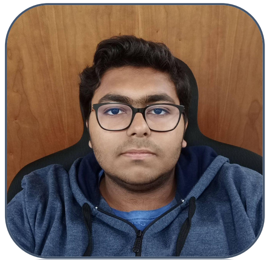
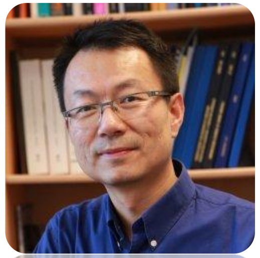
Exploiting Unlabelled Photos for Stronger Fine-Grained SBIR






Aneeshan Sain[a,b] Ayan Kumar Bhunia[a] Subhadeep Koley [a,b] Pinaki Nath Chowdhury [a,b] Soumitri Chattopadhyay[*] Tao Xiang[a,b] Yi-Zhe Song[a,b]

a. SketchX, CVSSP, University of Surrey, United Kingdom

b. iFlyTek-Surrey Joint Research Centre on Artificial Intelligence

* Interned with SketchX

Paper Tag: **TUE-PM-262**

SketchX

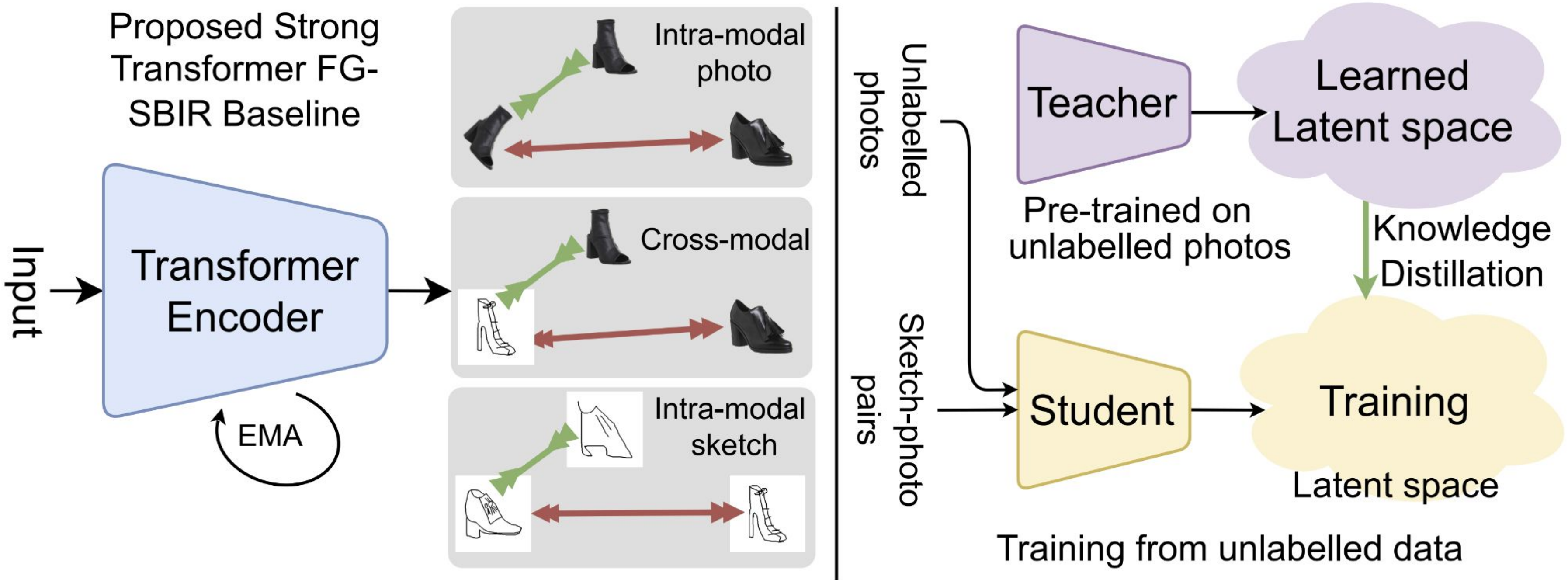
CVSSP Centre for Vision, Speech and Signal Processing

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

UNIVERSITY OF SURREY

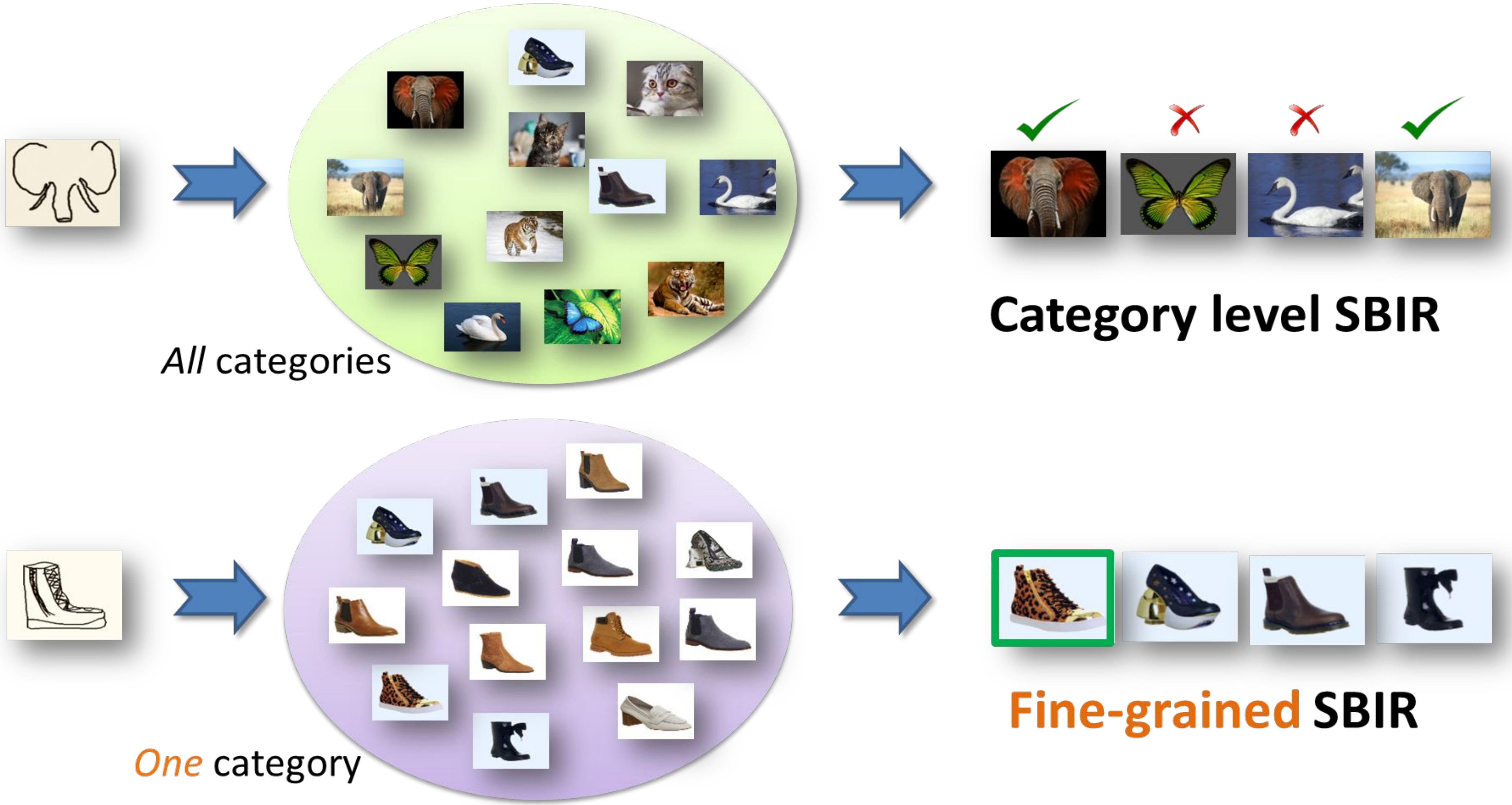
Overview

- We present a stronger baseline for fine-grained SBIR that addresses *two critical issues* facing the community – (i) **inadequate latent space feature separation** and (ii) **insufficient availability of paired sketches.**
- Specifically, we propose:
  - An ***intra-modal triplet objective*** in each modality that explicitly *enforces instance separation*
  - A novel paradigm to *leverage unlabeled data* in FG-SBIR by ***distilling knowledge from unlabelled photos***
  - A modified **PVT encoder** with a ***learnable distillation token*** that caters to the end-to-end learning approach.



- Our work overshoots prior state-of-the-arts by ≈11% and also yields satisfactory results on generalising to new classes, establishing itself as a ***stronger baseline*** for future fine-grained SBIR works.

Sketch-based Image Retrieval – Category-level to Fine-grained



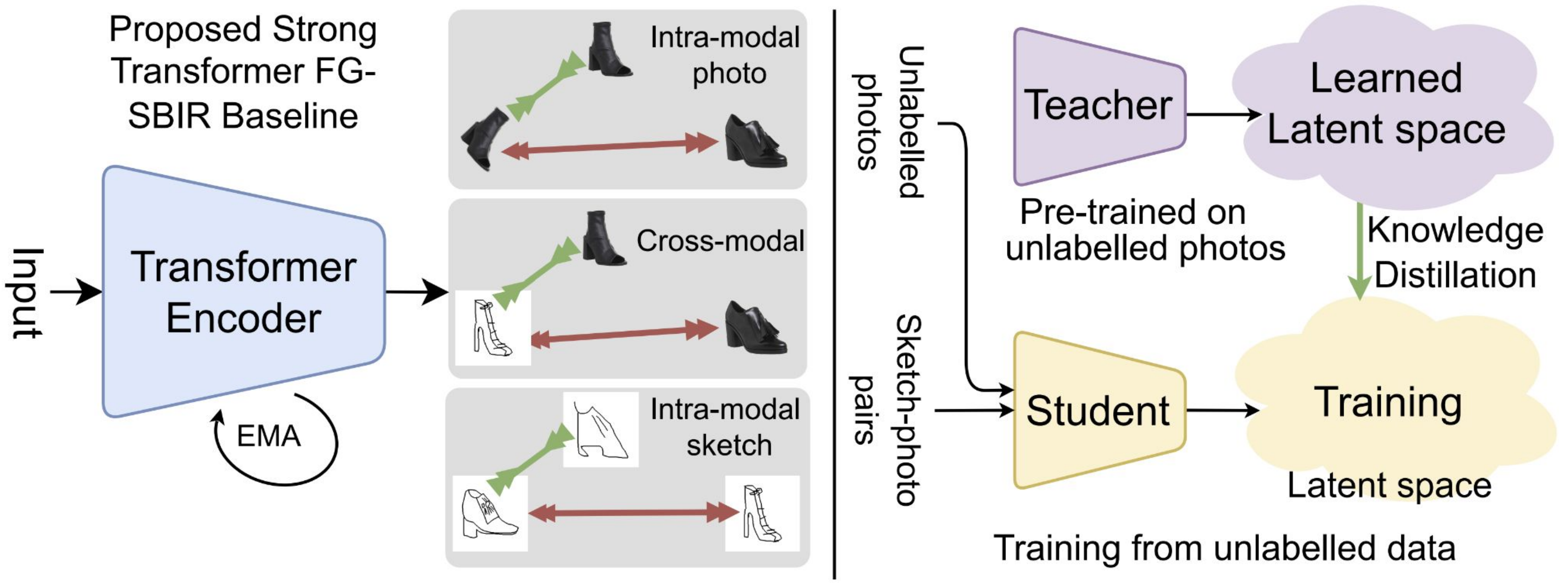
Motivation

Issues with existing fine-grained SBIR literature:

- the gold standard triplet loss does *not* enforce **holistic latent space geometry**
- **Insufficient availability** of *paired sketches* for SBIR training

What we need:

- Enforce **adequate latent space separation** amongst different photos or sketch instances
- Alleviate the constraint of sketch availability and **use unlabeled photos** to improve performance

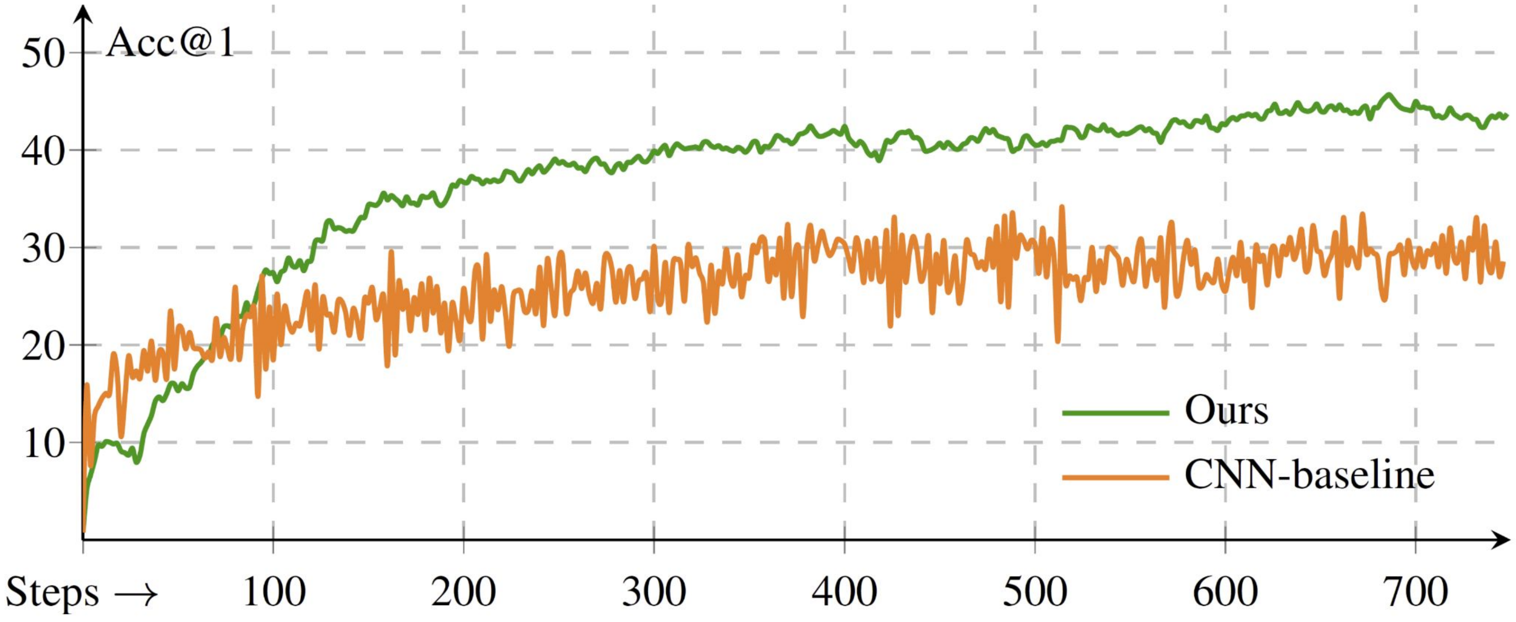


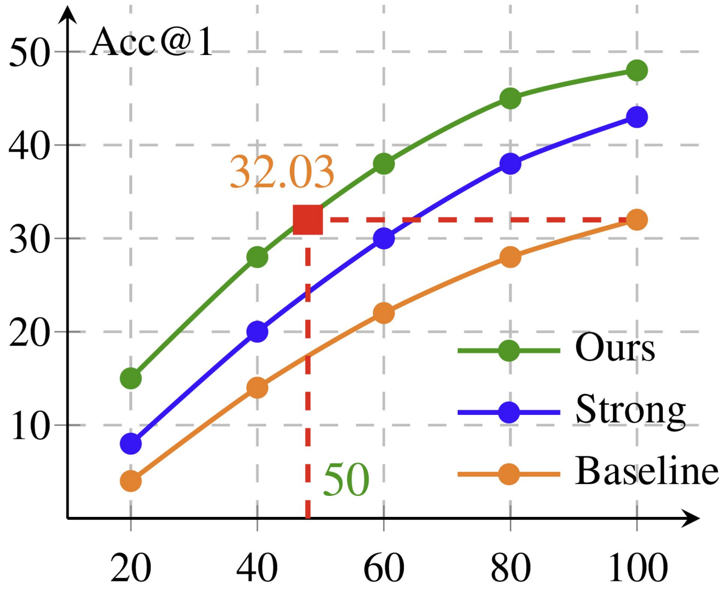
How we achieve:

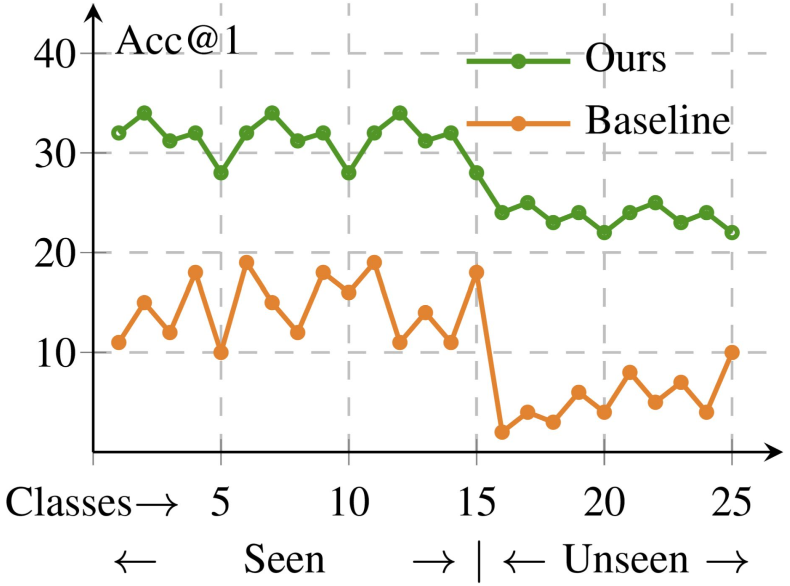
- We address the first issue by employing an **intra-modal triplet objective** in *both modalities*, which brings *matched sketches closer* as well as *pushes different sketch/photo instances farther* apart
- For the second issue, we adopt **knowledge distillation** – we train a model on *unlabelled photos* only *via an intra-modal triplet loss*, then **distill** its instance-wise discriminative knowledge to an *FG-SBIR model*.

Pilot Study

- Training stability
  - Baseline FG-SBIR shows *notorious instability* during training
  - For this, we introduce EMA in our learning paradigm which *imparts a stabilising effect* (see adjacent figure)
- Training dataset size
  - Our *stronger baseline* effectively utilises unlabeled photos and matches performance of existing baseline using *half the training data* (fig. bottom left)
- Unseen classes (cross-category)
  - Performance drop of our model on *unseen classes* is **less** compared to baseline, showing *greater generalisability across categories* (fig. bottom right)







Framework – Stronger FG-SBIR baseline model

- Vision Transformer backbone
  - *PVT* encoder ensures global receptive fields
- Cross-modal and Intra-modal losses
  - **Cross-modal**:
    – Traditional sketch-photo triplet loss

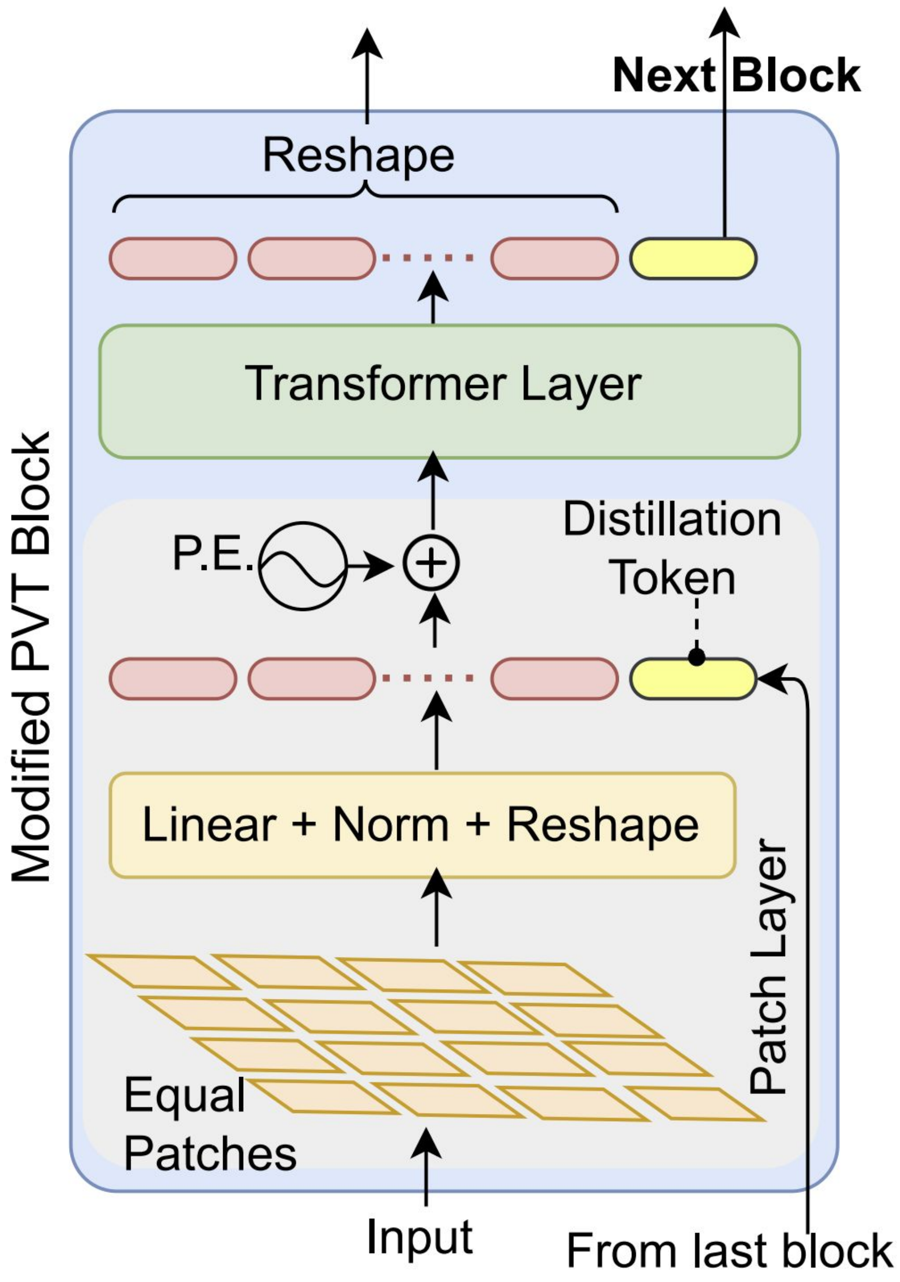
$$\mathcal{L}_{\text{Tri}}^{\text{CM}} = \max\{0, m_{\text{CM}} + \delta(f_s, f_p) - \delta(f_s, f_n)\}$$

  - **Intra-modal**:
    – Separates visually similar *sketch/photo instances*

$$\mathcal{L}_{\text{Tri}}^{\text{IM}_p} = \max\{0, m_{\text{IM}}^{p} + \delta(f_p, f_{p^t}) - \delta(f_p, f_n)\}$$
$$\mathcal{L}_{\text{Tri}}^{\text{IM}_s} = \max\{0, m_{\text{IM}}^{s} + \delta(f_s, f_{s^+}) - \delta(f_s, f_{s^-})\}$$

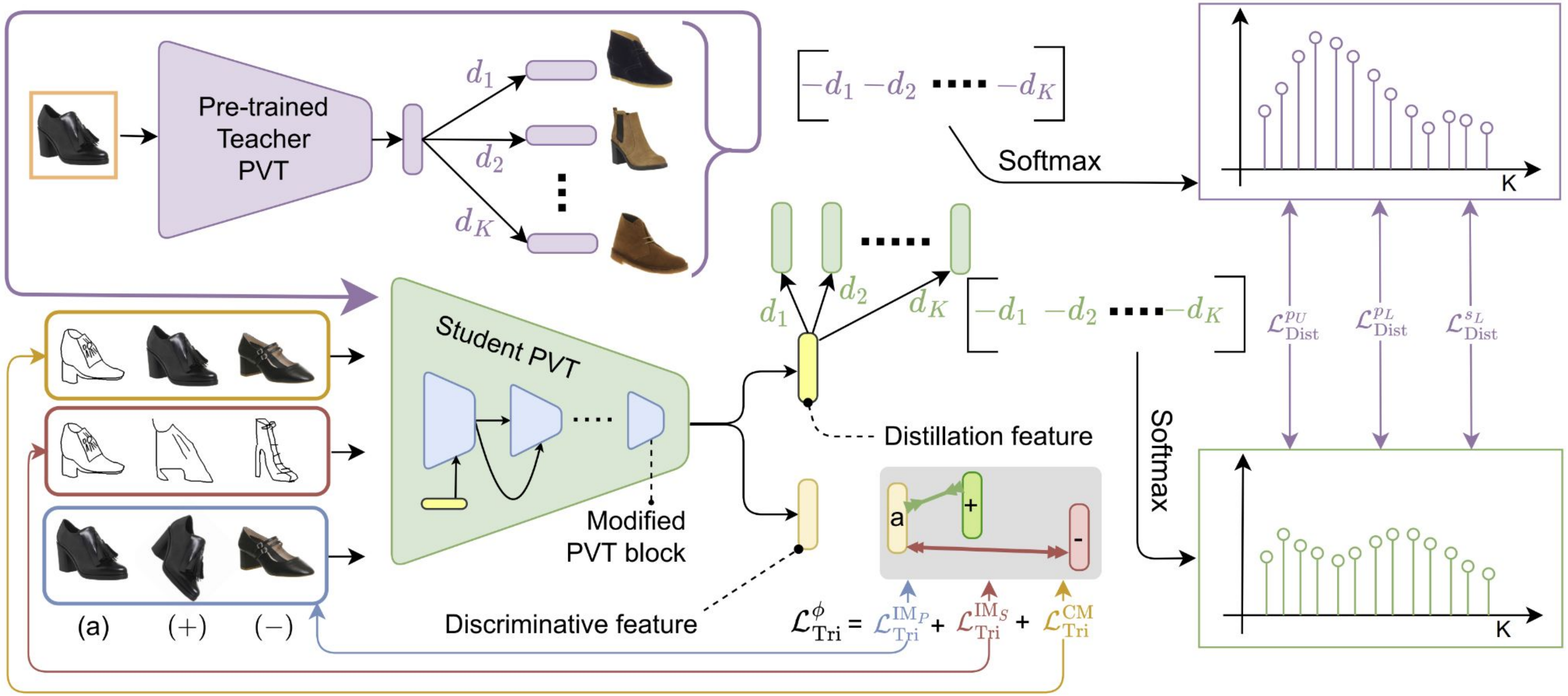
- Exponential Moving Average
  - *EMA* enhances training stability

$$\theta_{\text{EMA}}^{t} = \beta\theta_{\text{EMA}}^{t-1} + (1-\beta)\theta^{t}$$
$$\mathcal{L}_{Trn} = \mathcal{L}_{\text{Tri}}^{\text{CM}} + \lambda_1\mathcal{L}_{\text{Tri}}^{\text{IM}_p} + \lambda_2\mathcal{L}_{\text{Tri}}^{\text{IM}_s}$$



Framework – KD for unlabeled photos

- **Knowledge distillation paradigm**
- Knowledge transfer from a *photo instance discriminator* **teacher** to a *FG-SBIR* **student** model for cross-modal retrieval.
- **PVT** backbone: ***Learnable distillation token*** for knowledge transfer from teacher to student
- **Teacher** pre-training: ***Intra-modal photo loss*** (unlabelled photos)
- **Student** FG-SBIR training: ***Nearest neighbour pairwise distance based distillation***



$$\mathcal{L}^{\phi}_{\text{Disc}} = \mathcal{L}^{L}_{Trn} + \lambda_3 \mathcal{L}^{U}_{\text{Tri}}$$

$$\mathcal{S}_\tau(-\mathbf{D}^{\Omega}_{p^i})_{r^j} = \frac{\exp(-\delta(f^{\Omega}_{p^i}, f^{\Omega}_{p^{r_j}})/\tau)}{\sum_{k=1}^{r_K}\exp(-\delta(f^{\Omega}_{p^i}, f^{\Omega}_{p^{r_k}})/\tau)}$$

$$\mathcal{L}^{p_U}_{\text{KL}} = KL(\mathcal{S}_\tau(-\mathbf{D}^{\Omega}_{p^i}) \;||\; \mathcal{S}_\tau(-\mathbf{D}^{\phi}_{p^i}))$$

$$\mathcal{L}^{\phi}_{\text{Dist}} = \mathcal{L}^{p_L}_{\text{KL}} + \lambda_4\mathcal{L}^{s_L}_{\text{KL}} + \lambda_5\mathcal{L}^{p_U}_{\text{KL}}$$

$$\mathcal{L}^{\phi}_{trn} = \mathcal{L}^{\phi}_{\text{Disc}} + \lambda_6 \mathcal{L}^{\phi}_{\text{Dist}}$$

Experiments

- **Datasets used:**
  - QMUL-Chair-V2[1] - 2000 (400) sketch (photo) pairs
  - QMUL-Shoe-V2[1] - 6730 (2000) sketch (photo) pairs
  - Sketchy (Extended)[2] - 73K sketches across 125 categories.
  - UT-Zappos50K[3] - 50K unlabeled photos

- **Competitors:**
  - SOTA fine-grained SBIR methods – Triplet-SN[2], HOLEF-SN[4], Jigsaw-SN[5], OnTheFly[6], StyleMeUp[7]
  - SOTA methods augmented with our intra-modal triplet objectives (SOTA++)
  - Architectural variants (CNN and ViT alternatives)
  - Alternative baselines using unlabeled data for training

- **Evaluation protocol and metric:**
  - Acc.@q i.e. percentage of sketches having true matched photo in the top-q list

[1] Qian Yu, et al. Sketch me that shoe. In CVPR, 2016.
[2] Patsorn Sangkloy, et al. The sketchy database: learning to retrieve badly drawn bunnies. In ACM TOG, 2016.
[3] Aron Yuand, and Kristen Grauman. Fine-grained visual comparisons with local learning. In CVPR, 2014.
[4] Jifei Song, et al. Deep spatial-semantic attention for fine grained sketch-based image retrieval. In ICCV, 2017.
[5] Kaiyue Pang, et al. Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval. In CVPR, 2020.
[6] Ayan Kumar Bhunia, et al. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In CVPR, 2020.
[7] Aneeshan Sain, et al. Stylemeup: Towards style-agnostic sketch-based image retrieval. In CVPR, 2021.

Quantitative Results:

- Table on right shows results obtained on the ChairV2 and ShoeV2 datasets.
- Table below shows comparative results on Sketchy database.

Methods	Sketchy (%)		Methods	Sketchy (%)	
	Top-1	Top-5		Top-1	Top-5
Triplet-SN[1]	15.32	34.15	B-InceptionV3	28.71	71.56
HOLEF-SN[2]	16.71	35.92	B-VGG-16	18.84	38.63
Jigsaw-SN[3]	16.74	36.37	B-ViT	7.63	11.23
OnTheFly[4]	04.76	07.81	B-SWIN	32.14	57.68
StyleMeUp[5]	19.62	39.72	B-CoAtNet	33.63	59.31
Triplet-SN-ours	19.48	37.91	B-Edge-Pretrain	34.98	61.32
HOLEF-SN-ours	20.23	38.61	B-Edge2Sketch	35.81	61.74
Jigsaw-SN-ours	21.45	39.56	B-Regress	36.33	62.31
OnTheFly-ours	07.28	12.14	B-RKD	37.02	63.02
StyleMeUp-ours	22.95	45.84	B-PKT	38.62	63.94
Ours-Strong	34.72	65.10	**Ours-Full**	38.54	71.52

		Methods	Chair-V2 (%)		Shoe-V2 (%)	
			Top-1	Top-10	Top-1	Top-10
SOTA		Triplet-SN[1]	47.45	84.32	28.71	71.56
SOTA		HOLEF-SN[2]	50.41	86.31	31.24	74.61
SOTA		Jigsaw-SN[3]	53.41	87.56	33.51	76.86
SOTA		OnTheFly[4]	54.54	88.61	34.10	78.82
SOTA		StyleMeUp[5]	59.86	89.64	36.47	81.83
SOTA		Semi-sup-SN[6]	60.20	90.81	39.12	85.21
Stronger Baseline	SOTA++	Triplet-SN-ours	53.48	87.91	33.78	76.84
Stronger Baseline	SOTA++	HOLEF-SN-ours	55.23	88.61	35.41	78.85
Stronger Baseline	SOTA++	Jigsaw-SN-ours	58.51	88.78	37.64	79.78
Stronger Baseline	SOTA++	OnTheFly-ours	59.18	89.35	38.62	81.97
Stronger Baseline	SOTA++	StyleMeUp-ours	65.85	90.84	40.42	82.94
Stronger Baseline	SOTA++	Semi-sup-SN-ours	66.86	91.12	44.35	86.83
Stronger Baseline	Backbone Variants	B-ResNet-18	48.42	85.62	26.61	70.31
Stronger Baseline	Backbone Variants	B-ResNet-50	47.78	82.34	28.12	70.84
Stronger Baseline	Backbone Variants	B-InceptionV3	55.41	88.21	34.24	78.56
Stronger Baseline	Backbone Variants	B-VGG-16	58.23	88.78	35.85	80.92
Stronger Baseline	Backbone Variants	B-VGG-19	61.46	89.16	37.28	81.01
Stronger Baseline	Backbone Variants	B-ViT	38.71	72.65	16.28	53.42
Stronger Baseline	Backbone Variants	B-DeIT	56.25	87.72	35.62	79.05
Stronger Baseline	Backbone Variants	B-SWIN	66.34	91.03	40.71	82.57
Stronger Baseline	Backbone Variants	B-CvT	68.42	91.21	41.58	83.14
Stronger Baseline	Backbone Variants	B-CoAtNet	69.68	91.78	42.63	83.20
Stronger Baseline		**Ours-Strong**	**71.22**	**92.18**	**44.18**	**84.68**
Unlabelled		B-Edge-Pretrain	71.58	90.78	44.62	84.85
Unlabelled		B-Edge2Sketch	72.16	91.01	45.18	84.92
Unlabelled		B-Regress	72.65	91.32	45.45	85.01
Unlabelled		B-RKD	73.02	91.78	46.18	85.12
Unlabelled		B-PKT	73.45	91.89	46.66	85.47
Unlabelled		**Ours-Full**	**74.68**	**92.79**	**48.35**	**85.62**

[1] Sangkloy et al. The sketchy database: learning to retrieve badly drawn bunnies. In ACM TOG, 2016.

[2] Song et al. Deep spatial-semantic attention for fine grained sketch-based image retrieval. In ICCV, 2017.

[3] Pang et al. Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval. In CVPR, 2020.

[4] Bhunia et al. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In CVPR, 2020.

[5] Sain et al. Stylemeup: Towards style-agnostic sketch-based image retrieval. In CVPR, 2021.

[6] Bhunia et al. More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval. In CVPR, 2021.

Ablative Studies:

Ablation study of loss function on the QMUL-ShoeV2 dataset.

Type	$\mathcal{L}_{\text{Tri}}^{\text{CM}}$	$\mathcal{L}_{\text{Tri}}^{\text{IM}}$	EMA	$\mathcal{L}_{\text{KL}}^{pU}$	$\mathcal{L}_{\text{KL}}^{pL}$	$\mathcal{L}_{\text{KL}}^{sL}$	Top-1 (%)
I	✓	-	-	✓	✓	✓	43.28
II	✓	✓	-	✓	✓	✓	45.39
III	✓	✓	✓	✓	-	-	46.50
IV	✓	✓	✓	✓	✓	-	47.21
Ours-Full	✓	✓	✓	✓	✓	✓	**48.35**

Cross-category generalisation on Sketchy database.

Methods	Sketchy (%)		Methods	Sketchy (%)	
	Top-1	Top-5		Top-1	Top-5
Jigsaw-SN[1]	23.16	44.63	B-Edge-Pretrain	24.81	46.24
Adaptive-SN[2]	32.71	53.42	B-Edge2Sketch	25.74	48.36
CC-Gen[3]	22.73	42.32	**Ours-Full**	**30.24**	**51.65**

[1] Pang et al. Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval. In CVPR, 2020.
[2] Bhunia et al. Adaptive fine-grained sketch-based image retrieval. In ECCV, 2022.
[3] Pang et al. Generalising fine-grained sketch-based image retrieval. In CVPR, 2019.

Thank You!

SketchX

http://sketchx.ai

aneeshan95.github.io/Sketch_PVT