Cross-Modal Hierarchical Modelling
for Fine-Grained Sketch Based Image Retrieval

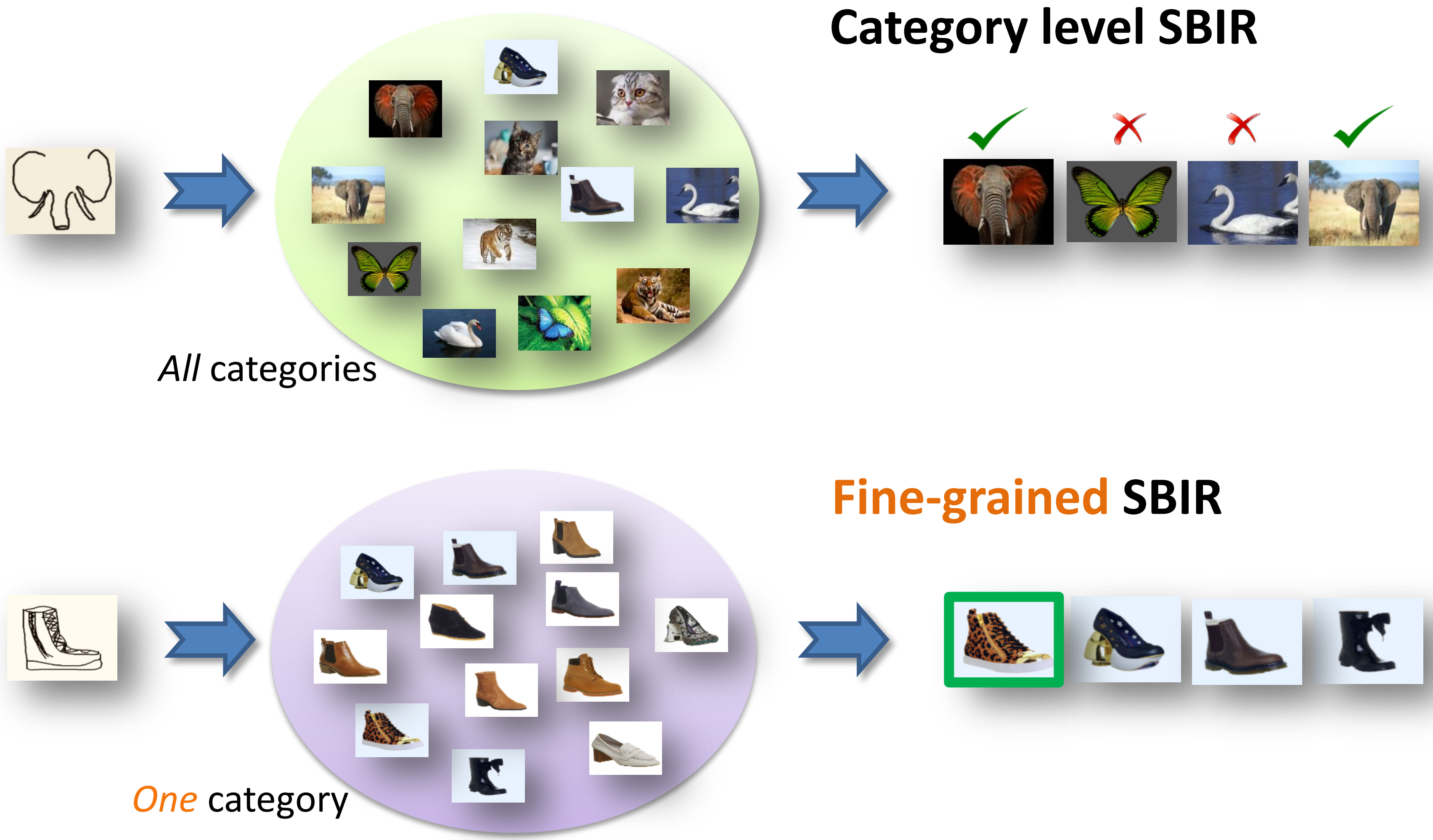




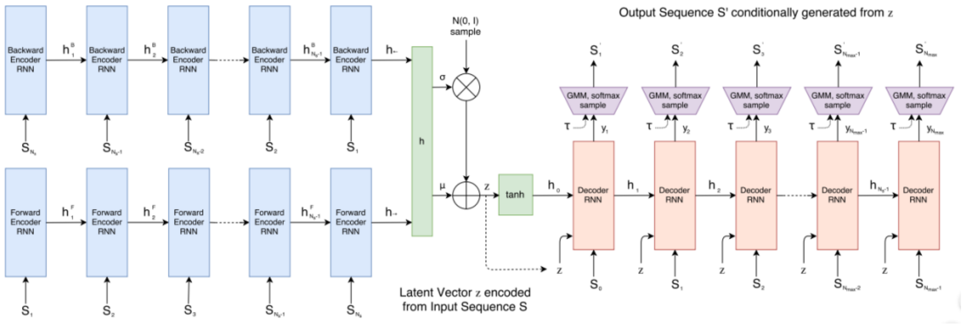
Aneeshan Sain [1,2]   Ayan Kumar Bhunia [1]   Yongxin Yang [1,2]   Tao Xiang [1,2]   Yi-Zhe Song [1,2]

[1] SketchX Lab, CVSSP, University of Surrey, UK
[2] iFlyTek-Surrey Joint Research Centre on Artificial Intelligence
http://sketchx.ai

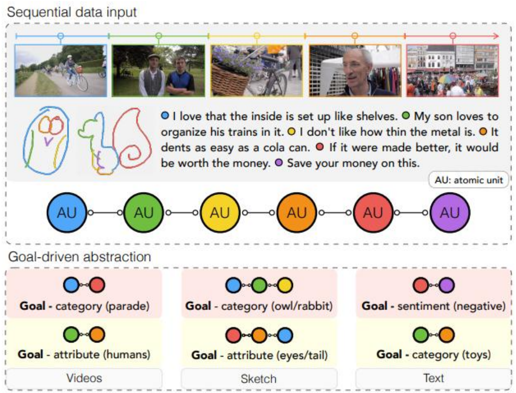
SketchX   BMVC 2020 MANCHESTER   CVSSP Centre for Vision, Speech and Signal Processing   UNIVERSITY OF SURREY


Category level SBIR
All categories
Fine-grained SBIR
One category

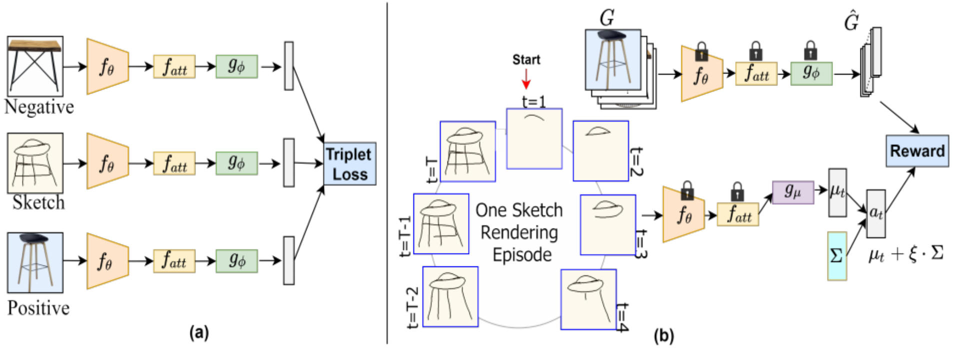
Explored traits in sketches



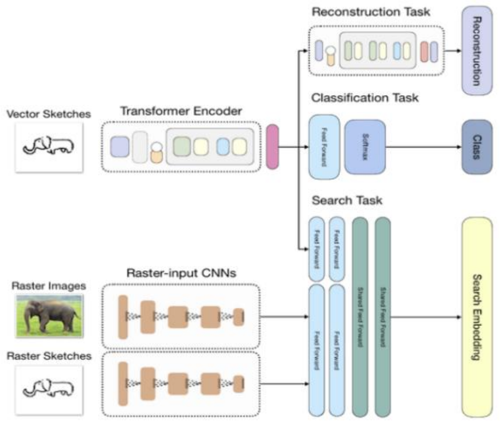
Sequential[1]



Abstract[2]



Stroke-wise[3]



Combined[4]

[1] David Ha and Douglas Eck. A neural representation of sketch drawings. In ICLR, 2018.
[2] Umar Riaz Muhammad, Yongxin Yang, Timothy Hospedales, Tao Xiang, and Yi-Zhe Song. Goal-driven sequential data abstraction. In ICCV, 2019.
[3] Ayan Kumar Bhunia, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In CVPR, 2020
[4] Leo Sampaio, Ferraz Ribeiro, Tu Bui, John Collomosse, and Moacir Ponti. Sketchformer:Transformer-based representation for sketched structure. In CVPR, 2020

Motivation:

- **Extent of details** being sketched **can vary** from coarse to fine.
- Sketches are **hierarchical** in terms of extent of detail sketched.
- **Capturing hierarchical cross-modal correspondence** between a sketch and its matching photo would therefore improve retrieval accuracy.

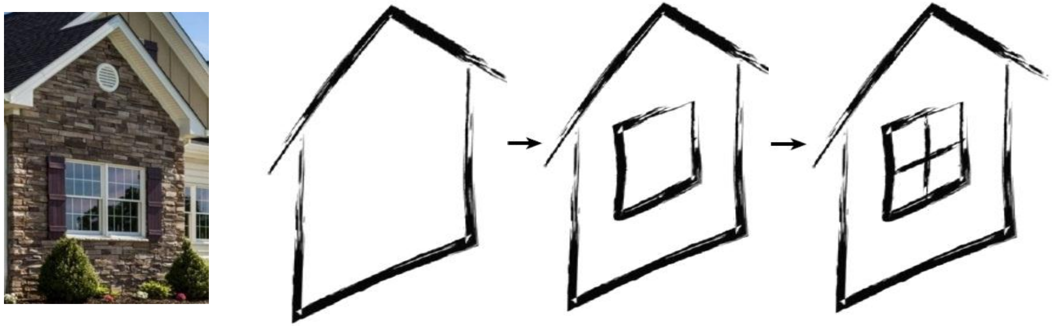


Photo Level 1 Level 2 Level 3

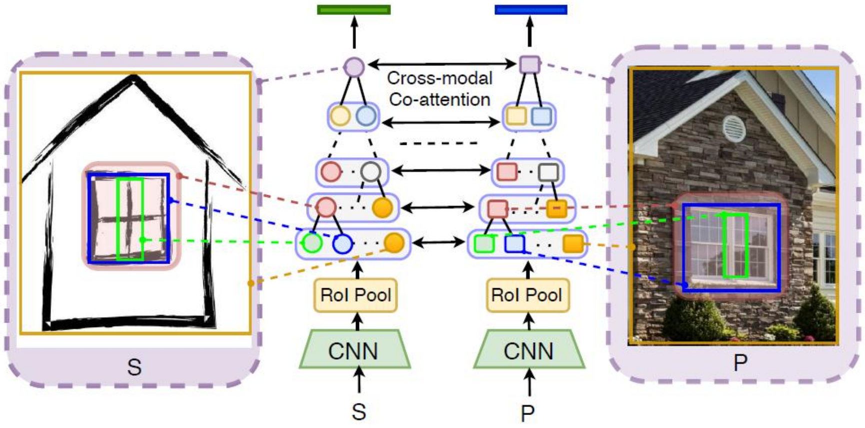
No matter the extent of detail drawn, we can fetch the right match!

Why Challenging?

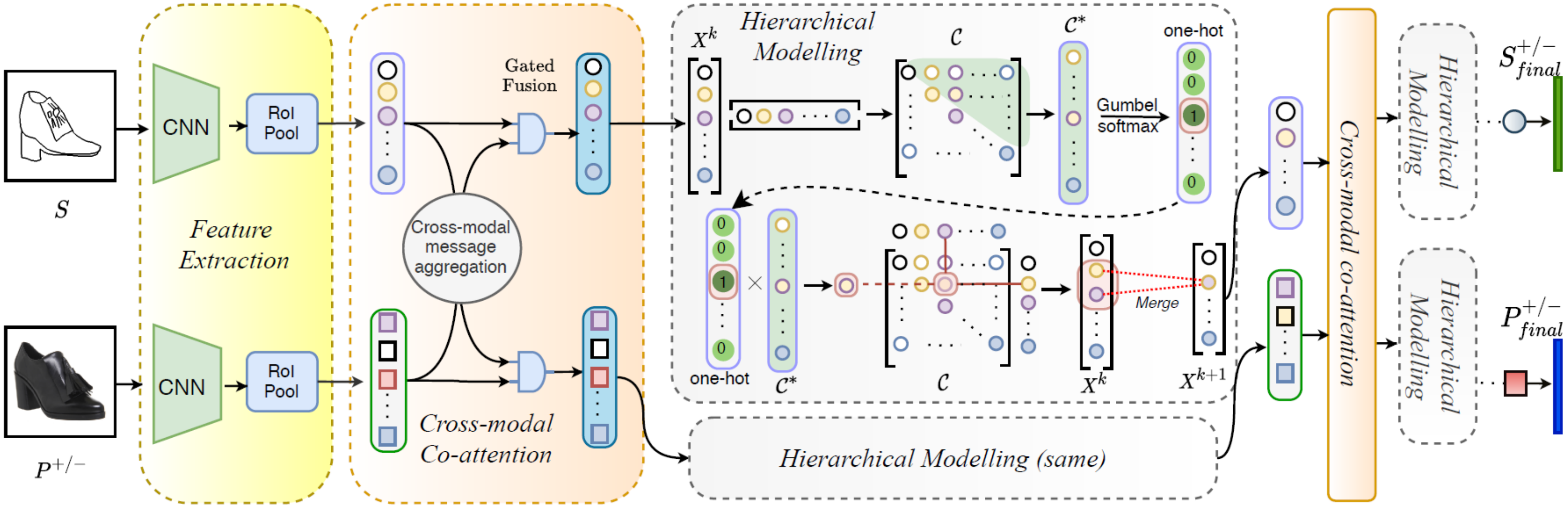
- Absence of predefined composition rules between sketch strokes.
- Unexplored cross-modal interaction between sketches and photos.

Contributions:

- End-to-end trainable framework that enables the **discovery of the underlying hierarchy** in a sketch.
- **Cross-modal co-attention module** to facilitate cross-modal hierarchy construction.
- Unique perspective of **utilising hierarchies for FG-SBIR**.



Overall Framework



Cross-modal Co-attention module

Aim : Enrich sketch representation with knowledge from its corresponding photo and vice-versa.

Method:

- Calculate a stroke-region affinity matrix

$$\mathbf{A} = \left(S_r . W_\psi^S\right) . \left(P_r . W_\psi^P\right)^T ; \ \mathbf{A} \in \mathbb{R}^{N_S \times N_P}$$

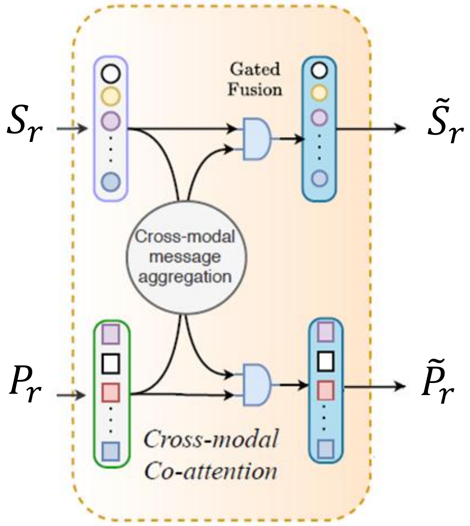
- Using co-attention, accumulate information to be fused.

$$\mathbf{A}_P^* = \text{softmax}\left(\frac{\mathbf{A}}{\sqrt{d_h}}\right) ; \quad P_r^S = \mathbf{A}_P^* . P_r , \quad P_r^S \in \mathbb{R}^{N_S \times d}$$

- *Adaptively fuse* the aggregated information onto its respective branch via a gating mechanism.

$$\mathrm{G}^S = \text{sigmoid}\left([S_r , P_r^S] . W_\mathrm{G}^S\right) , \quad W_\mathrm{G}^S \in \mathbb{R}^{2d \times d}$$

$$\tilde{S}_r = Z_S(\mathrm{G}^S \odot (S_r \oplus P_r^S)) \oplus S_r , \quad \tilde{S}_r \in \mathbb{R}^{N_S \times d}$$

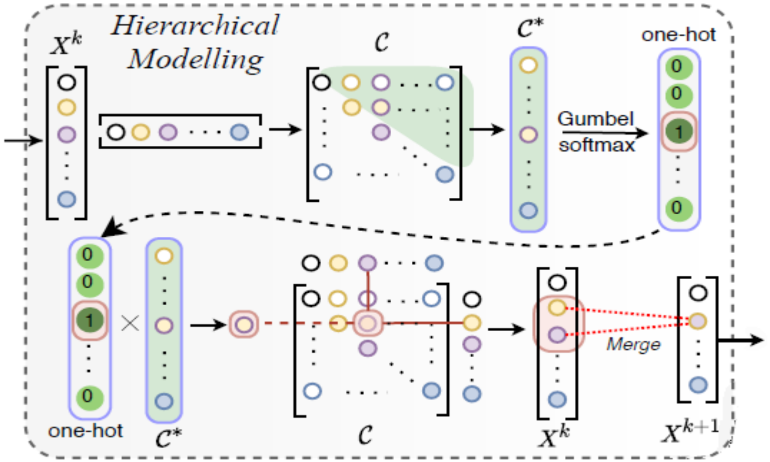


Hierarchical Modelling

Formulate intra-regional compatibility matrix

$$\mathbf{C} = \left(X^k . W_\phi^C\right) . \left(X^k . W_\phi^C\right)^T, \mathbf{C} \in \mathbb{R}^{N^k \times N^k}$$

$$\mathbf{C}^* = \text{flatten}\left(\text{UpTri}(\mathbf{C})\right)$$



Modeling **discrete decision** via **Gumbel-softmax** [1] distribution

$$q_i = \frac{\exp\left(\frac{\log \pi_i + g_i}{\tau}\right)}{\sum_{j=1}^{H^k} \exp\left(\frac{\log \pi_j + g_j}{\tau}\right)}$$

$$q^{ST} = \left(q_1^{ST}, q_2^{ST}, \ldots q_{H^k}^{ST}\right); \quad q_i^{ST} = 1_{[i = argmax_j(q_j)]}$$

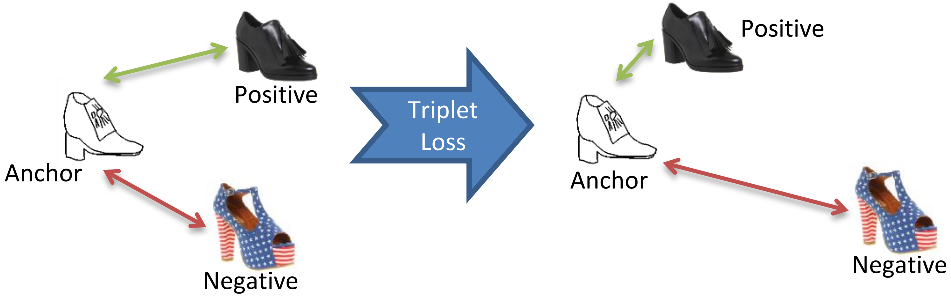
Fusion: $\hat{x}_{a,b} = \text{ReLU}\left(W_\phi^F . [x_a, x_b]\right), \quad W_\phi^F \in \mathbb{R}^{d \times 2d}$

Updation: $X^{k+1} := X^k - \{x_a, x_b\} + \{\hat{x}_{a,b}\}; \; x_a, x_b \in X^k$

[1] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel softmax. In ICLR, 2017.

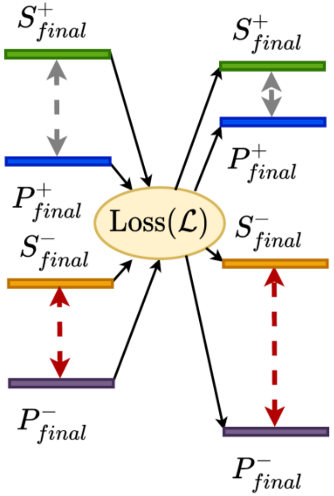
Learning Objective

Triplet Loss : $\mathcal{L}_\theta(s, p^+, p^-) = \max\{0, \Delta + D(f_\theta(s), f_\theta(p^+)) - D(f_\theta(s), f_\theta(p^-))\}$ [1]



Our formulation:

$$\mathcal{L}(S^+_{final}, S^-_{final}, P^+_{final}, P^-_{final}) = \max\{0, \Delta + D(S^+_{final}, P^-_{final}) - D(S^-_{final}, P^-_{final})\}$$



[1] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In CVPR, 2016.

Experiments

- **Datasets:** QMUL-Shoe-V2, QMUL-Chair-V2, SWIRE
- **Evaluation Metric:**
  - top-1 accuracy (acc@1), top-10 accuracy (acc@10)
- **Competitors**:
  - Contemporary state-of-the-arts:
    - Triplet-SN [1]
    - Triplet-Attn-SN [2]
    - SWIRE [3]
  - Baselines:
    - *B-Siamese* : Siamese triplet network with Inception V3 backbone.
    - *B-Gated-Siamese*: Network involving paired embedding by employing a matching gate [4].
    - B-Localised-Coattn: Framework with paired embedding having interaction between local photo-sketch sub-regions, without hierarchy.
    - B-Graph-Hierarchy : A framework modelling a graph-based method inspired by DIFFPOOL [5].
  - Other contemporary SBIR pipelines:
    - SketchFormer-variant : Based on Sketch-former architecture [6].
    - SketchBERT-variant: Based on Sketch-BERT architecture [7].

[1] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In CVPR, 2016.
[2] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In ICCV, 2017.
[3] Forrest Huang, John F Canny, and Jeffrey Nichols. Swire: Sketch-based user interface retrieval. In CHI, 2019.
[4] Rahul Rama Varior, Mrinal Haloi, and GangWang. Gated siamese convolutional neural network architecture for human re-identification. In ECCV, 2016
[5] Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. In NeurIPS, 2018
[6] Leo Sampaio Ferraz Ribeiro, Tu Bui, John Collomosse, and Moacir Ponti. Sketchformer: Transformer-based representation for sketched structure. In CVPR, 2020.
[7] Hangyu Lin, Yanwei Fu, Yu-Gang Jiang, and Xiangyang Xue. Sketch-bert: Learning sketch bidirectional encoder representation from transformers by self-supervised learning of sketch gestalt. In CVPR, 2020

Performance Analysis

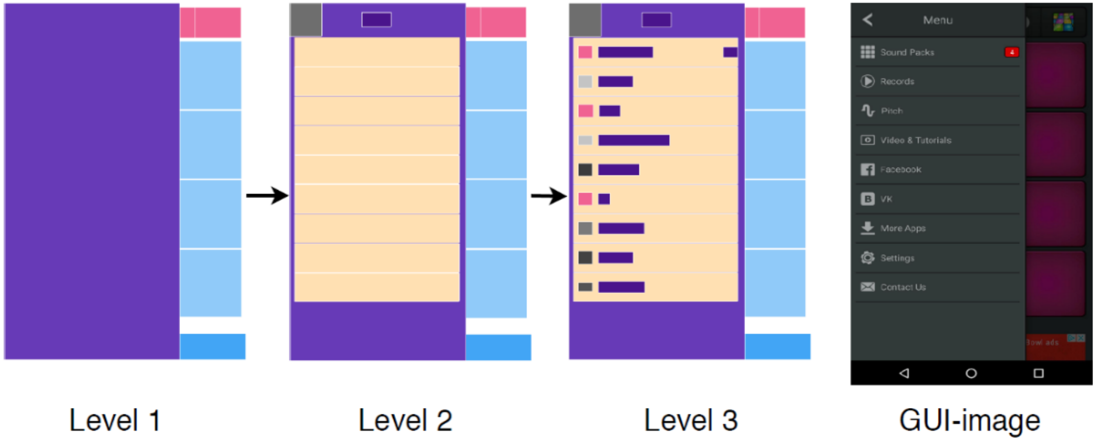
Methods		Chair-V2		Shoe-V2		SWIRE	
		acc.@1	acc.@10	acc.@1	acc.@10	acc.@1	acc.@10
State-of-the-arts	Triplet-SN	45.65	84.24	28.71	71.56	-	-
	Triplet-Attn-SN	56.54	88.15	31.74	74.78	-	-
	SWIRE	-	-	-	-	15.90	60.90
Baselines	B-Siamese	49.54	85.98	30.96	72.54	54.21	82.15
	B-Gated-Siamese	53.08	86.34	32.65	74.24	62.12	85.65
	B-Localised-Coattn	55.24	88.21	33.21	77.83	65.48	88.65
	B-Graph-Hierarchy	58.22	89.97	34.05	79.54	66.18	89.32
Others	SketchBERT-Variant	13.54	54.78	8.15	48.23	-	-
	SketchFormer-Variant	32.54	84.82	26.21	65.34	-	-
	Proposed	**62.45**	**90.71**	**36.27**	**80.65**	**67.23**	**90.11**

Ablation Study

Is hierarchy useful for FG-SBIR?

- Design elements in GUIs have a hierarchy defined by **containment**.
- Larger rectangular boxes encompass smaller ones (buttons) within.

- If hierarchy is at all useful, it should be **most helpful** in sketch based GUI image retrieval task, as there exists a pre-defined rule here.



Layout order in a GUI

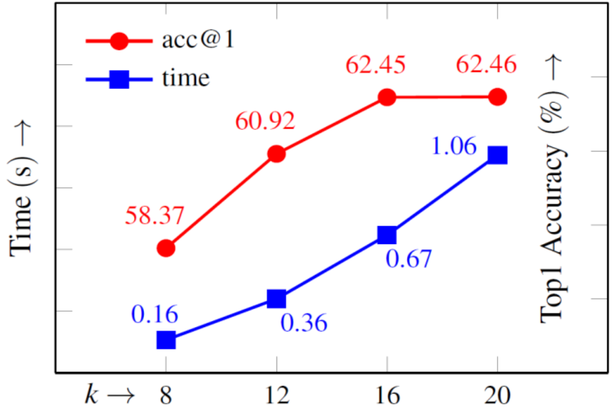
Further Analysis

Table 2: Ablative Study (acc.@1)

Methods	Chair-V2	Shoe-V2	SWIRE
Explicit Hierarchy	-	-	**71.54**
w/o Localised-Coattn	51.85	31.82	60.32
w/o Hierarchy	55.24	33.21	65.48
Sketch-coarse	47.64	31.83	51.26
Sketch-coarse++	42.33	24.11	45.33
Proposed	**62.45**	**36.27**	67.23

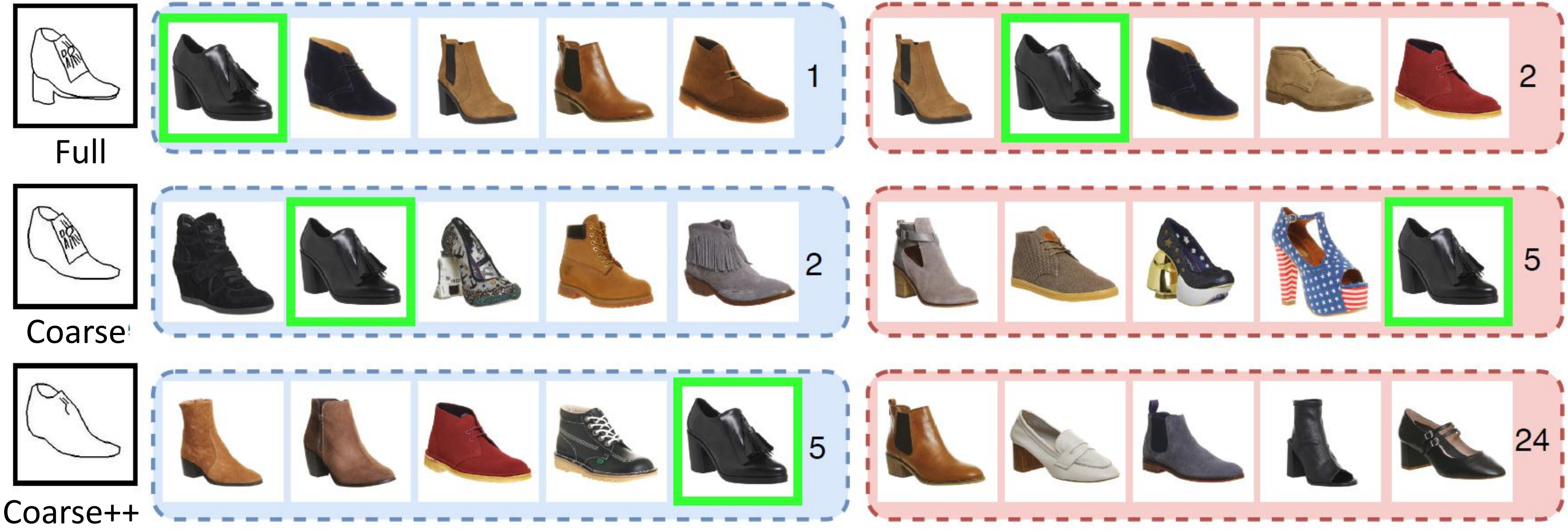
Table 3: Retrieval performance on varying extent of detail (Acc@10)

		ChairV2	ShoeV2
B-Siamese	Sketch-coarse	75.32	62.68
	Sketch-coarse++	65.31	54.32
	Full sketch	85.98	72.54
Proposed Method	Sketch-coarse	87.58	77.23
	Sketch-coarse++	85.64	75.91
	Full sketch	90.71	80.65



Study on impact of number of regions chosen for feature extraction in image branch

Qualitative Results



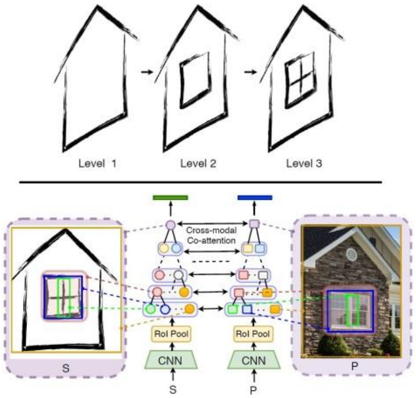
QMUL Shoe-V2 Dataset

Qualitative Results (contd.)

SWIRE Dataset

SketchX

http://sketchx.ai



https://aneeshan95.github.io/Cross-modal_Hierarchy_FGSBIR