correspondence (CVPR'16 Oral, SIGGRAPH'16, ICCV'17, 3xECCV'18, CVPR'19 Oral, CVPR'20)

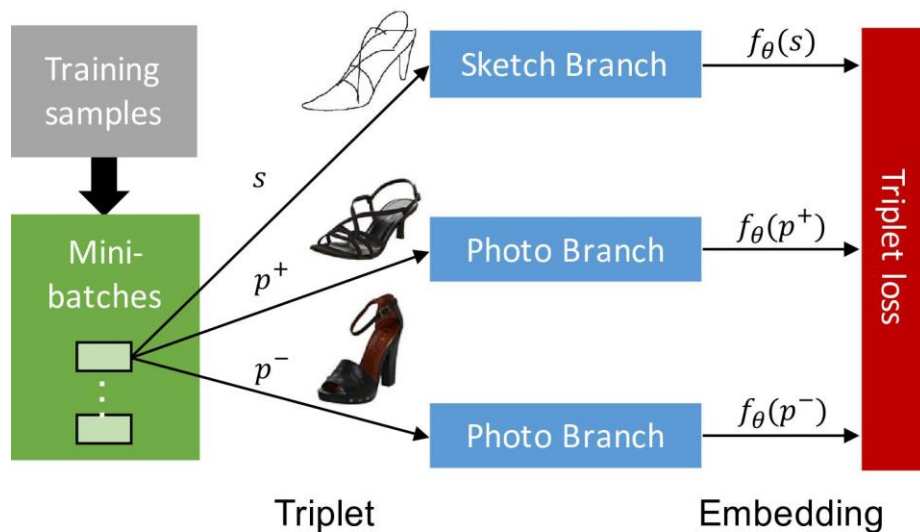


FG-SBIR 3.0 – on-the-fly retrieval (CVPR'20 Oral)



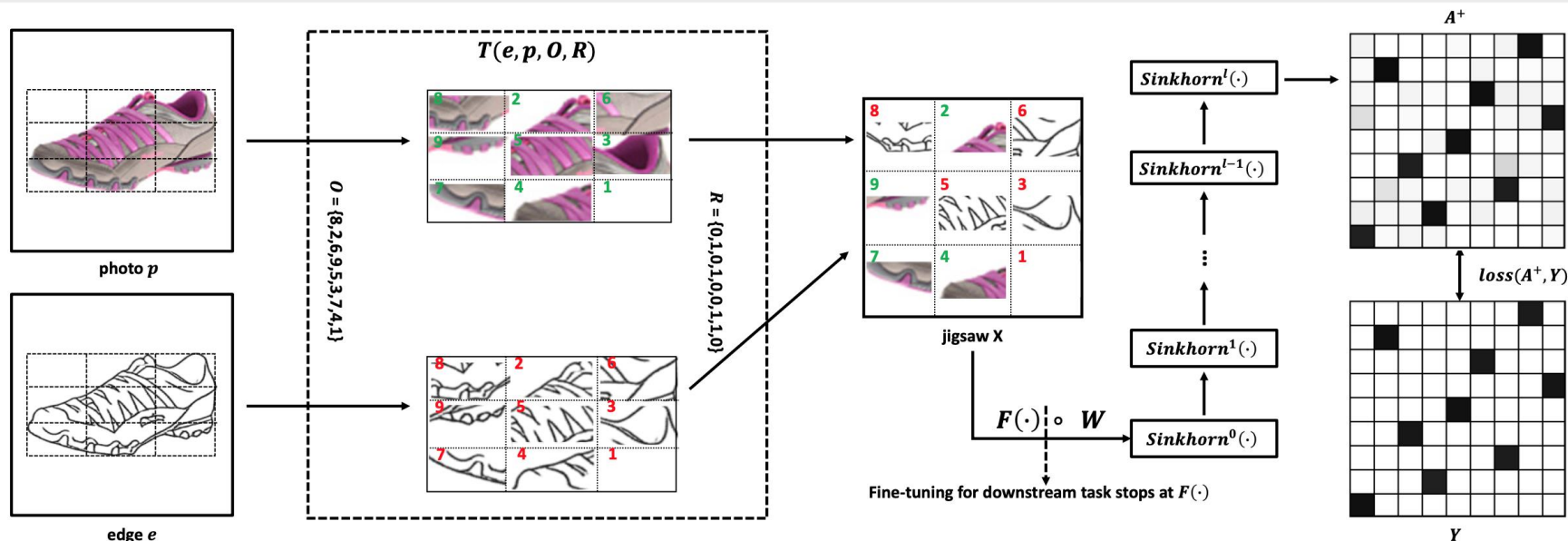
FG-SBIR: Fine-Grained Sketch-Based Image Retrieval

- Dataset usually very small
 - ImageNet pre-training is thus a must + fine-tuning.
- Triplet Ranking Network
 - pushing positive sketch-photo pairs near, and negatives apart.



FG-SBIR: The Role of Jigsaw

- Jigsaw puzzles helps with fine-grained [1]
 - See also [2] for classification



[1] Kaiyue Pang, Yongxin Yang, Timothy Hospedales, Tao Xiang, Yi-Zhe Song, *Solving Mixed-modal Jigsaw Puzzle for Fine-Grained Sketch-Based Image Retrieval*, CVPR 2020

[2] Ruoyi Du, Dongliang Chang, Ayan Kumar Bhunia, Jiyang Xie, Yi-Zhe Song, Zhanyu Ma, Jun Guo. *Fine-Grained Visual Classification via Progressive Multi-Granularity Training of Jigsaw Patches*, ECCV 2020

FG-SBIR: The Role of Jigsaw

- Solving a mixed-modality jigsaw model requires learning to:
 - Bridge the domain discrepancy
 - Understand holistic object configuration
 - Encode fine-grained detail.
- A permutation inference problem
 - Normalisation via Sinkhorn iterations
- Great performance boost to long standing practice of ImageNet pre-training.

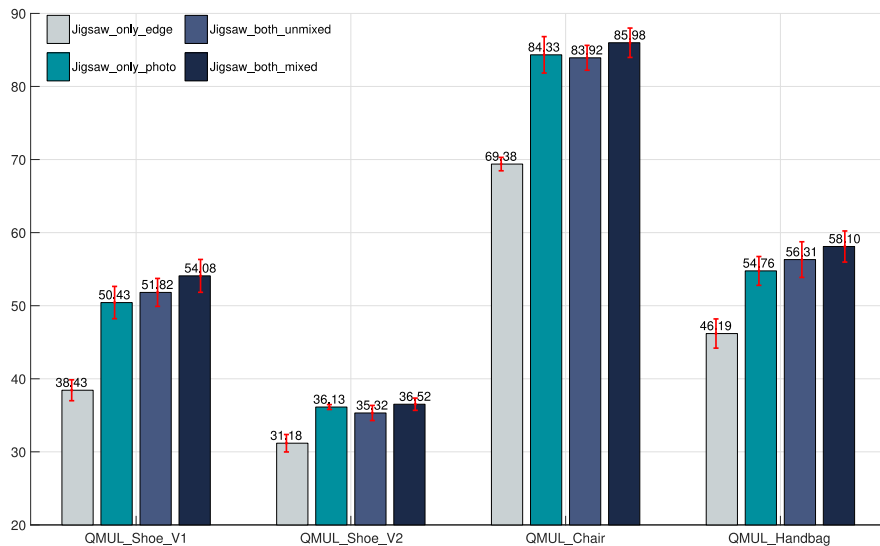
FG-SBIR: The Role of Jigsaw

Pre-training		FG-SBIR Dataset			
Method	Self-supervised?	QMUL_Shoe_V1 ^{4×4}	QMUL_Shoe_V2 ^{3×3}	QMUL_Chair ^{3×3}	QMUL_Handbag ^{4×4}
Counting [16]	✓	41.74%± 2.30	30.42%± 0.54	72.78%± 4.35	54.05%± 2.77
Rotation [9]	✓	32.17%± 2.68	28.83%± 0.40	70.31%± 3.45	38.33%± 1.86
CPC [17]	✓	21.91%± 1.69	8.65%± 0.34	35.24%± 0.42	15.36%± 0.69
Matching [20]	✓	39.13%± 0.87	31.05%± 0.84	75.69%± 1.53	50.36%± 0.68
ImageNet [29]	✗	43.48%± 1.74	33.99%± 1.09	85.16%± 1.56	52.62%± 2.04
Ours/1000-way	✓	42.78%± 3.75	30.24%± 1.74	79.59%± 1.53	49.40%± 3.97
Ours/ImageNet	✗✓	48.00%± 2.91	31.26%± 0.65	79.59%± 1.34	61.07%± 1.50
Ours	✓	56.52%± 2.75	36.52%± 0.84	85.98%± 2.01	62.97%± 2.04

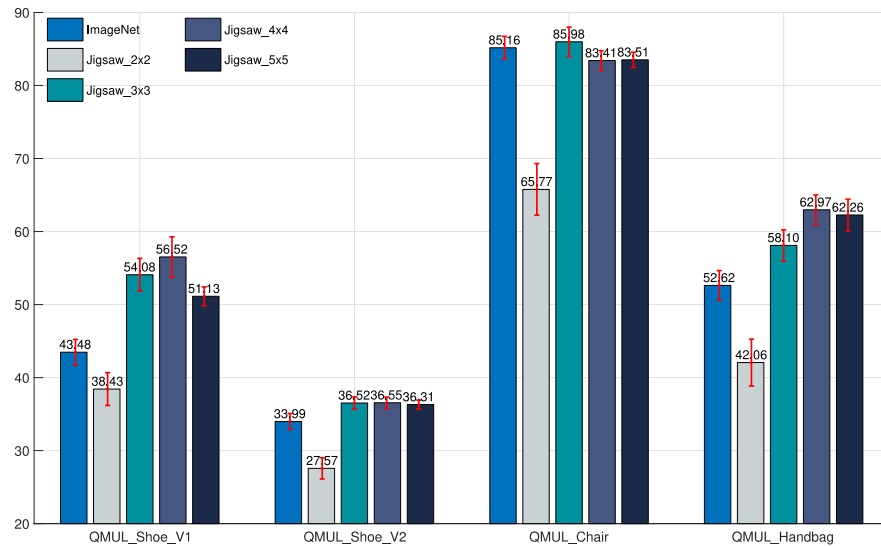
NOTE: **opposite conclusions** for category-level task!

FG-SBIR: The Role of Jigsaw

Effect of jigsaw modality



Effect of jigsaw granularity



- mixed-modal Jigsaw is the best
- granularity of jigsaw not crucial

FG-SBIR: On-the-Fly



Problem – “I can’t sketch”

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

FG-SBIR: On-the-Fly

Old Setup: sketch first, *then* retrieve



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



Less is more!

FG-SBIR: On-the-Fly

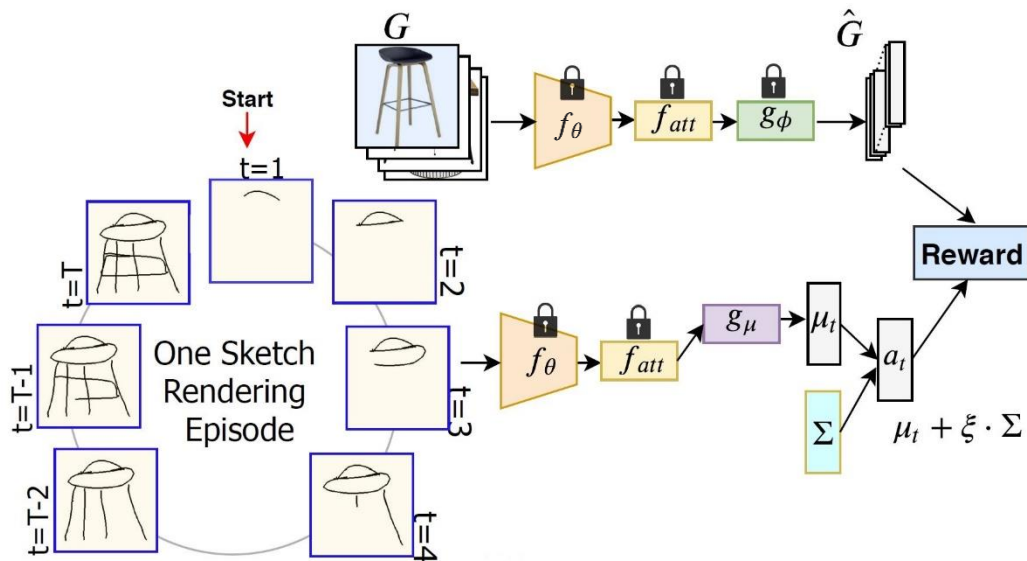
- **Natural:** incomplete sketches can *already* retrieve!
- **Faster:** *no need* to sketch the whole thing
- **More accurate:** modelling the *sketching process* does help

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



FG-SBIR: On-the-Fly

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

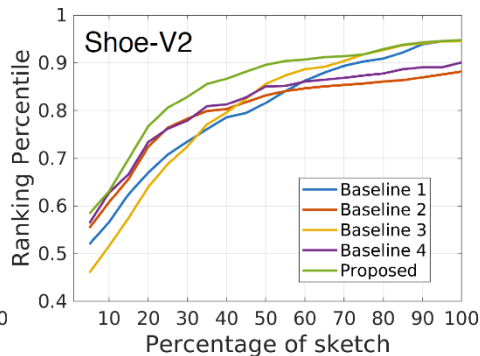
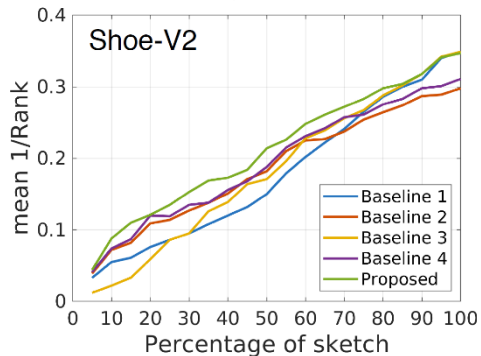


FG-SBIR: On-the-Fly

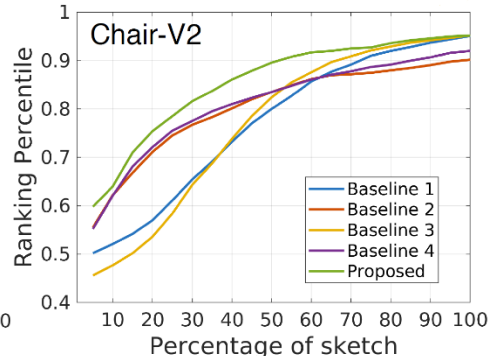
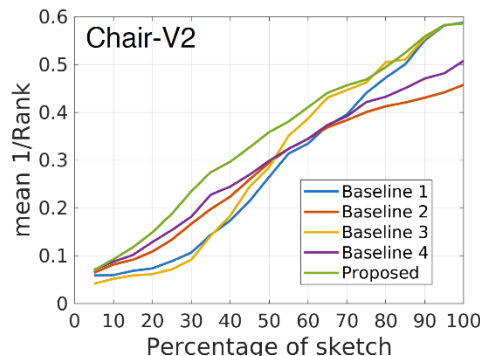
Quantitative Results vs Different Baselines (A@q, m@A, and m@B)

	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)

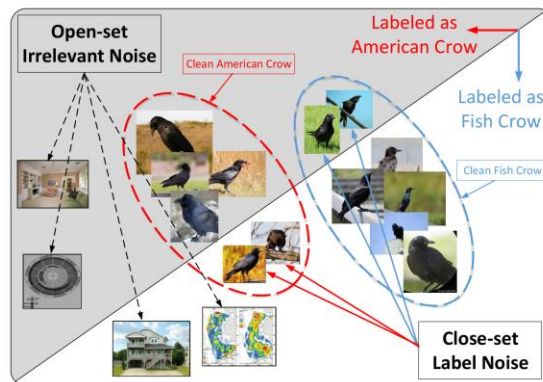


Percentage-wise Results for Chair-V2 (m@A, and m@B)



Classification ⊗ Retrieval

- Classification → Retrieval
 - Obvious
- Retrieval → Classification
 - Cure for web data?
 - Sub-class discovery?



[1]



Conclusion

- Fine-grained is important!
- Classification bottlenecked
- Retrieval needs more work
 - Unique challenges
 - Practical applications
 - Can help classification
- Beyond 2D!

