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

Supplementary Material

Abhra Chaudhuri[1]
ac1151@exeter.ac.uk
Massimiliano Mancini[2]
Yanbei Chen[2]
Zeynep Akata[2,3,4]
Anjan Dutta[5]

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

Appendix

1 Further Experimental Details and Findings

1.1 Platform Details

We implement our XModalViT model using the PyTorch [3] deep learning framework, on an Ubuntu 20.04 workstation with a single Nvidia GeForce RTX 3090 GPU, an 8-core Intel Xeon processor and 32 GBs of RAM. Since we are using a fixed-size queue to store XMA representations, we do not have a dependency on batch-size for the purpose of negative sampling as part of our contrastive learning phase, which enables us to train both the teacher and the students on a single GPU.

1.2 Dataset Statistics

Dataset	#Classes	#Photos	#Sketches	#Sketches/Photos	Test Fraction
QMUL-Shoe-V2	–	2000	6648	2 to 4	0.1
QMUL-Chair-V2	–	400	1275	2 to 4	0.1
Sketchy	125	12,500	75,471	5 to 9	0.1

Table 1: Details of the datasets used for experimental evaluation.

1.3 Effects of Varying the Queue Sizes in XAQC

We vary the XAQC queue sizes as 2^m, where m $\in \{1, 2, ...\lfloor \log_2 N \rfloor\}$, and N is the total number of datapoints. The observed trend has been graphically depicted in Figure 1, where

it can be seen that the accuracy is minimum at $m = 1$, when the objective is equivalent to InfoNCE, which only performs reasonably with larger batch-sizes. However, the accuracy begins to saturate near its maximum at $m = \lfloor \log_2 N \rfloor$.

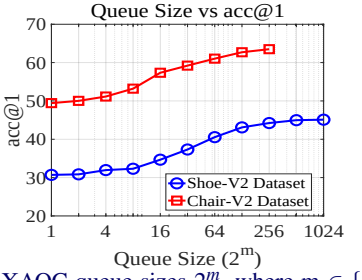


Figure 1: Acc@1 with varying XAQC queue sizes 2^m, where m $\in \{1, 2, ..., \lfloor \log_2 N \rfloor\}$, and N is the total number of datapoints.

1.4 ViT+CNN Backbones

While keeping the ViT-B teacher backbones, we switched the student backbones to ResNet18 [1] and obtained acc@1 of 44.9% and 63.21% on the Shoe-V2 and Chair-V2 datasets respectively. However, with the InceptionV3 [2] network as the student backbones, the acc@1 on the same datasets drop to 42.5% and 62.07% respectively. This goes on to show that, once our ViT-Base networks as the teacher encoders learn the fused cross-modal attention representations, the correct inductive biases in CNNs can be used to approximate them. However, for the teacher network, there is no straightforward way to formulate the XMA operator that would make sense over the space of CNN feature maps. The cross-modal interaction between the global class-token of one modality and the local patch embeddings of the other is something that can be more naturally defined for Vision Transformers.

2 Pseudocodes

This section provides algorithmic pseudocodes for the core components of the XModalViT framework.

Notations like tuple assignment:

$$(a, b) = (b, a) \tag{1}$$

are inspired by Python's syntax. The construct "with no gradient" signifies that the output of the operation performed within the block is treated like a constant, and its computation does not affect the gradient of any of the learnable parameters that might have been used for that purpose.

2.1 Modality Fusion Network

The objective of the MODALITY-FUSION-NETWORK is to train the cross-modal attention (XMA) based teacher network, Γ, that fuses information across the photo and the sketch modalities. Algorithm 1 provides a pseudocode for the same.

Algorithm 1: MODALITY-FUSION-NETWORK: Train the cross-modal attention (XMA) based teacher network, Γ, that fuses information across the photo and the sketch modalities.

Input : A set of photos $\mathcal{P}$ and corresponding sketches $\mathcal{S}$; Cross-attention queue size k; Learning rate η; Number of epochs N

Output: Cross-Modal Fusion Network (Teacher): Γ

1 $\mathcal{Q} \longleftarrow$ Randomly initialized queue of size k
2 **for** *epoch* $\leftarrow 1$ **to** N **do**
3 $\quad p \sim \mathcal{P},\ (s_1, s_2) \sim \mathcal{S} \mid p \leftrightarrow (s_1, s_2)$
4 $\quad (\mathbf{x}_p^1, \mathbf{x}_s^1) \longleftarrow \Gamma(p, s_1)$
5 $\quad$ `// Outputs are treated as constants`
6 $\quad$ with no gradient:
7 $\quad\quad (\mathbf{x}_p^2, \mathbf{x}_s^2) \longleftarrow \Gamma(p, s_2)$
8 $\quad \mathcal{L}_{\text{teacher}} \longleftarrow \textbf{XAQC}(\mathbf{x}_p^1, \mathbf{x}_s^2, \mathcal{Q})$
9 $\quad \mathcal{Q} \longleftarrow \mathcal{Q}.\text{enqueue}(\mathbf{x}_s^2)$
10 $\quad \Gamma \longleftarrow \Gamma - \eta \nabla_\Gamma \mathcal{L}_{\text{teacher}}$

$\mathcal{Q}$ is initialized as a queue of size k containing random vectors, which is later used for storing XMA-sketch representations. We sample a photo p and 2 of its corresponding sketches s_1 and s_2. In line-5, (p, s_1) is propagated through the teacher network for obtaining its XMA-photo embedding, $\mathbf{x}_p^1$. The XMA-sketch embedding, $\mathbf{x}_s^2$, for (p, s_1) is computed in this manner. We then compute the XAQC loss for the teacher by treating $\mathbf{x}_s^2$ as a soft-target for $\mathbf{x}_p^1$ and all the representations in $\mathbf{x}_s^1$ as negatives. We then enqueue $\mathbf{x}_s^1$ into $\mathcal{Q}$ in line-9. We finally update the teacher network with the gradient of the XAQC loss with respect to its parameters.

2.2 Cross-Modal Knowledge Distillation

The objective of cross-modal knowledge distillation is to decouple the input-space of the modality fusion network (teacher), Γ, into independent, modality-specific encoders, ξ_{photo} and ξ_{sketch}. Algorithm 2 describes the process in the form of a pseudocode.

We start by sampling 3 corresponding photo-sketch pairs $(p_1, s_1), (p_2, s_2)$ and (p_3, s_3). We propagate these pairs through the modality fusion network (teacher), Γ, to obtain their XMA-photo and XMA-sketch embeddings. The photos and the sketches are then separately encoded via the photo and the sketch students respectively, to obtain the approximate versions of their XMA representations, *i.e.*, $\mathbf{z}_{p_i}$ and $\mathbf{z}_{s_i}$. With the objective of aligning these z_* representations with the true XMA representations, x_*, obtained from the teacher, we minimize the contrastive XAQC loss between the two (lines 11 and 12). We also aim to preserve the geometry of the teacher's representation space in that of the students' by distilling distance and angular relationships between arbitrary k-tuples of datapoints. This process of cross-modal relation distillation (XMRD) is depicted in lines 16-18. m, here, has been introduced for the purpose of conciseness, serving as an abstract notation for modality, *i.e.*, standing for both photos and sketches. ψ_1 and ψ_2 are the distance and angle relation functions respectively, and δ is the Huber loss, as described in the main text. The total student loss is computed as the sum of the losses from the individual students and the XMRD loss (weighted by a balancing factor of λ). We finally update the individual students by the gradient of the total student loss with respect to the weights of the corresponding student

Algorithm 2: CROSS-MODAL-KNOWLEDGE-DISTILLATION: Decouple the domain of Γ by transferring its representations to independent encoders, ξ_{photo} and ξ_{sketch}.

Input : A set of photos $\mathcal{P}$ and corresponding sketches $\mathcal{S}$; Teacher network Γ; Cross-attention queue size k; Learning rate η; Number of epochs N

Output: Student Networks - Photo encoder ξ_{photo} and sketch encoder ξ_{sketch}

$\mathcal{Q}_{\mathcal{P}}, \mathcal{Q}_{\mathcal{S}} \longleftarrow$ Randomly initialized queues of size k

for *epoch* $\leftarrow 1$ **to** N **do**

  $p_1 \sim \mathcal{P},\ s_1 \sim \mathcal{S} \mid p_1 \leftrightarrow s_1$

  $p_2, p_3 \sim \mathcal{P},\ s_2, s_3 \sim \mathcal{S} \mid p_2 \leftrightarrow s_2, p_3 \leftrightarrow s_3$

  `// Outputs are treated as constants`

  with no gradient:

  for $i \leftarrow 1$ **to** 3 **do**

    $(\mathbf{x}_{p_i}, \mathbf{x}_{s_i}) \longleftarrow \Gamma(p_i, s_i)$

  for $i \leftarrow 1$ **to** 3 **do**

    $(\mathbf{z}_{p_i}, \mathbf{z}_{s_i}) \longleftarrow \xi_{\text{photo}}(p_i), \xi_{\text{sketch}}(s_i)$

  $\mathcal{L}_{\text{XAQC}}^{\text{photo-student}} \longleftarrow \textbf{XAQC}(\mathbf{x}_{p_1}, \mathbf{z}_{p_1}, \mathcal{Q}_{\mathcal{P}})$

  $\mathcal{L}_{\text{XAQC}}^{\text{sketch-student}} \longleftarrow \textbf{XAQC}(\mathbf{x}_{s_1}, \mathbf{z}_{s_1}, \mathcal{Q}_{\mathcal{S}})$

  $\mathcal{Q}_{\mathcal{P}} \longleftarrow \mathcal{Q}_{\mathcal{P}}.\text{enqueue}(\mathbf{x}_{p_1})$

  $\mathcal{Q}_{\mathcal{S}} \longleftarrow \mathcal{Q}_{\mathcal{S}}.\text{enqueue}(\mathbf{x}_{s_1})$

  // $m \in \{p, m\}$

  $\pi_m^{\text{teacher}} \longleftarrow \psi_1(\mathbf{x}_{m_1}, \mathbf{x}_{m_2}) + \psi_2(\mathbf{x}_{m_1}, \mathbf{x}_{m_2}, \mathbf{x}_{m_3})$

  $\pi_m^{\text{student}} \longleftarrow \psi_1(\mathbf{z}_{m_1}, \mathbf{z}_{m_2}) + \psi_2(\mathbf{z}_{m_1}, \mathbf{z}_{m_2}, \mathbf{z}_{m_3})$

  $\mathcal{L}_{\text{XMRD}} \longleftarrow \delta(\pi_p^{\text{teacher}}, \pi_p^{\text{student}}) + \delta(\pi_s^{\text{teacher}}, \pi_s^{\text{student}})$

  $\mathcal{L}_{\text{student}} \longleftarrow \mathcal{L}_{\text{XAQC}}^{\text{sketch-student}} + \mathcal{L}_{\text{XAQC}}^{\text{photo-student}} + \lambda \cdot \mathcal{L}_{\text{XMRD}}$

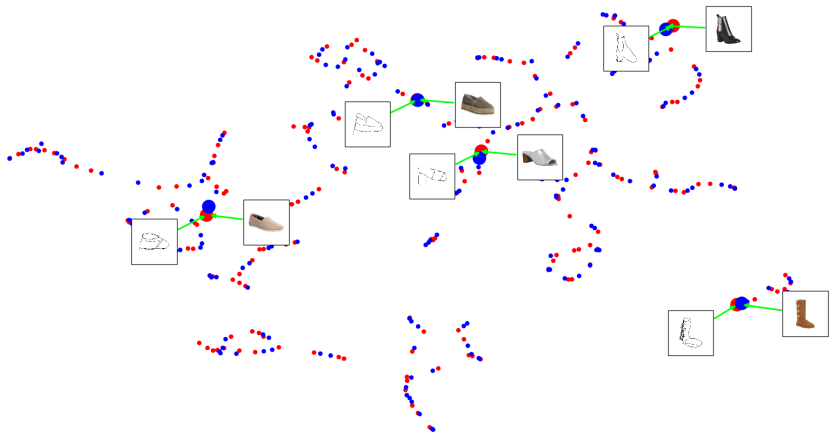
  $\xi_{\text{photo}} \longleftarrow \xi_{\text{photo}} - \nabla_{\xi_{\text{photo}}} \mathcal{L}_{\text{student}}$

  $\xi_{\text{sketch}} \longleftarrow \xi_{\text{sketch}} - \nabla_{\xi_{\text{sketch}}} \mathcal{L}_{\text{student}}$

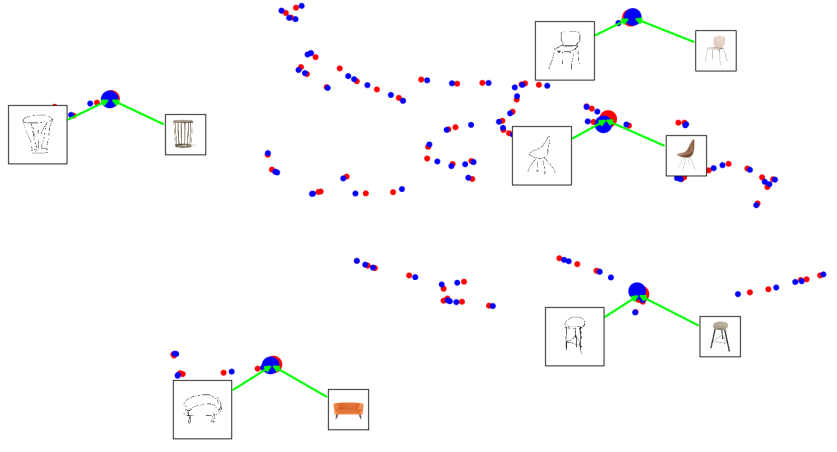
encoders.

3 Embedding Visualizations

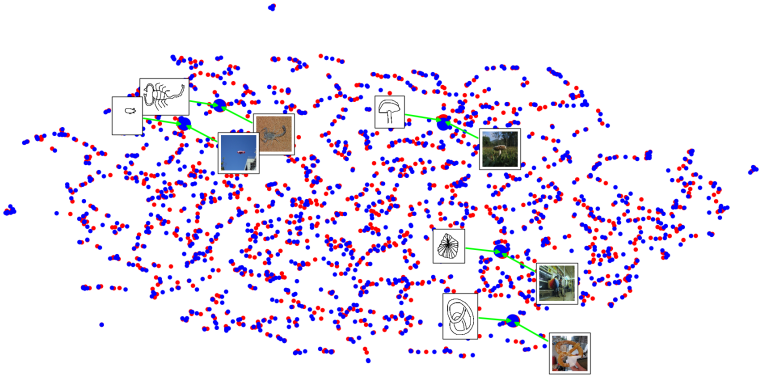
Figure 2 shows the test-set embeddings of the photo (red) and sketch (blue) student encoders, projected onto a 2-dimensional space via UMAP [2]. The layout of the datapoints across the two modalities can be seen as being very similar, indicating that both the encoders have learned to model the distribution of the underlying shared abstract concept. As a result of this, and by the virtue of the instance-discriminative XAQC loss, photo-sketch pairs of the same instance get mapped close to each other, a phenomenon that has also been depicted in the visualizations.



(a)



(b)



(c)

Figure 2: UMAP [] visualizations of photo-student (red) and sketch-student (blue) embeddings on the (a) QMUL-Shoe-V2 (b) QMUL-Chair-V2 and (c) Sketchy datasets.

4 Retrieval Results

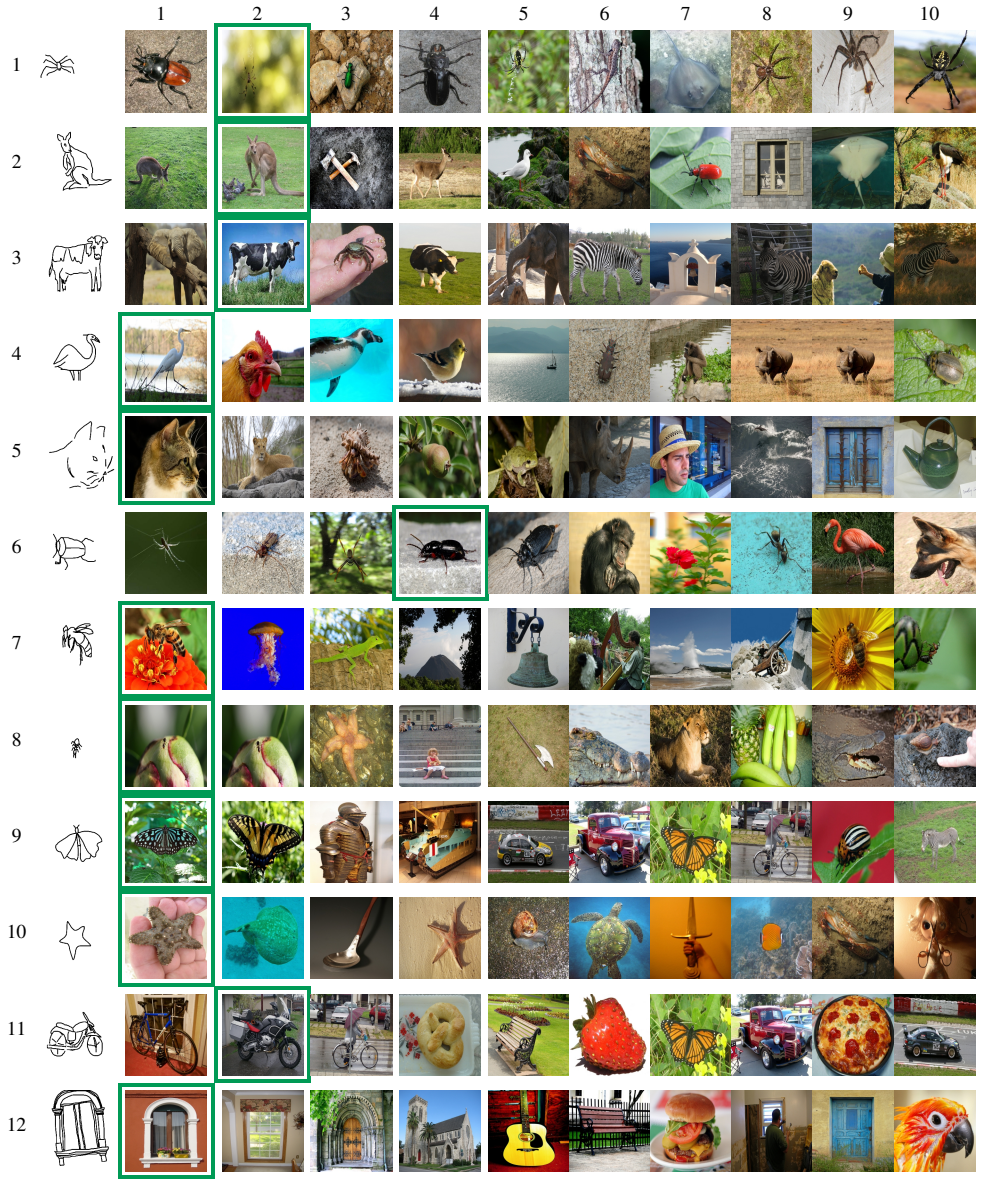


Figure 3: Qualitative fine-grained SBIR results on Sketchy dataset.

4.1 Sketchy Dataset

Apart from the instance-discriminative fine-grained features, the network learns attribute information such as classes with features that are closely related visually (rows 1, 3, 6, 11

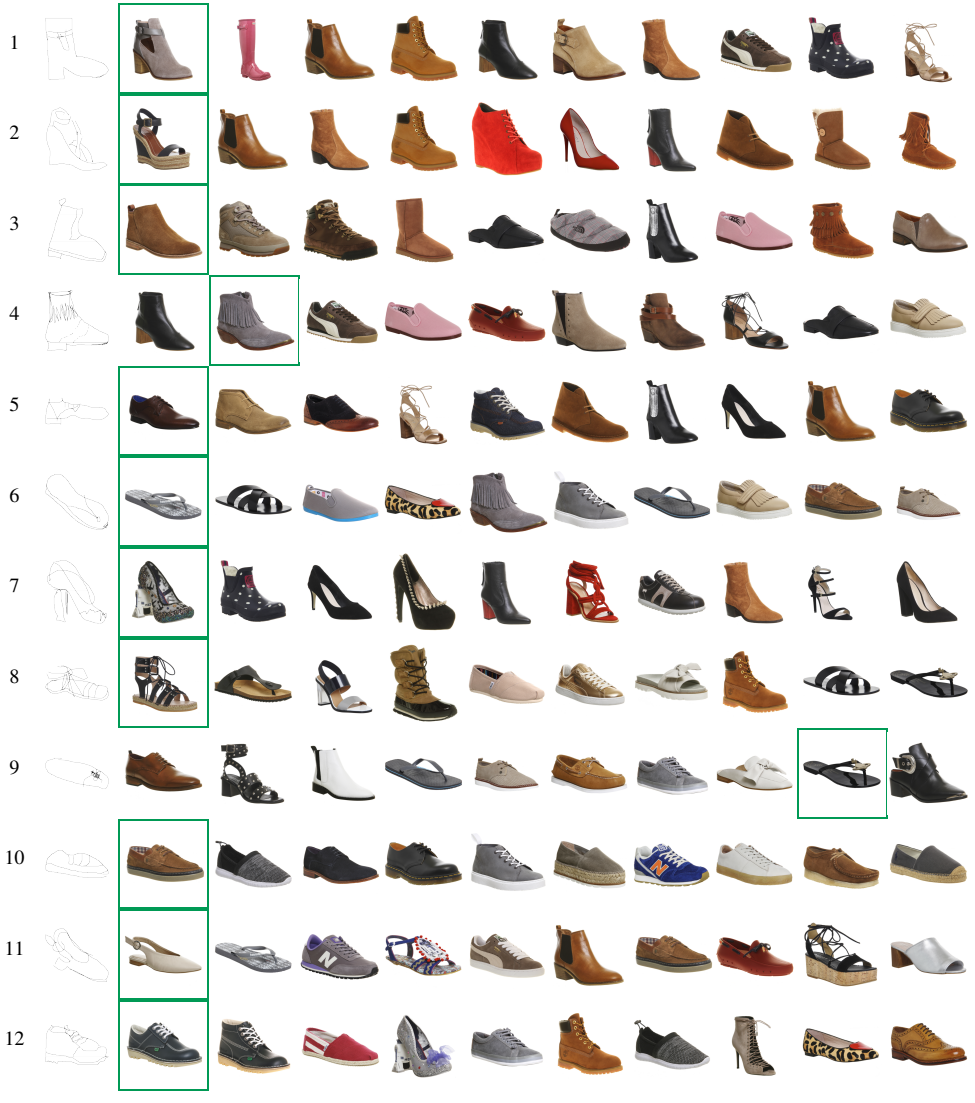


Figure 4: Qualitative fine-grained SBIR results on Shoe-V2 dataset.

of Figure 3), object orientation and geometry (rows 8, 5, 4, 10 Figure 3), and even naturally occurring relationships (rows 5.1 and 5.2 Figure 3).

4.2 QMUL-Shoe-V2 Dataset

Most instances with a sufficient number of fine-grained instance-discriminative features can be seen to appear in the top 1. However, for the ones that do get demoted in rank, (row 4 of Figure 4), are preceded by an instance that have noticeable features in common that could cause confusion. Some false positive top-1 results might occur as a side-effect of modality fusion (row 10 of Figure 4), where the embedding space also captures the texture information from the photo modality, which may not always be necessarily relevant (causing a shoe with a similar shiny texture as the ground-truth being returned as the first retrieval result in row

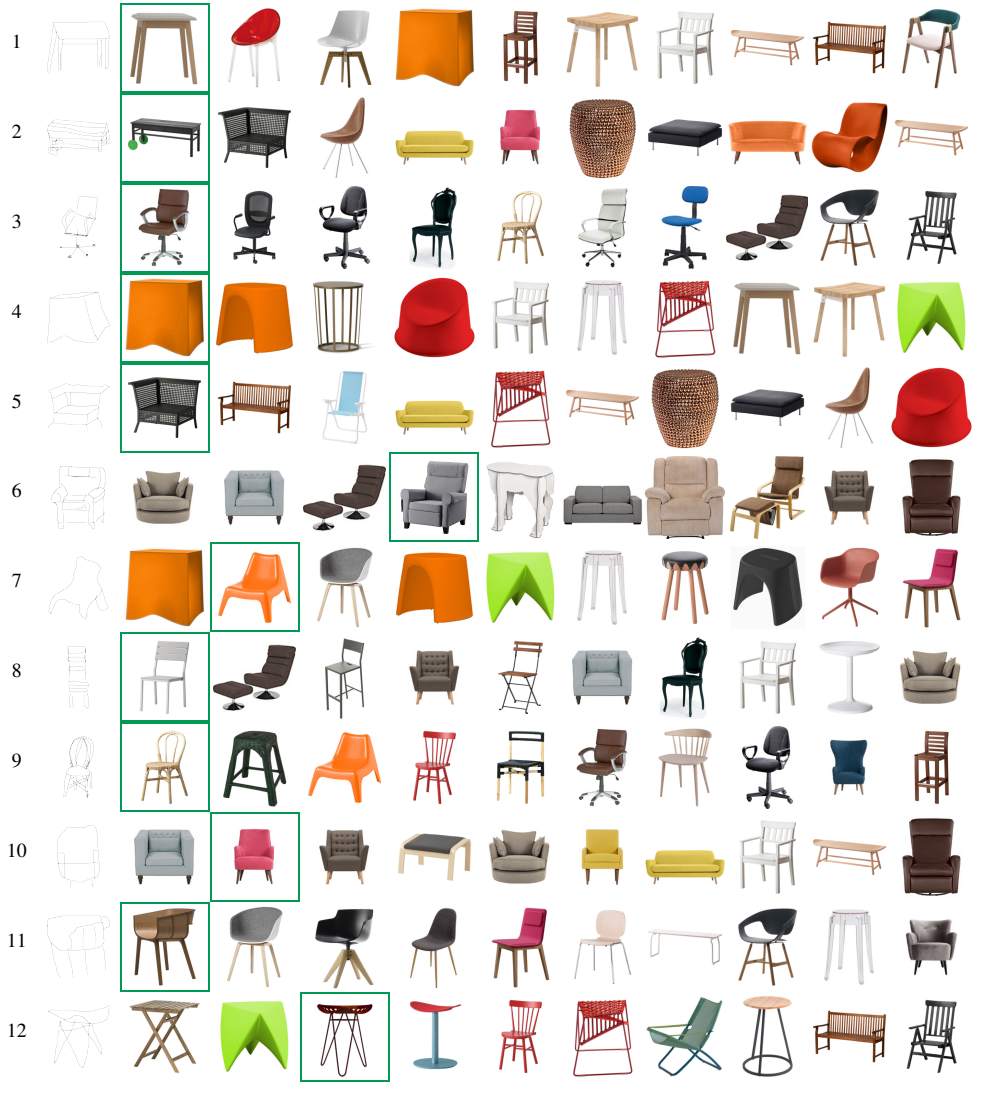


Figure 5: Qualitative fine-grained SBIR results on Chair-V2 dataset.

10 of Figure 4).

4.3 QMUL-Chair-V2 Dataset

The modality fusion operator is able to capture cross-modal information such as color and texture, which are beyond the geometric structure of sketches (rows 4, 5, 7 of Figure 5), while being able to able to attend to the modality-native attributes, such as structural geometry and orientation (rows 3, 8, 9, 10, 11 of Figure 5).

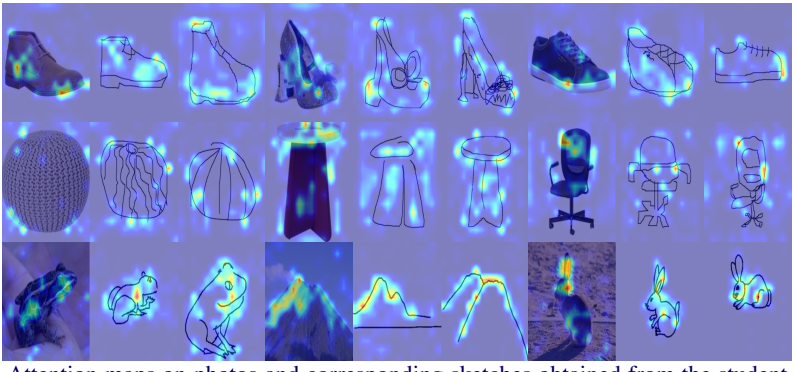


Figure 6: Attention maps on photos and corresponding sketches obtained from the student encoders (examples from the Shoe-V2, Chair-V2 and Sketchy datasets).

5 Attention Maps

Figure 6 depicts attention maps for photos and two of their corresponding sketches obtained from the photo and the sketch students respectively. Both the networks can be seen to generally focus on the same object localities irrespective of the modality. Also, within the sketch modality, the regions attended to by the sketch encoder are quite stable, indicating that the network has learned to focus more on the structural information and is robust to the variations in sketching style.

References

[1] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016.

[2] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. UMAP: Uniform Manifold Approximation and Projection. *JOSS*, 2018.

[3] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPSW*, 2017.

[4] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.