Freeview Sketching

View-Aware Fine-Grained Sketch-Based Image Retrieval

Aneeshan Sain

Pinaki Nath Chowdhury

Subhadeep Koley

Ayan Kumar Bhunia

Yi-Zhe Song

SketchX, CVSSP, University of Surrey, United Kingdom

SketchX

CVSSP
Centre for Vision, Speech and Signal Processing

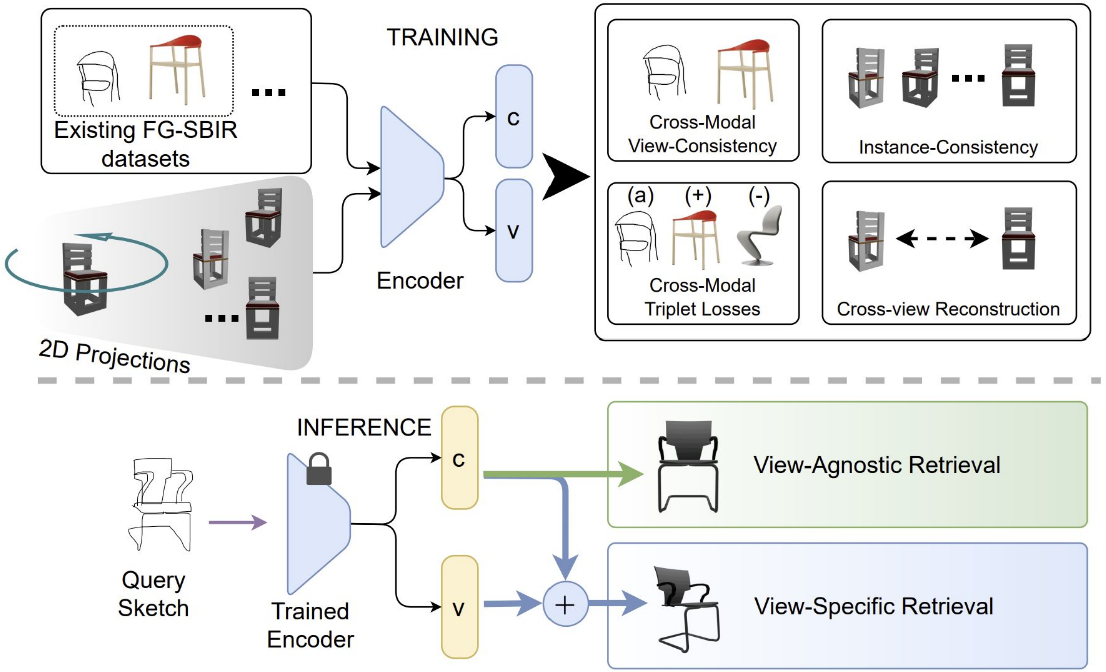
ECCV
EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO 2024

UNIVERSITY OF SURREY

Overview

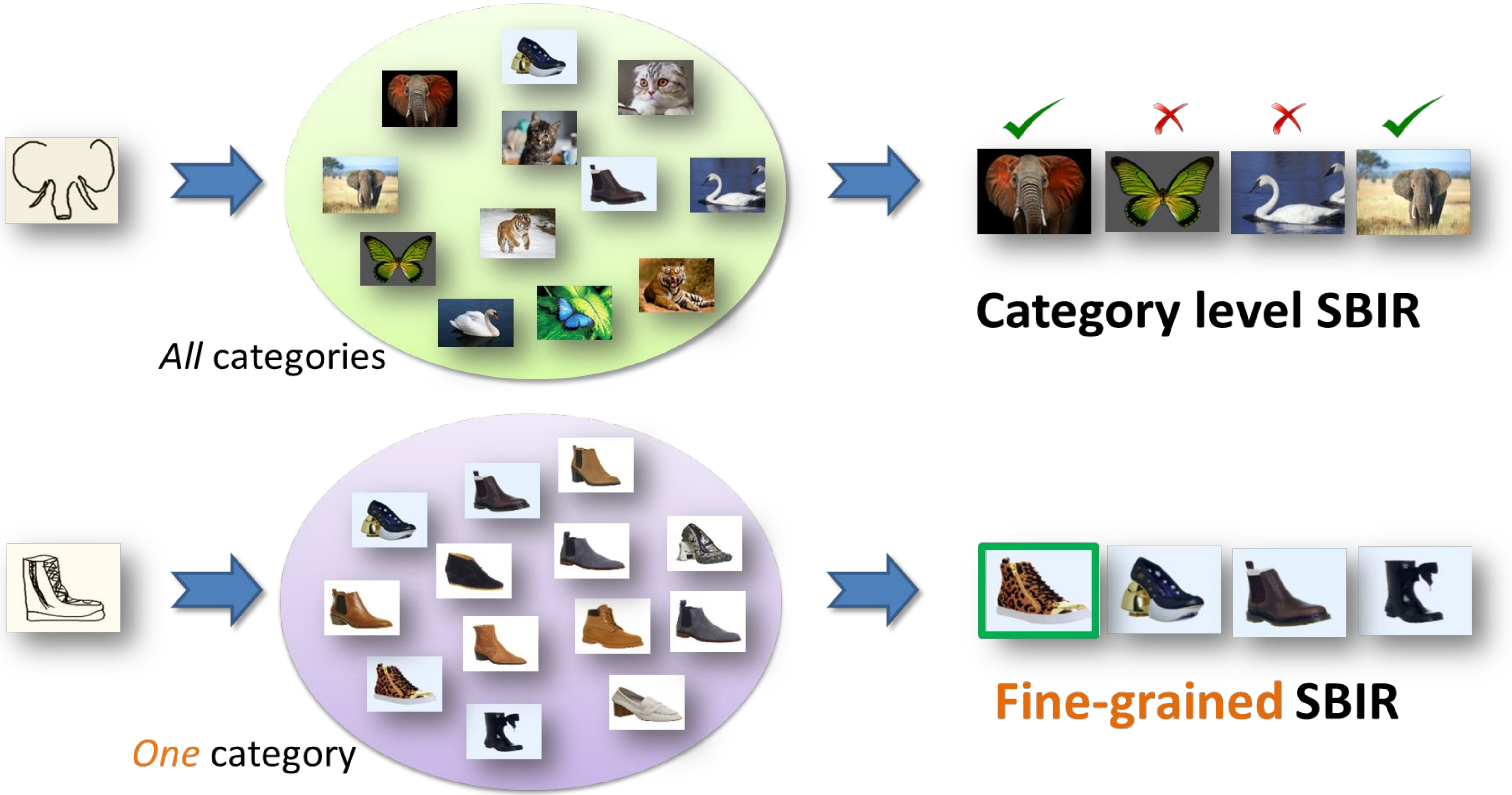
- We address a critical yet overlooked aspect of Fine-Grained Sketch-Based Image Retrieval (FG-SBIR) – the choice of viewpoint while sketching. With **very limited data** collected from ***fixed*** viewpoint, sketch systems often **mismatch** between *view* of the user's sketch and that of its image in the gallery, leading to **inaccurate** retrievals.

- Specifically, we:
  - propose a ***view-aware system*** designed to accommodate both view-agnostic and view-specific FG-SBIR seamlessly.
  - introduce the use of ***multi-view 2D rendered projections*** of 3D objects promoting cross-modal view awareness.
  - present a ***customisable cross-modal feature*** through a disentanglement framework, allowing an easy switch between view-agnostic and view-specific retrieval modes.



- Satisfactory qualitative and quantitative results show this avenue of view-aware FG-SBIR to be a promising direction of future research.

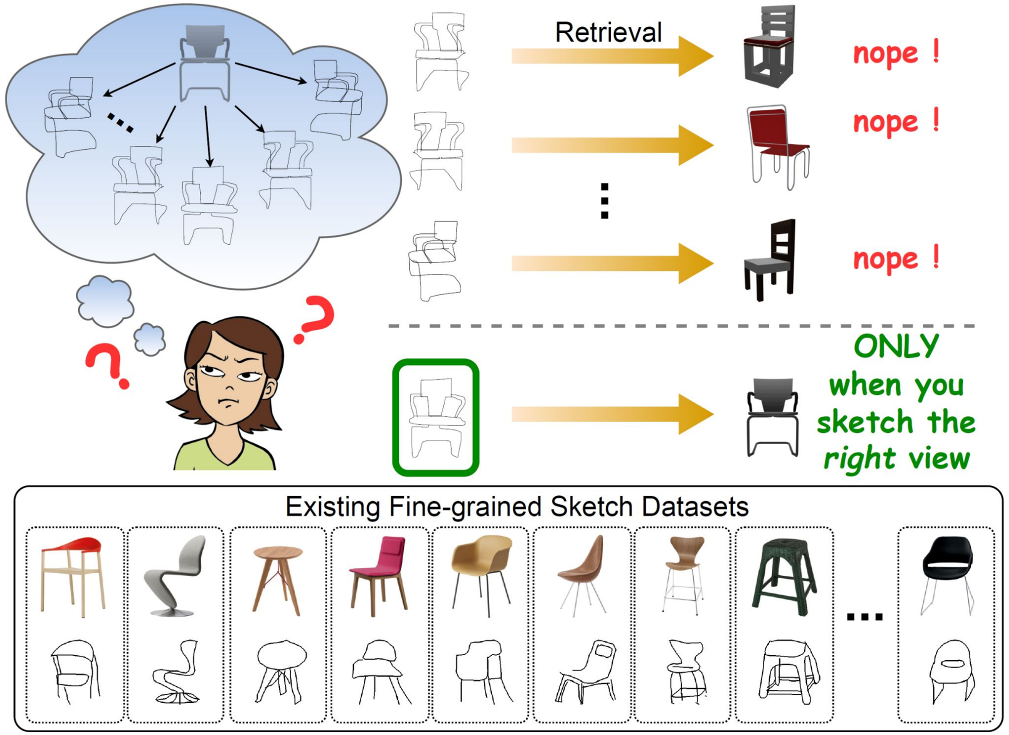
Sketch-based Image Retrieval – Category-level to Fine-grained



Motivation

Issues with existing fine-grained SBIR literature:

➔ While searching a specific photo, users are often ***confused*** as to 'Which view should I sketch?'
➔ Existing sketch-photo pairs are matched against a ***single fixed view***. So a sketch drawn from a different view retrieves an **incorrect** image.
➔ ***Absence of view-specific annotations*** in sketch-photo pairs.



What we desire:

➔ Incorporate **view-awareness** in the task of FG-SBIR.
➔ **Alleviate the challenge of data-scarcity** in having multi-view sketches of the same object for training.

How we achieve:

➔ Introduce the use of **multi-view 2D rendered projections** of 3D objects to overcome the limitations of existing datasets,
➔ **Carefully design a training paradigm** involving cross-modal discriminative guiding signals as well as reconstruction objectives catered to distill view-aware knowledge simultaneously.
➔ Present a **customisable cross-modal feature through a disentanglement framework**, that can switch between retrieving the target instance in any view or in the specific view as that of the sketch.

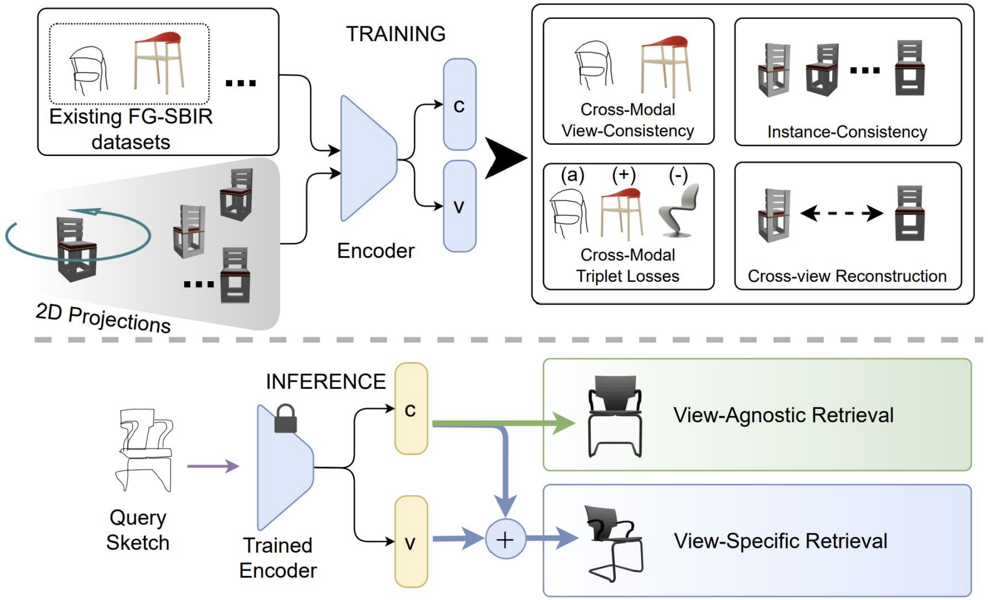
Pilot Study

- We use a FG-SBIR model pre-trained on fixed single-view sketch-photo pairs.
- Experiment Setup:
    - Existing – Photos of target instance matching the ***exact*** view of query-sketches (0°, 30° and 75°) are **present** in the test-gallery alongside other views.
    - Pilot – Photos of target instance matching the ***exact*** view of the sketch are **absent**, but other views are present.
- Result:
    - Existing : 58.25% Top-1 Accuracy.
    - Pilot: 31.10% Top-1 Accuracy.
- Inference: FG-SBIR models trained on **fixed single-view** sketch-photo pairs cannot generalise to sketches whose view ***doesn't match*** its target photo.

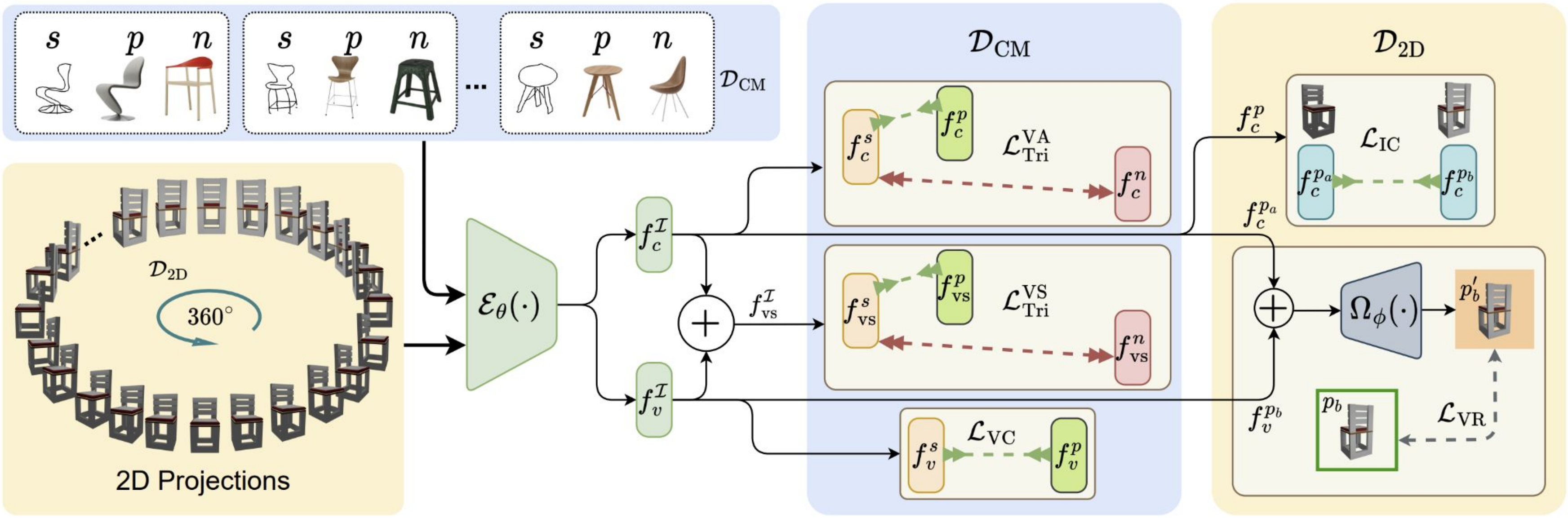
Problem Definition

➔ We recognise and put forth two novel tasks of FG-SBIR :

- ◆ View-*Agnostic* FG-SBIR : retrieve a photo that matches ***any*** view of the target instance.
- ◆ View-*Specific* FG-SBIR: retrieve that photo of the target instance whose view matches ***exactly*** with the query sketch.



Training



Architecture and problem setup:

➔ We employ a simple VGG-16 network as backbone feature extractor.

➔ The encoder extracts and disentangles an input image into two component features: *view* and *content.*

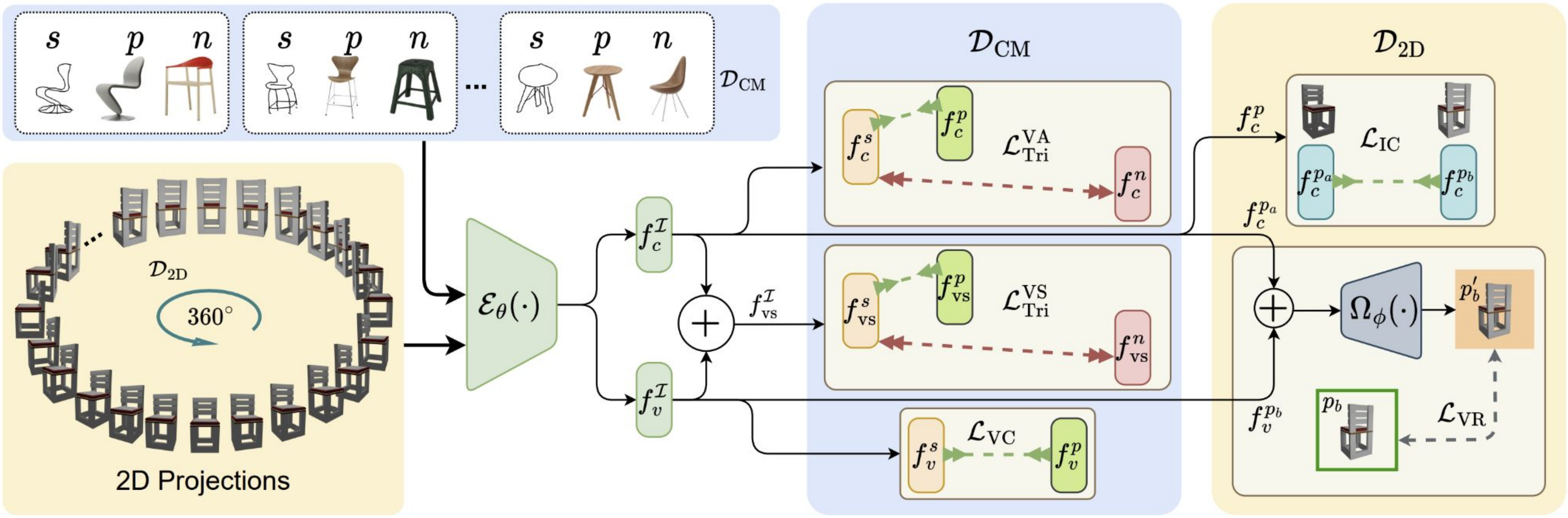
➔ To avoid high complexity of *3D*-based training paradigm, we **restrict our setup** to a *2D* learning paradigm.

➔ To alleviate data-scarcity, we use two datasets:

a. Standard FG-SBIR dataset for cross-modal sketch-photo pairs (D_{CM}) with fixed view.

b. 2D projections freely rendered from pre-existing 3D shapes (D_{2D}) to harness view-aware knowledge.

Training (Contd.)



Learning objectives:

- Cross-modal Discriminative Learning for fine-grained matching:
  - Traditional sketch-photo triplet loss on content for view-agnostic discriminative learning.
  - Loss for view-specific feature $f^{\mathcal{I}}_{\mathrm{vs}} = f^{\mathcal{I}}_c + f^{\mathcal{I}}_v$.
- Learning from 2D Projections to alleviate data-scarcity:
  - Instance-consistency loss across different projections of the same photo to train for view-agnostic retrieval.
  - Cross-view reconstruction loss to enrich latent space on view-specific knowledge from photos
  - Cross-modal View Consistency loss to instill cross-modal sketch-photo view-awareness.

$$\mathcal{L}^{\mathrm{VA}}_{\mathrm{Tri}} = \max\{0, \mu_c + \delta(f^s_c, f^p_c) - \delta(f^s_c, f^n_c)\}$$

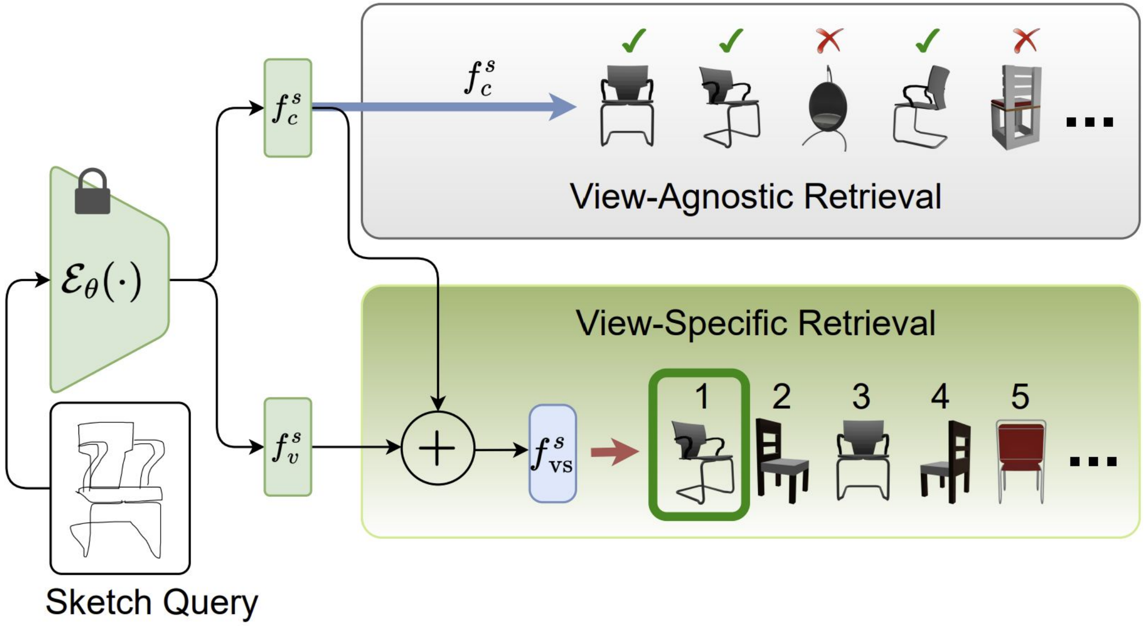
$$\mathcal{L}^{\mathrm{VS}}_{\mathrm{Tri}} = \max\{0, \mu_{\mathrm{vs}} + \delta(f^s_{\mathrm{vs}}, f^p_{\mathrm{vs}}) - \delta(f^s_{\mathrm{vs}}, f^n_{\mathrm{vs}})\}$$

$$\mathcal{L}_{\mathrm{IC}} = \frac{1}{\binom{\mathrm{M}_i}{2}} \sum_{a=1}^{\mathrm{M}_i-1} \sum_{b=a+1}^{\mathrm{M}_i} \left|\left| f^{p_a}_c - f^{p_b}_c \right|\right|_2$$

$$\mathcal{L}_{\mathrm{VR}} = \frac{1}{\binom{\mathrm{M}_i}{2}} \sum_{a=1}^{\mathrm{M}_i-1} \sum_{b=a+1}^{\mathrm{M}_i} \left|\left| p'_b - p_b \right|\right|_2$$

$$\mathcal{L}_{\mathrm{VC}} = ||f^s_v - f^p_v||_2$$

Evaluation



We use ***only* one** model for **two** types of evaluation setups, just via feature selection:

- View-Agnostic FG-SBIR **:**
  - ***Only*** content feature $f_c^{\mathcal{I}}$ is used as '*view*' is not necessary.

- View-Specific **:**
  - View-specific feature $f_{\text{vs}}^{\mathcal{I}}$ is used as besides *content* it **also** holds the *view*, which is needed for additionally matching the *view*.

Experiments

- **Datasets used:**
  - Dataset by Qi et al.[8] – 555 and 1005 sketch/3D-shape quadruplets of 'lamps' and 'chairs'.
  - QMUL-Chair-V2[1] – 2000 (400) sketch (photo) pairs.

- **Competitors:**
  - SOTA FG-SBIR methods – Triplet-SN[2], HOLEF-SN[4], Jigsaw-CM[5], Triplet-OTF[6], StyleVAE[7], StrongPVT[9].
  - Variants with different backbones (CNN and vision transformer alternatives).
  - Alternative baselines using different mechanisms for disentanglement.
  - A few probable alternative ideas towards our motivation.

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

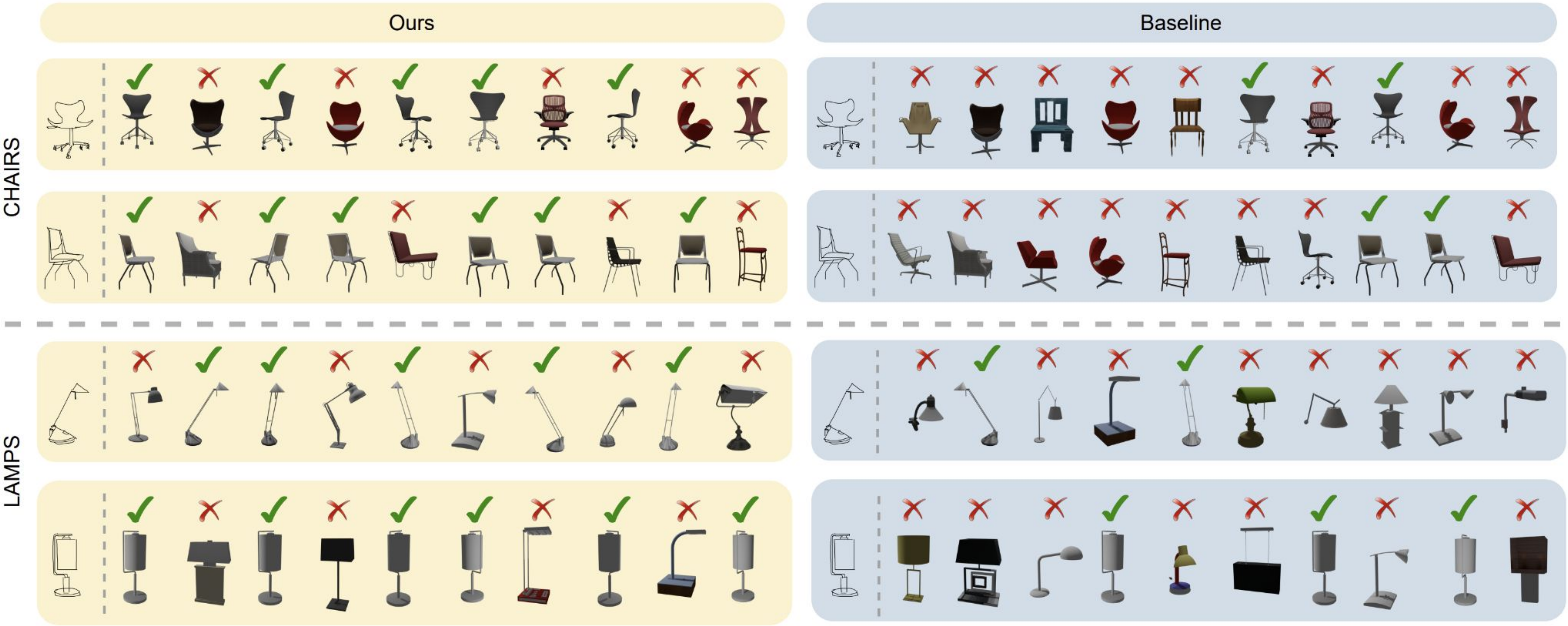
[1] Qian Yu, et al. Sketch me that shoe. In CVPR, 2016.
[2] Patsorn Sangkloy, et al. The sketchy database: learning to retrieve badly drawn bunnies. In ACM TOG, 2016.
[3] Aron Yuand, and Kristen Grauman. Fine-grained visual comparisons with local learning. In CVPR, 2014.
[4] Jifei Song, et al. Deep spatial-semantic 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.
[8] Anran Qi, et. al. Toward Fine-Grained Sketch-Based 3D Shape Retrieval. In TIP, 2021.
[9] Aneeshan Sain, et al. Exploiting Unlabelled Photos for Stronger Fine-Grained SBIR. In CVPR, 2022.

Quantitative Results:

Quantitative evaluation for View-Aware FG-SBIR

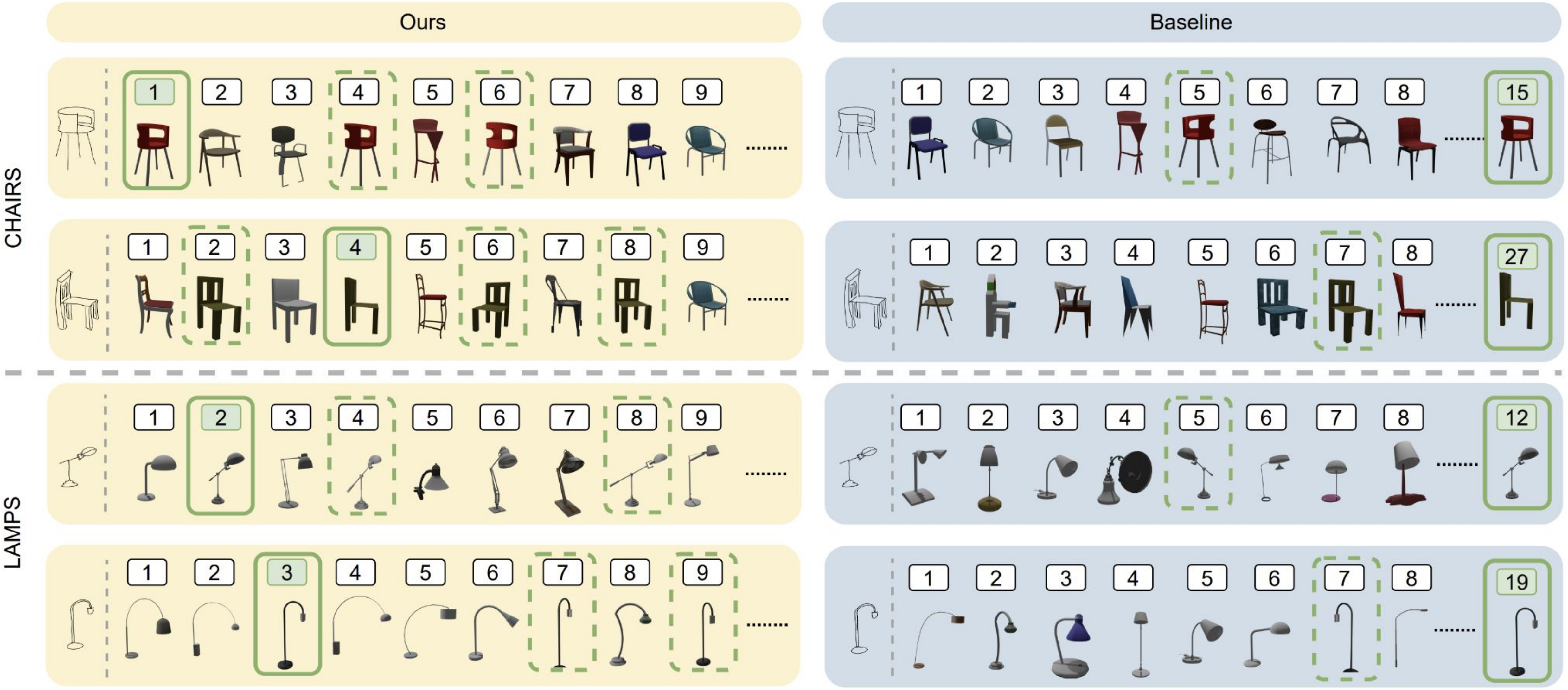
Methods		View-Agnostic				View-Specific			
		Chairs		Lamps		Chairs		Lamps	
		mAP@all	P@100	mAP@all	P@100	Top-1	Top-10	Top-1	Top-10
SoTA	Triplet-SN [75]	0.379	0.447	–	–	34.88	76.62	–	–
	HOLEF-SN [58]	0.398	0.454	–	–	37.23	78.63	–	–
	Jigsaw-CM [43]	0.432	0.525	–	–	41.14	81.78	–	–
	Triplet-OTF [6]	0.447	0.514	–	–	42.21	82.79	–	–
	StyleVAE [53]	0.523	0.602	–	–	46.19	87.66	–	–
	StrongPVT [50]	0.569	0.624	–	–	55.93	90.78	–	–
B-Backbones (Ours)	ViT [19]	0.385	0.415	0.338	0.399	34.38	76.18	33.96	75.53
	ResNet-50 [22]	0.451	0.536	0.415	0.511	47.15	88.02	46.21	87.11
	Inception-V3 [6]	0.512	0.573	0.468	0.542	50.18	90.19	49.11	89.23
	VGG-16 [56]	0.615	0.693	0.552	0.664	60.71	91.18	60.56	90.62
	PVT [63]	0.689	0.742	0.628	0.716	67.11	91.78	65.35	92.97
B-Disentangle	B-TVAE [27]	0.394	0.449	0.345	0.414	36.89	77.91	35.28	74.92
	B-DVML [36]	0.417	0.478	0.381	0.458	45.63	86.94	43.21	84.11
	B-Trio [10]	0.572	0.629	0.501	0.582	58.68	90.85	55.63	89.02
B-Misc	B-Single [37]	0.221	0.281	0.184	0.233	18.68	45.68	17.91	44.69
	B-Pivot [12]	0.316	0.401	0.295	0.362	55.92	90.62	53.62	88.65
	B-TwoModel	0.421	0.498	0.382	0.459	48.23	89.21	46.93	87.75
	B-NoProjection	0.592	0.667	0.529	0.611	50.79	89.93	48.73	87.98
SoTA++ ($\mathcal{D}_{\text{CM}}^{\text{Chair*}}$)	Triplet-SN [75]	0.416	0.476	0.378	0.451	43.09	83.29	41.32	81.48
	HOLEF-SN [58]	0.428	0.502	0.387	0.466	45.78	87.33	43.89	85.42
	Jigsaw-CM [43]	0.492	0.539	0.442	0.518	48.51	88.59	46.51	86.67
	Triplet-OTF [6]	0.521	0.591	0.476	0.571	49.53	89.66	47.49	87.71
	StyleVAE [53]	0.618	0.675	0.553	0.644	54.36	90.71	52.12	88.73
	StrongPVT [50]	0.641	0.708	0.584	0.677	64.68	91.15	62.02	90.15
	Ours-PVT [63]	0.702	0.771	0.681	0.749	70.26	92.86	68.32	93.04

Qualitative Results:



Qualitative Results of View-Agnostic FG-SBIR

Qualitative Results:

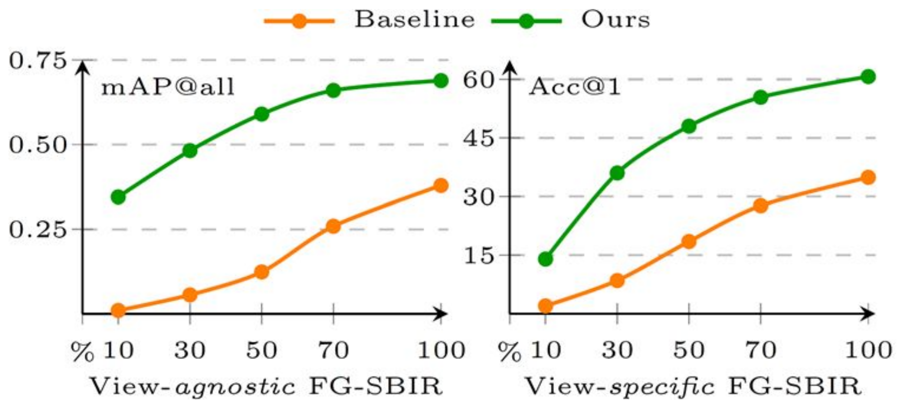


Qualitative Results of View-Specific FG-SBIR

Ablative Studies:

Objective-stripped	$\mathcal{L}_{\mathrm{Tri}}^{\mathrm{VA}}$	$\mathcal{L}_{\mathrm{Tri}}^{\mathrm{VS}}$	$\mathcal{L}_{\mathrm{VC}}$	$\mathcal{L}_{\mathrm{IC}}$	$\mathcal{L}_{\mathrm{VR}}$	Ours-VGG-16
[VS] Top-1 (%)	25.56	21.12	55.69	52.71	56.26	60.71
[VA] mAP@all	0.104	0.416	0.541	0.520	0.565	0.615

Ablation of Loss Objectives on ‘Chairs’



Performance under Low Data regime.

Other findings:

- Optimal feature dimension for both *content* and *view* features were empirically found to be 128.
- Our-VGG16 utilises 14.71 mil. params with ~ 40.18 GFLOPs.
- It takes 0.16ms (0.21ms) for view-specific (agnostic) retrieval per query during evaluation.

Thank You!

SketchX

http://sketchx.ai

https://aneeshan95.github.io/Sketch_Freeview/