Cross-Modal Fusion Distillation for Fine-Grained Sketch-Based Image Retrieval

Abhra Chaudhuri [1] Massimiliano Mancini [2] Yanbei Chen [2] Zeynep Akata [2 3 4] Anjan Dutta [5]

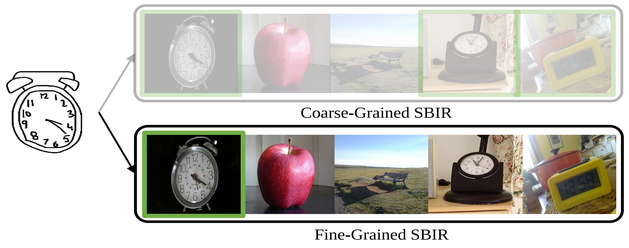
[1]University of Exeter [2]University of Tübingen [3]MPI for Informatics [4]MPI for Intelligent Systems [5]University of Surrey

BMVC 2022

TL;DR

- **Observation:** *Modality-specific* features can also be *instance-discriminative*.
- **Action:** Capture them via *cross-attention*. Bypass test-time overhead via *knowledge distillation*.
- **Results:** SOTA on benchmark FG-SBIR datasets.

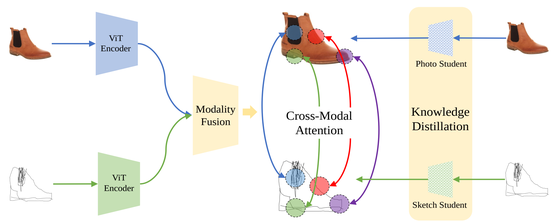
Coarse-Grained SBIR *vs.* Fine-Grained SBIR



Fine-grained SBIR aims to retrieve a *specific* photo corresponding to a query sketch, unlike the coarse-grained setting, where all photos of the corresponding class are returned.

Types of Cross-Modal Features

1. **Modality-shared** (useful for retrieval).
2. **Modality-specific**, that represents a *shared underlying concept*, but manifests differently in the two modalities (**useful for retrieval**).
3. **Modality-specific**, that does not represent a shared concept (**not useful for retrieval**).



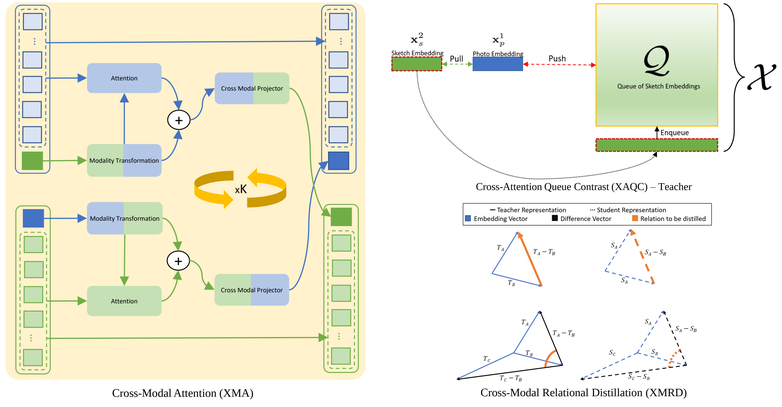
Color contrast may be depicted in sketches via scribbles (area under blue circles).

We simultaneously preserve (1) and (2) while discarding (3) via **Cross-Modal Attention (XMA)**. We bypass the test-time computational overhead of XMA via **Cross-Modal Knowledge Distillation**.

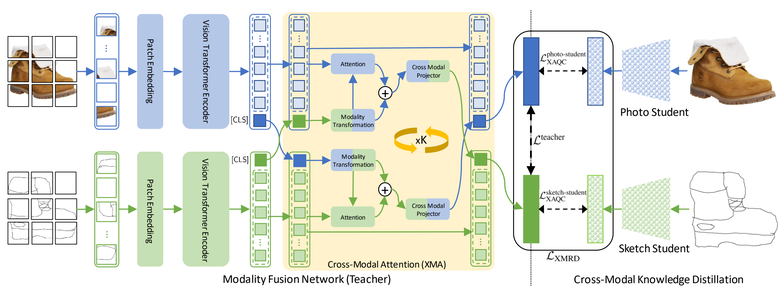
Contributions

- Leveraging **modality-specific, instance-discriminative** features for FG-SBIR.
- Cross-Modal Knowledge Distillation for **fast retrieval without test-time cross-attention** [1].
- **State-of-the-art results** on **benchmark FG-SBIR** datasets.

The XModalViT Framework



XMA → Modality-Fusion | XAQC → Ranking, XModal KD | XMRD → Relational Invariants



XModalViT: **Modality Fusion** (XMA) followed by **disentanglement** (XMKD) for **fast retrieval**.

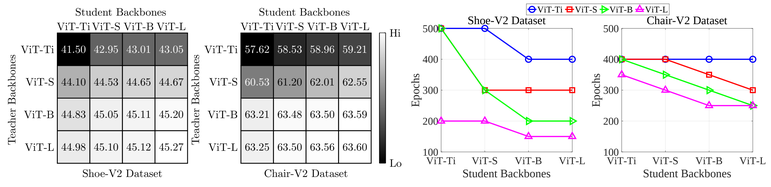
Comparison with SOTA

Method	Shoe-V2 acc@1	Shoe-V2 acc@10	Chair-V2 acc@1	Chair-V2 acc@10
Yu *et al.*, CVPR '16	28.71	71.56	47.65	84.24
Yang *et al.*, ICCV '21	32.33	79.63	52.89	94.88
Sain *et al.*, CVPR '21	36.47	81.83	62.86	91.14
Bhunia *et al.*, CVPR '21	39.10	87.50	62.20	90.80
Chowdhury *et al.*, CVPR '22	39.90	82.90	-	-
Bhunia *et al.*, CVPR '22	43.70	-	64.80	-
Ours (XModalViT)	**45.05**	**90.23**	63.48	**95.02**

Method	Sketchy acc@1	Sketchy acc@10
Human [2]	54.27	–
Yu *et al.*, CVPR '16	25.87	–
Sangkloy *et al.*, SIGGRAPH '16	37.10	–
Pang *et al.*, BMVC '17	50.14	–
Wang *et al.*, PR '20 (S+I)	40.16	92.00
Wang *et al.*, PR '20 (S+I+D)	46.20	96.49
Ours (XModalViT)	**56.15**	**96.86**

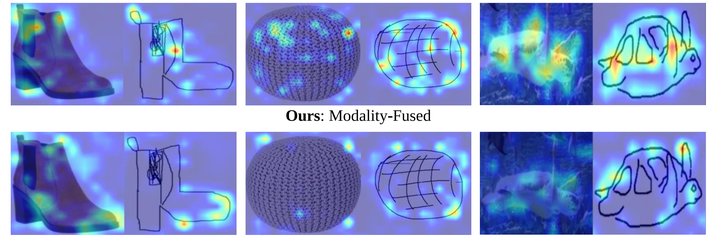
Surpasses **SOTA** on both **single-class** (Shoe-V2 & Chair-V2) & **multi-class** (Sketchy) benchmarks.

Effect of Varying Backbone Sizes



(Left) **Variation in acc@1** and (right) **student convergence time** (# epochs) with encoder size.

Qualitative Results



Ours: Modality-Fused

Modality-Shared Only

Our method attends to **modality-specific features** like **scribbles** (Shoe), **grids/meshes** (Chair), or **primitive shapes** like square/triangle (Turtle), that are **instance discriminative**.

Sample **top-10** retrieval results in order obtained using our method (green → ground-truth).

References

[1] Chun-Fu (Richard) Chen, Quanfu Fan, and Rameswar Panda. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In *ICCV*, 2021.

[2] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: Learning to retrieve badly drawn bunnies. *ACM SIGGRAPH*, 2016.

https://github.com/abhrac/xmodal-vit

arXiv https://arxiv.org/abs/2210.10486

Abhra Chaudhuri: ac1151@exeter.ac.uk