Paper Tag: **TUE-AM-263**

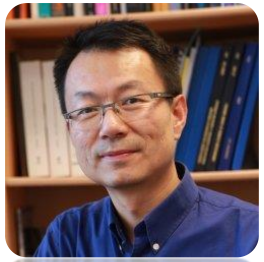
CLIP for All Things Zero-Shot Sketch-Based Image Retrieval, Fine-Grained or Not


Aneeshan Sain [1,2]


Ayan Kumar Bhunia [1]


Pinaki Nath Chowdhury [a,b]


Subhadeep Koley[a,b]

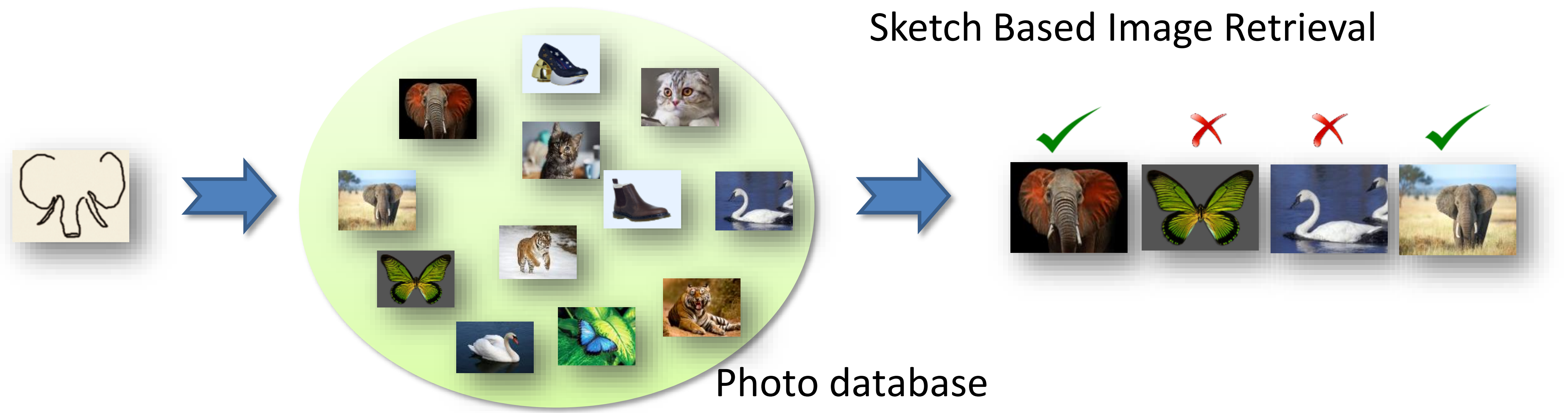

Tao Xiang [1,2]


Yi-Zhe Song [1,2]

[1] SketchX Lab, CVSSP, University of Surrey, UK

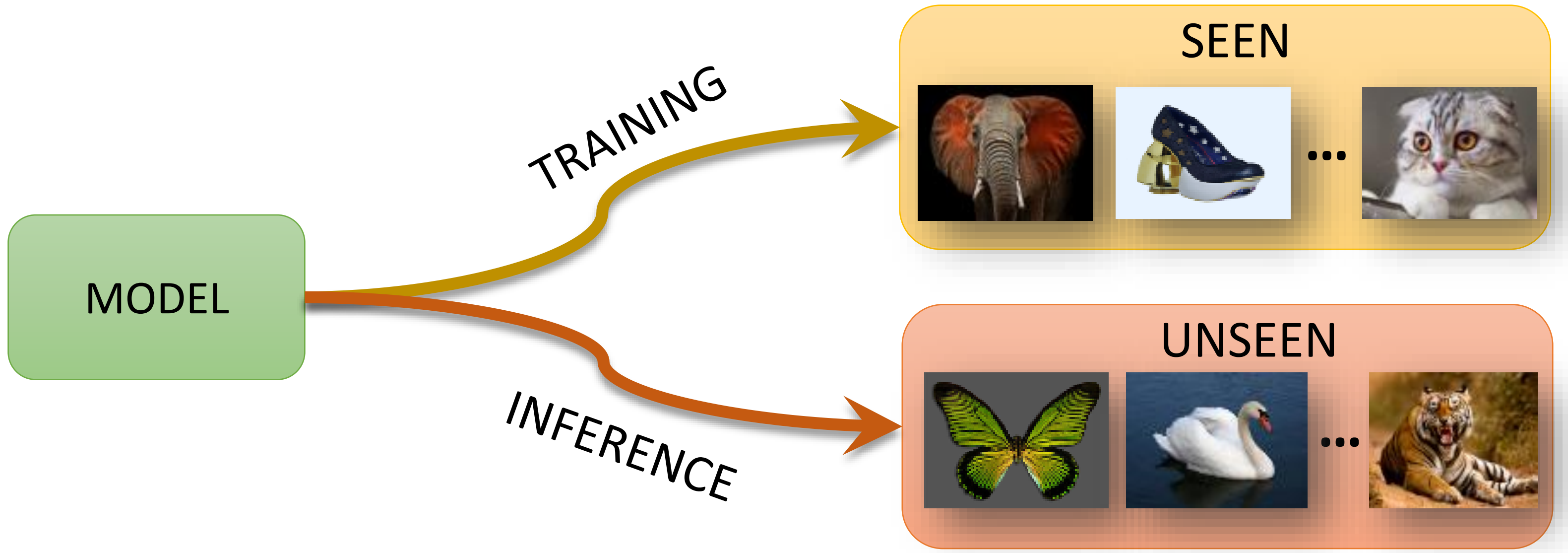
[2] iFlyTek-Surrey Joint Research Centre on Artificial Intelligence

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA



Zero-Shot Learning Setup

- Trained on **seen** classes .
- Evaluated on **unseen** classes .
- No additional data .

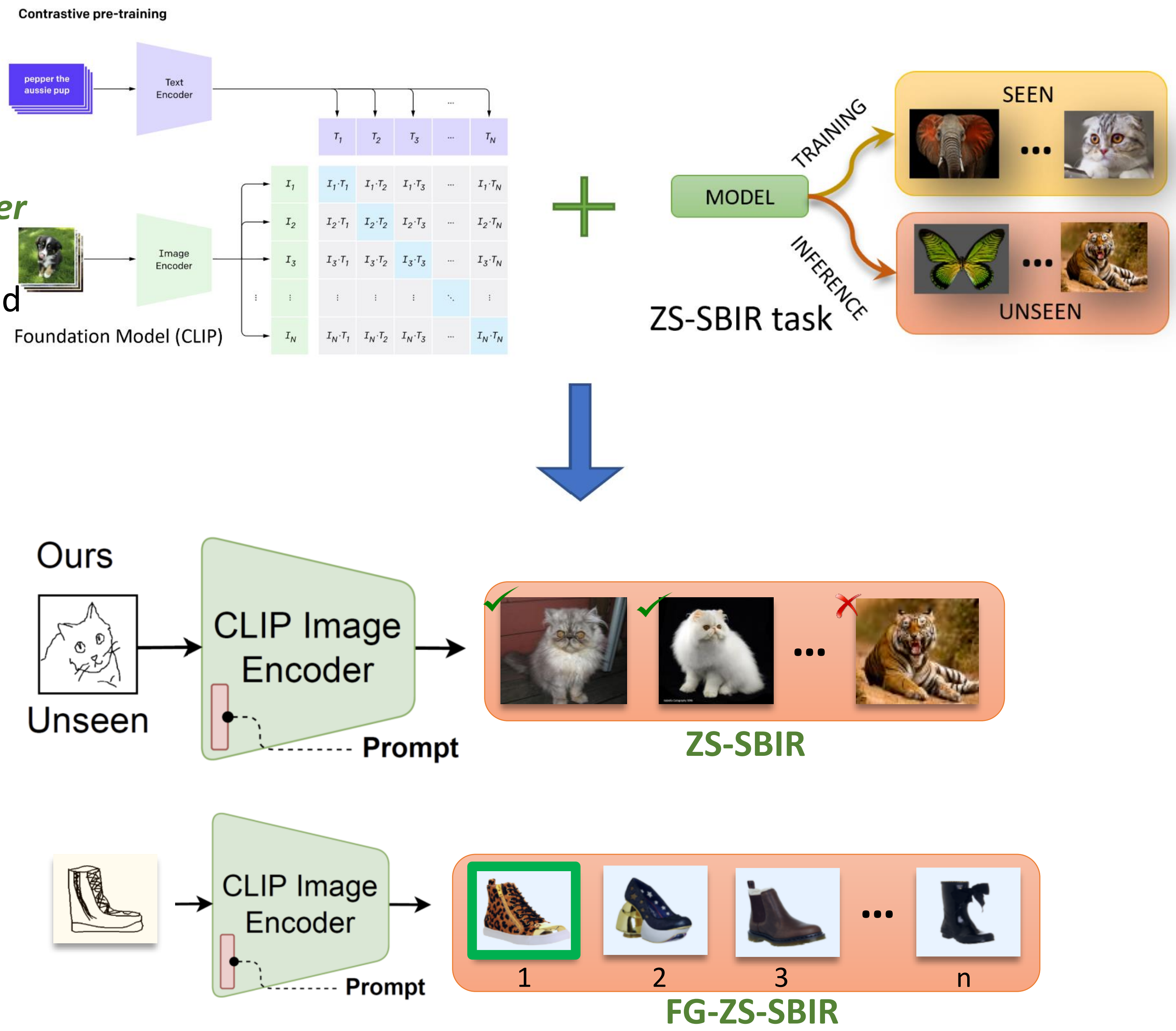


Motivation and objective

Overview:
- Cross-category and cross-modal ***Semantic transfer***
- Usage of ***standard word embeddings***.
- ***Synergy*** between foundation models like CLIP and the cross-modal problem of ZS-SBIR.

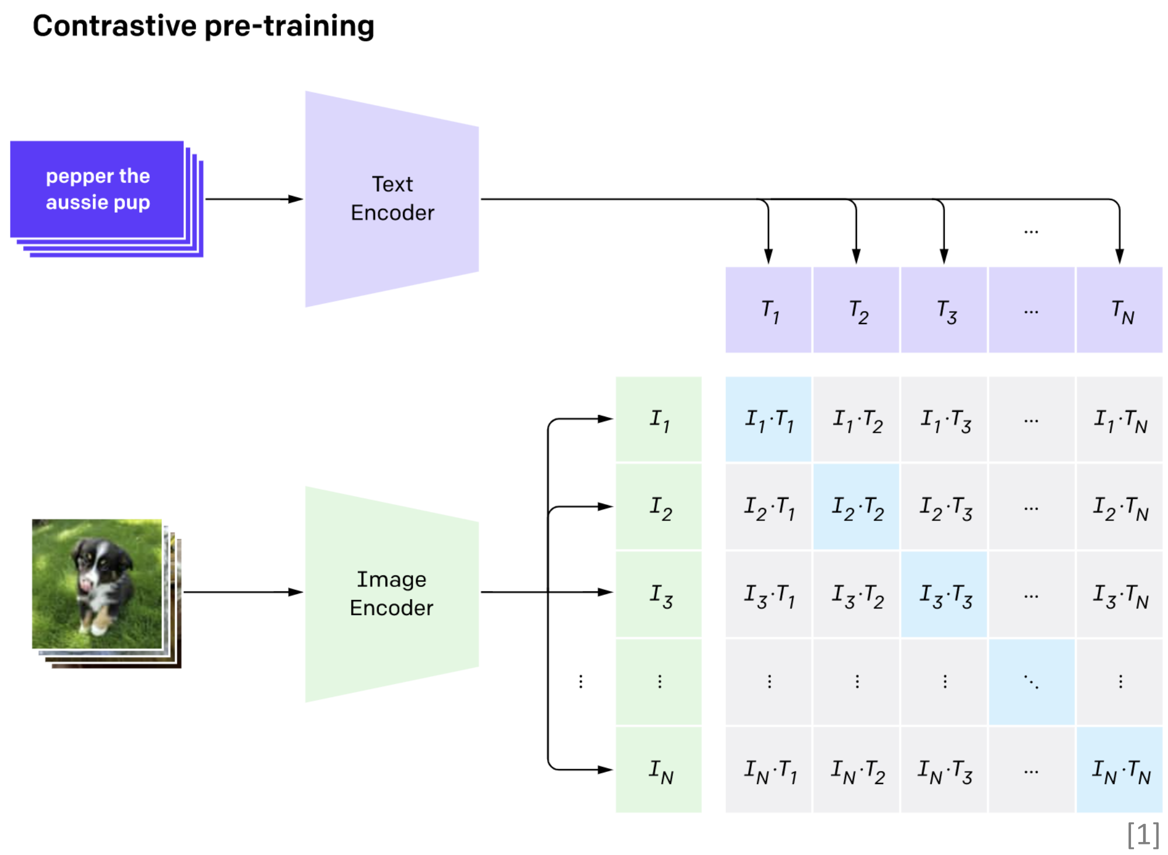
Proposal:
- Foundation models –
  - highly ***enriched*** semantic latent space
  - ***encapsulates*** cross-modal knowledge.
- ***Visual prompts*** **–**
  - **adapts** CLIP to SBIR tasks.
  - ***preserves*** its ***generalizability***.

Further goals:
- Extend to even more difficult setup of ***Fine-grained*** ZS-SBIR.



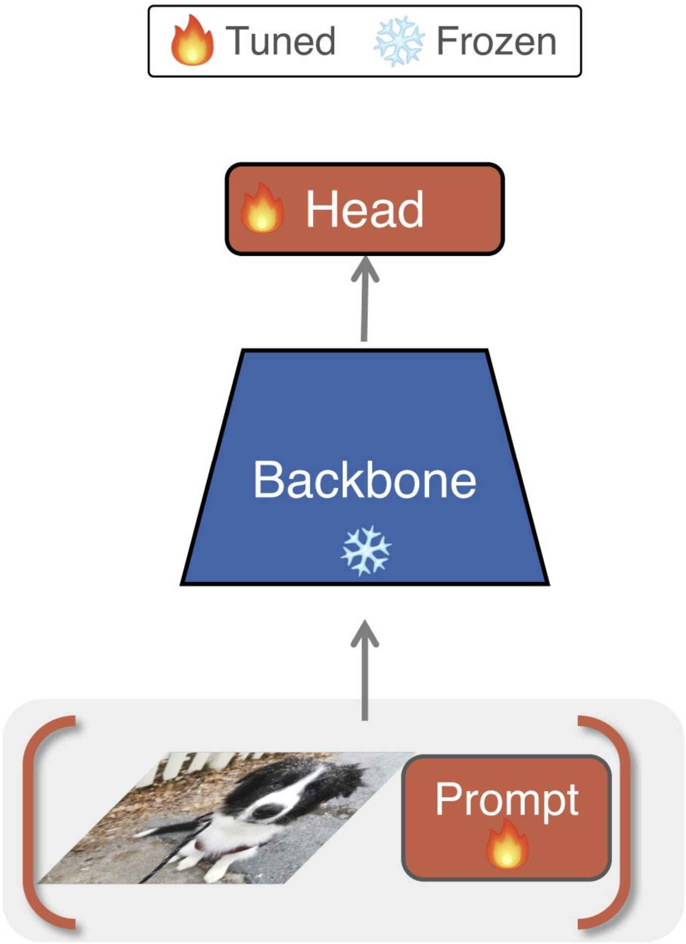
Background

Foundation Model -- CLIP



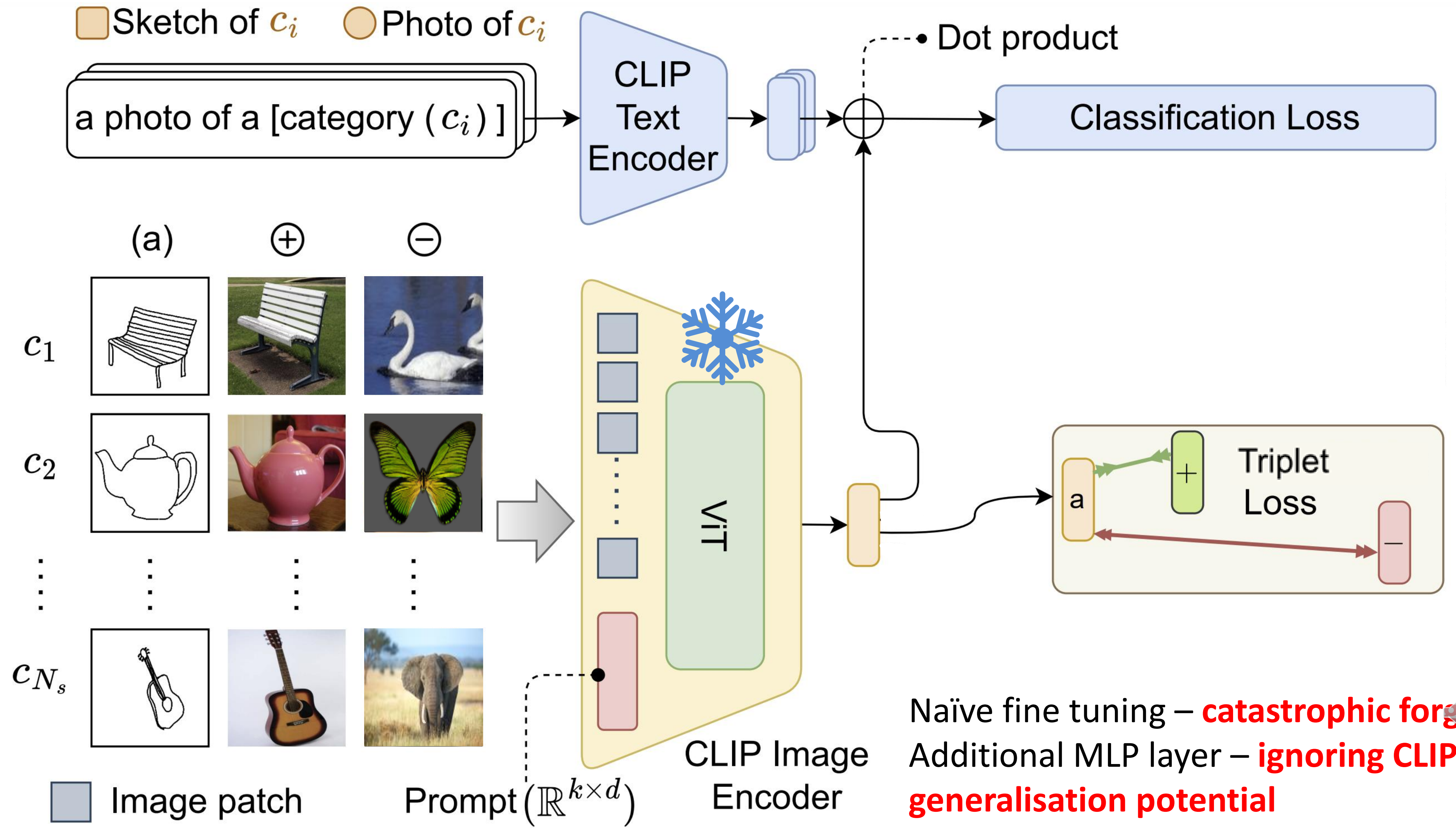
- Maximize cosine similarity for matching pairs, minimise otherwise

Prompt Learning (in our context)

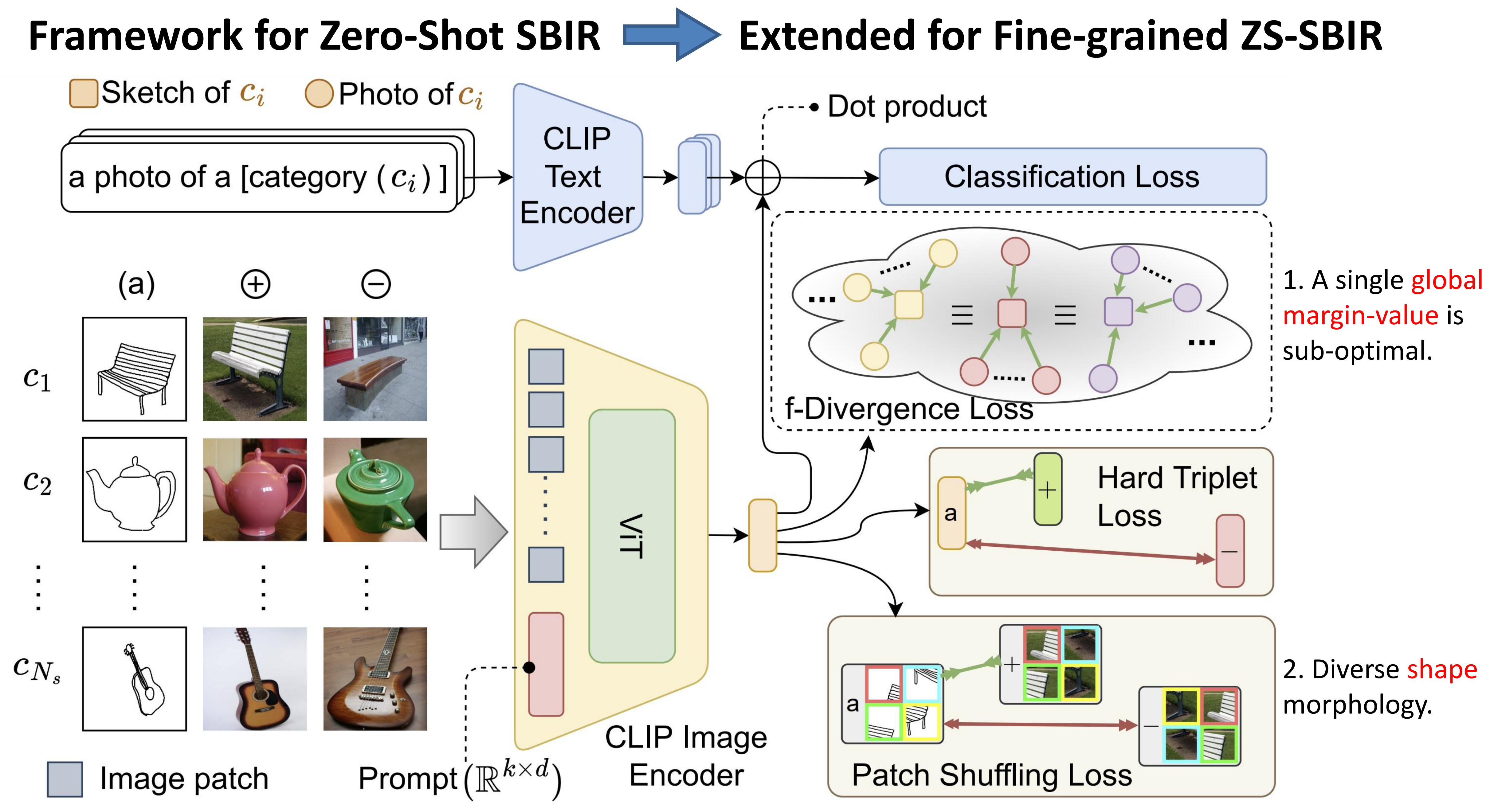


[1] Alec Radford et. al. Learning transferable visual models from natural language supervision. In ICML, 2021.

Framework for Zero-Shot SBIR



Naïve fine tuning – **catastrophic forgetting**
Additional MLP layer – **ignoring CLIP's generalisation potential**


Framework for Zero-Shot SBIR
Extended for Fine-grained ZS-SBIR
Sketch of c_i
Photo of c_i
a photo of a [category (c_i)]
CLIP Text Encoder
Dot product
Classification Loss
f-Divergence Loss
1. A single global margin-value is sub-optimal.
(a)
c_1
c_2
c_{N_s}
ViT
Hard Triplet Loss
a
CLIP Image Encoder
Patch Shuffling Loss
2. Diverse shape morphology.
Image patch
Prompt $(\mathbb{R}^{k\times d})$

Experiments

- **Datasets used:**
  - Sketchy (both basic and Extended)[1] – 73K sketches across 125 categories.
  - TU-Berlin (Extended)[2] – 20K sketches across 250 categories.
  - QuickDraw[3] – We use a subset of 110 categories with 330K sketches and 204K photos.

- **Competitors:**
  - State of the art Zero-shot SBIR (ZS-SBIR) methods.
  - CLIP-based ZS-SBIR baselines.
  - CLIP-based Fine-Grained ZS-SBIR (FG-ZS-SBIR) baselines.

- **Evaluation protocol and metric:**
  - **ZS-SBIR**
    - Mean average precision mAP@All.
    - Precision top 200 retrievals P@200.
  - **Fine-Grained ZS-SBIR**
    - Acc@Q : Percentage of sketches having true-matched photos in the top-Q list.

[1] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. ACM TOG, 2016.
[2] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? ACM TOG, 2012.
[3] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and YiZhe Song. Doodle to search: Practical zero-shot sketchbased image retrieval. In CVPR, 2019.

Quantitative Analysis

Zero-Shot SBIR

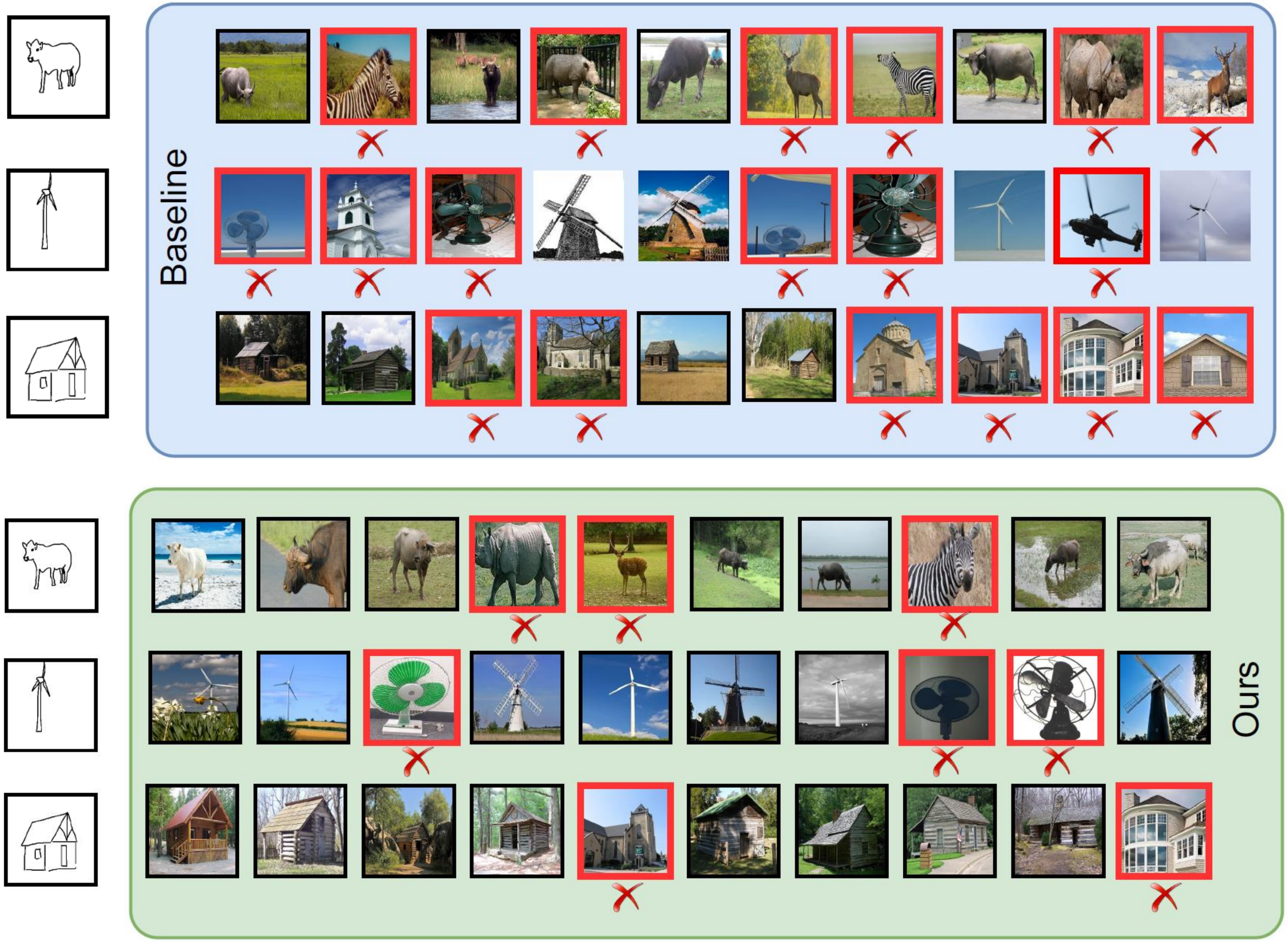
	Venue	Methods	Sketchy mAP@200	Sketchy P@200	TU-Berlin mAP@all	TU-Berlin P@100	QuickDraw mAP@all	QuickDraw P@200
ZS-SOTA	ECCV '18	ZS-CAAE[1]	0.156	0.260	0.005	0.003	–	–
	ECCV '18	ZS-CVAE[1]	0.225	0.333	0.005	0.001	0.003	0.003
	CVPR '19	ZS-CCGAN[2]	–	–	0.297	0.426	–	–
	CVPR '19	ZS-GRL[3]	0.369	0.370	0.110	0.121	0.075	0.068
	ICCV'19	ZS-SAKE[4]	0.497	0.598	0.475	0.599	–	–
	AAAI '20	ZS-GCN[5]	0.568	0.487	0.110	0.121	–	–
	NeurIPS '20	ZS-IIAE[6]	0.373	0.485	0.412	0.503	–	–
	TPAMI '21	ZS-TCN[7]	0.516	0.608	0.495	0.616	0.140	0.298
	AAAI '22	ZS-TVT[8]	0.531	0.618	0.484	0.662	0.149	0.293
	ACM MM '22	ZS-PSKD[ViT][9]	0.560	0.645	0.502	0.662	0.150	0.298
	CVPR '22	ZS-Sketch3T[10]	0.579	0.648	0.507	0.671	–	–
B-CLIP		B-FT	0.102	0.166	0.003	0.001	0.001	0.001
		B-Lin	0.422	0.512	0.398	0.557	0.082	0.098
		B-Cond	0.618	0.675	0.562	0.648	0.159	0.312
		B-IP	0.691	0.711	0.628	0.702	0.182	0.361
		B-MM	0.685	0.691	0.604	0.678	0.171	0.347
		B-Deep	0.702	0.718	0.637	0.718	0.188	0.375
		Ours	**0.723**	**0.725**	**0.651**	**0.732**	**0.202**	**0.388**

Cross-category Zero-Shot FG-SBIR

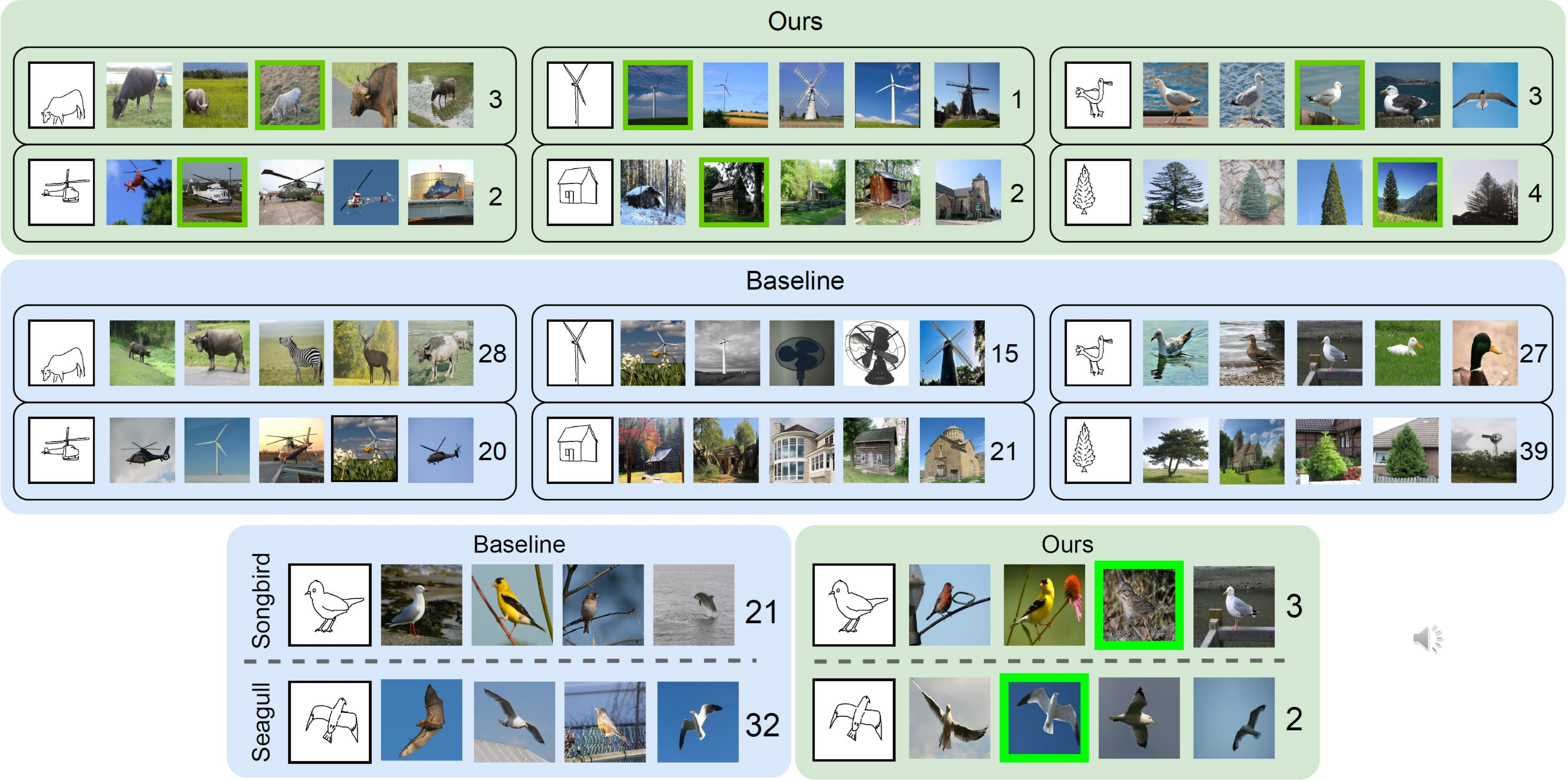
Methods	Sketchy Top-1	Sketchy Top-5
Cross-GRAD[11]	13.4	34.90
CC-DG[12]	22.6	49.00
B-FG-FT	1.23	4.56
B-FG-Lin	15.75	39.63
B-FG-Cond	25.98	54.38
B-FG-IP	26.69	56.08
B-FG-MM	27.16	59.46
B-FG-Deep	27.62	61.56
Ours	**28.68**	**62.34**

[1] Yelamarthi, Sasi Kiran, et al. A zero-shot framework for sketch based image retrieval. In ECCV, 2018.
[2] Dutta, Anjan, and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval." In CVPR, 2019.
[3] Dey, Sounak, et al. Doodle to search: Practical zero-shot sketch-based image retrieval." In CVPR, 2019.
[4] Liu, Qing, et al. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In ICCV, 2019.
[5] Zhaolong Zhang et. al. Zero-shot sketch-based image retrieval via graph convolution network. In AAAI, 2020.
[6] HyeongJoo Hwang et. al. Variational interaction information maximization for cross-domain disentanglement. In NeurIPS, 2020.
[7] Hao Wang et. al. Transferable coupled network for zero-shot sketch-based image retrieval. IEEE TPAMI, 2021.
[8] Jialin Tian et. al. TVT: Three-way vision transformer through multimodal hypersphere learning for zero-shot sketch-based image retrieval. In AAAI, 2022.
[9] Kai Wang et. al. Prototype-based selective knowledge distillation for zero-shot sketch based image retrieval. In ACM MM, 2022.
[10] Aneeshan Sain et. al. Sketch3t: Test-time training for zero-shot SBIR. In CVPR, 2022.
[11] Shiv Shankar et. al. Generalizing across domains via cross-gradient training. In ICLR, 2018.
[12] Kaiyue Pang et. al. Generalising fine-grained sketch-based image retrieval. In CVPR, 2019.

Qualitative Zero-Shot SBIR Results on Sketchy



Qualitative Fine Grained Zero-Shot SBIR Results on Sketchy



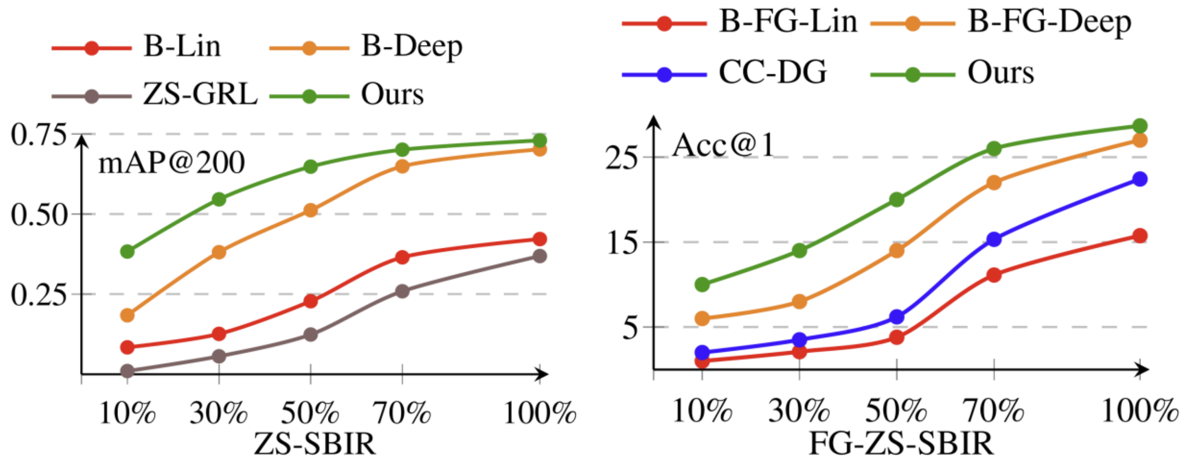
Ablation and Further Analysis

Methods	ZS-SBIR		FG-ZS-SBIR	
	mAP@all	P@200	Top-1	Top-5
w/o LayerNorm	0.698	0.701	27.18	59.55
w/o Classification ($\mathcal{L}_{\text{cls}}^{\mathcal{I}}$)	0.703	0.710	10.69	16.32
w/o Patch-Shuffling ($\mathcal{L}_{\text{PS}}$)	–	–	25.18	53.07
w/o f-Divergence ($\mathcal{L}_{\delta}$)	–	–	24.93	53.72
Ours	**0.723**	**0.725**	**28.68**	**62.34**

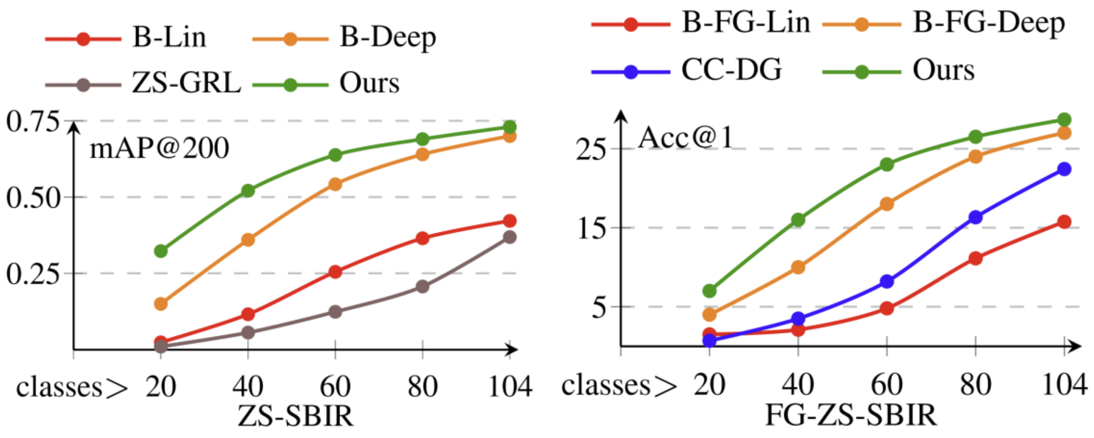
Towards alleviating data-scarcity for sketch-based applications !!

Ablation on model components, dropping one component at a time

Generalisation Potential of our model



Performance across varying training data-size



Performance across varying number of seen classes http://sketchx.ai

Please visit our project page for more:
https://aneeshan95.github.io/Sketch_LVM/