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

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

Abstract

Representation learning for sketch-based image retrieval has mostly been tackled by learning embeddings that discard modality-specific information. As instances from different modalities can often provide complementary information describing the underlying concept, we propose a cross-attention framework for Vision Transformers (XModalViT) that fuses modality-specific information instead of discarding them. Our framework first maps paired datapoints from the individual photo and sketch modalities to fused representations that unify information from both modalities. We then decouple the input space of the aforementioned modality fusion network into independent encoders of the individual modalities via contrastive and relational cross-modal knowledge distillation. Such encoders can then be applied to downstream tasks like cross-modal retrieval. We demonstrate the expressive capacity of the learned representations by performing a wide range of experiments and achieving state-of-the-art results on three fine-grained sketch-based image retrieval benchmarks: Shoe-V2, Chair-V2 and Sketchy. Implementation is available at https://github.com/abhrac/xmodal-vit.

1 Introduction

Fine-grained sketch-based image retrieval (FG-SBIR) [2, 3] is a particular setting of SBIR [2, 3, 8, 10, 11, 25, 28] that aims to retrieve a *specific* photo based on a query sketch. Classical metric-learning based literature on FG-SBIR directly employs a contrastive learning strategy to estimate the modality-invariant component in a sketch-photo pair, optimizing an objective that aligns embeddings of similar photo-sketch pairs closer to and dissimilar pairs away from each other [25, 35, 36, 52]. While such an approach only takes into account the shared mutual semantics between the two modalities, it does not consider how information specific to a modality might manifest itself upon being translated to the other modality. This is particularly important for free-hand sketches where the modality-gap is large due to imprecise depiction of object attributes. For example, as depicted in Figure 1, a region with a dark shade, may be distinguished by the sketcher, from one with a lighter shade, using a sparse collection of zigzag lines spanning the dark region (highlighted with a blue circle). Such nuances get eliminated during the modality filtering stage, but should ideally be preserved in an optimal fine-grained representation.

To model the large sketch-photo modality gap, we start with viewing sketches and photos as instantiations of an abstract conceptual representation derived from the interrelationships between local, spatial regions of the underlying object and their higher order interactions. The goal then is to derive the aforementioned representation by modeling the interrelationships and higher order interactions between the different localities across the two modalities. Hence, we design our learning objective so as to unify the instance-discriminative modality-specific information into the encoded representations rather than discard them. By accounting for such variations across modalities, a representation would correspond to the complete abstract higher-order concept of which photos and sketches are different manifestations. As elaborated above through the example in Figure 1, in the cross-modal setting, a sketch-photo pair has the following three kinds of 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). Different from existing methods, we simultaneously preserve (1) and (2) while discarding (3) via a Modality Fusion Network and thus, are able to substitute modality alignment with a *modality-fused instance alignment* by minimizing a cross-modal contrastive loss. We decouple the fused space into independent, modality-specific encoders via a novel cross-modal knowledge distillation strategy, thereby avoiding the need for performing the expensive cross-modal fusion operation at runtime.

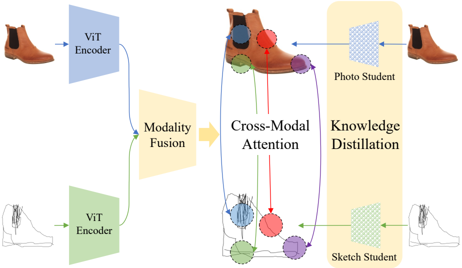

Figure 1: Our XModalViT retains semantically relevant, modality-specific features by learning a fused representation space, while bypassing the expensive cross-attention computation at runtime via cross-modal knowledge distillation.

We encapsulate all of the above steps by designing the XModalViT framework centered on our novel cross-modal attention operation for Vision Transformers. In this paper, we make the following contributions: (1) A novel approach to the cross-modal visual representation learning task for FG-SBIR by designing a modality fusion operator for ViT based on the cross-attention mechanism, which unifies complementary information across modalities while being instance-discriminative at the same time. (2) A cross-modal distillation technique to train independent encoders that can leverage a modality-fused representation space, without having to perform a computationally expensive cross-attention. (3) State-of-the-art results from a wide range of experiments conducted on three benchmark datasets for the task of fine-grained SBIR, which further strengthens our claim in favor of preserving instance-discriminative, modality-specific information in the learned representations.

2 Related Work

Below we review the literature on FG-SBIR, Cross-Attention and Knowledge Distillation.

Fine-Grained Sketch-Based Image Retrieval: Early works on (FG-)SBIR [44] were mainly focused on hand-crafted features, such as gradient field HOG [15], deformable parts model [18], histogram of edge local orientations [31], learned key shapes [32], which were limited

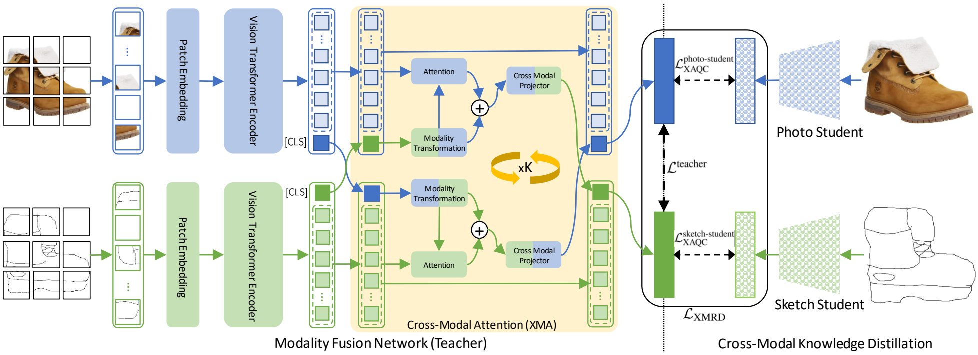
Patch Embedding
Vision Transformer Encoder
[CLS]
Attention
Modality Transformation
Cross Modal Projector
xK
Modality Transformation
Attention
Cross Modal Projector
[CLS]
Patch Embedding
Vision Transformer Encoder
Cross-Modal Attention (XMA)
Modality Fusion Network (Teacher)
$\mathcal{L}^{\text{photo-student}}_{\text{XAQC}}$
$\mathcal{L}^{\text{teacher}}$
$\mathcal{L}^{\text{sketch-student}}_{\text{XAQC}}$
$\mathcal{L}_{\text{XMRD}}$
Photo Student
Sketch Student
Cross-Modal Knowledge Distillation

Figure 2: Our XModalViT framework: The modality fusion network (teacher) takes as input, a positive photo-sketch pair and computes XMA embeddings for the individual modalities. During the cross-modal knowledge distillation phase, the weights of the teacher are frozen and the independent student networks are tasked with mapping the same pair of datapoints to the corresponding XMA branch of the teacher, thereby decoupling the domain of the XMA operator.

by the domain gap between sketches and photos. To address this issue, deep FG-SBIR models employing a deep triplet network to learn a common embedding space were proposed in [35, 52], an idea that was extended by the introduction of spatial attention [36], self-supervised pre-training tasks [4, 25], or mining fine-grained local-features [48]. Generative models have also shown promising results [3, 23, 34], employing ideas like style-semantic disentanglement, cross-domain image synthesis, and reinforcement learning. To the best of our knowledge, we are the first to study the effects of fusing information across modalities on fine-grained representations for FG-SBIR.

Cross-Attention: Dosovitskiy *et al.* [9] introduced the idea of visual self-attention to be an effective strategy to learn local and global representations of an image. Cross-attention extends this idea to incorporate multiple embedding spaces, which can arise from different modalities [20, 22], and can be used to perform tasks like capturing relationships between sentence words and image regions [43], unified, modality-agnostic classification [21], or multi-scale feature learning [6]. In contrast, we design a cross-modal attention mechanism that facilitates the retention of semantically relevant, modality-specific features.

Knowledge Distillation: Knowledge Distillation [14] is the technique for transferring representations learned by one model (*teacher*) to another (*student*). Originally formulated as a KL-Divergence minimization problem between the teacher-student outputs [14], making the student equal to, or greater in size than the teacher [47], or maintaining geometric and relational invariants between the representation spaces of the two networks [26, 29, 49] provide improved approximation performance. The idea of knowledge distillation was adapted in domains like image-to-video person ReID [12, 27], as well as for retrieval tasks by minimizing various distance metrics between the teacher and the student embeddings [51], or by formulating the problem in contrastive learning terms [13, 40, 41, 46]. Our work presents a novel and significantly more challenging application domain for knowledge distillation, with the requirement of transferring fused cross-modal representations to independent, modality specific encoders.

3 The XModalViT Framework

Let $\mathcal{P}$ and $\mathcal{S}$ denote the sets of photos and sketches respectively. Consider a function $\mathcal{P} \rightarrow \mathcal{S}$, that takes a photo as input and returns the set of corresponding sketches, defined as $p_i \in \mathcal{P} \longmapsto \{s_{ij} \in \mathcal{S} \mid p_i \leftrightarrow s_{ij}\}$, where $\leftrightarrow$ denotes the correspondence relation between a photo-sketch pair. The FG-SBIR task is then to derive the function $g(s) = p_i,\ \forall s \in \{s_{ij}\}$ This can be achieved by projecting the photo-sketch pairs into a common dot-product space $\mathbb{X}$, *i.e.*, a cross-modal representation space, such that $\forall p, p' \in \mathcal{P}$, where $p \leftrightarrow s$ and $p \neq p'$, we have, $\xi_{\text{photo}}(p) \cdot \xi_{\text{sketch}}(s) > \xi_{\text{photo}}(p') \cdot \xi_{\text{sketch}}(s)$. The encoders ξ_{photo} and ξ_{sketch} represent the respective photo and sketch projection functions. Thus, given a query sketch s_q, the retrieval result would be: $\arg\max_{p \in \mathcal{P}} \xi_{\text{photo}}(p) \cdot \xi_{\text{sketch}}(s_q)$, *i.e.*, the photo $p \in \mathcal{P}$ that has the highest similarity with s_q in $\mathbb{X}$.

We organize the cross-modal representation learning into a 2-step process, which we term XModalViT. First, we train a Vision Transformer based modality fusion network as a *teacher* by taking photo-sketch pairs as inputs, to learn unified representations by fusing information from both the modalities. Next, we decouple the input space of this teacher network into independent, modality-specific encoders (*students*) via knowledge distillation [14]. Our end-to-end framework is graphically depicted in Figure 2. Pseudocodes for training our models are provided in the supplementary material.

3.1 Modality Fusion Network

With the objective of fusing instance-discriminative features from multiple modalities, we propose a Vision Transformer based modality fusion network. We obtain patch embeddings for an input image by dividing it into patches of size 16×16 and propagating their flattened versions through a linear layer. A learnable vector, that we call the instance token, of the same size is prepended to the patch embeddings; this serves as the output representation of the ViT. Abstractly, the teacher network can be defined as a binary function $\Gamma(p, s)$, that takes a photo p, and a sketch s (of the same instance), as inputs and returns two vectors $\mathbf{x}_p$ and $\mathbf{x}_s$, the sketch-to-photo and photo-to-sketch cross-attention representations, respectively.

We now proceed to a more concrete definition of the sketch-to-photo cross-attention fusion operation, and the photo-to-sketch version will be analogous and symmetric. Let the instance tokens of the photo and the sketch branches be represented by $\mathbf{p}_c$ and $\mathbf{s}_c$ respectively, and the set of patch tokens for the sketch branch by $\mathbf{s}_{patch}$. We first propagate $\mathbf{p}_c$ through a modality transformation layer $t_{\mathcal{P} \rightarrow \mathcal{S}}$ that maps it from the photo-modality to the sketch-modality, producing $\overline{\mathbf{p}}_c = t_{\mathcal{P} \rightarrow \mathcal{S}}(\mathbf{p}_c)$. We then concatenate $\overline{\mathbf{p}}_c$ to $\mathbf{s}_{\text{patch}}$ to obtain $\overline{\mathbf{s}} = [\overline{\mathbf{p}}_c, \mathbf{s}_{\text{patch}}]$. Learnable matrices $\mathbf{W}_q$ and $\mathbf{W}_k$ are used to project $\overline{\mathbf{p}}_c$ and $\overline{\mathbf{s}}$ respectively, onto the same dot-product space with D dimensions where the attention scores for $\overline{\mathbf{p}}_c$ are computed, followed by a subsequent *softmax* on their product. The cross-modal-attention embedding for the sketch-to-photo branch can then be computed as:

$$\mathbf{a} = \sigma\left((\overline{\mathbf{p}}_c \mathbf{W}_q) \cdot (\overline{\mathbf{s}} \mathbf{W}_k)^{\mathrm{T}} / \sqrt{D}\right), \ \mathbf{x}_p = \Phi_{\mathbb{X}}^{\mathcal{S} \rightarrow \mathcal{P}} \left(\overline{\mathbf{p}}_c + l_{norm}(\mathbf{a} \cdot \overline{\mathbf{s}} \mathbf{W}_k)\right)$$

where $\Phi_{\mathbb{X}}^{*}$ is a fully-connected projection head, mapping the sketch-to-photo and photo-to-sketch embeddings to the same representation space $\mathbb{X}$, which we term as the cross-modal attention space, and $l_{norm}(\cdot)$ is the layer normalization operation [1]. Since cross-modal attention follows the same operational semantics as that of self-attention, it can also be computed for multiple attention heads [6, 9]. For m heads, m cross-modal attention operations are performed in parallel followed by a concatenation and projection of their outputs, and D is set to D/m to keep compute and number of parameters constant.

Learning Objective: We design an objective termed Cross-Attention Queue Contrast (XAQC), to learn the cross-modal attention space $\mathbb{X}$. XAQC aims to bring cross-modal attention (XMA) representations $\mathbf{x}_p$ (XMA-photo) and $\mathbf{x}_s$ (XMA-sketch) for the positive sketch-photo pairs close together and push those of different instances away from each other. This alignment is achieved by making the XMA-sketch representations act as soft targets for the XMA-photo representations for the same instance. XMA-sketch representations for other instances are treated as negatives and a $(k+1)$-way softmax-based binary cross-entropy loss is minimized under this setting, where k is the number of negatives.

Consider a photo p, and two of its corresponding sketches s_1 and s_2. The cross-attention representations of the sketch-photo pairs (p, s_1) and (p, s_2) are $(\mathbf{x}_p^1, \mathbf{x}_s^1)$ and $(\mathbf{x}_p^2, \mathbf{x}_s^2)$. The loss $\texttt{XAQC}(\mathbf{x}_p^1, \mathbf{x}_s^2, \mathcal{Q})$ is then formulated as:

$$\mathcal{L}^{\text{teacher}} = \texttt{XAQC}(\mathbf{x}_p^1, \mathbf{x}_s^2, \mathcal{Q}) = -\log \frac{\exp\left(\mathbf{x}_p^1 \cdot \mathbf{x}_s^2/\tau\right)}{\sum\limits_{\forall \mathbf{h}_s \in \mathcal{X}} \exp\left((\mathbf{x}_p^1 \cdot \mathbf{h}_s)/\tau\right)} \tag{1}$$

where $\mathcal{X} = \mathcal{Q} \cup \{\mathbf{x}_s^2\}$, $\mathcal{Q}$ is a fixed-size dynamic queue of XMA-sketch representations from previous mini-batches, and τ is a hyperparameter controlling the concentration of the distribution (higher values produce softer distributions). For each new sample, we enqueue its photo-to-sketch representation $\mathbf{x}_s^2$ into $\mathcal{Q}$ after computing Equation (1)[1]. We freeze the weights of the network while computing x_s^2. While training, we randomly sample the sketch pairs, so $\texttt{XAQC}(\mathbf{x}_p^2, \mathbf{x}_s^1, \mathcal{Q})$ would also be invoked at some point during the training. Since the dot-product is a symmetric similarity metric, the corresponding symmetric property of a sketch representation being closer to its photo representation rather than to the representations of other photos is also satisfied by minimizing the $\texttt{XAQC}$ loss.

3.2 Cross-Modal Fusion Distillation

To bypass the expensive cross-modal attention computation at test-time, we propose a strategy to decouple the joint photo-sketch input space of the modality fusion *teacher* network, Γ, by training simple uni-modal CNNs or ViTs (*students*) to align their output with that of the corresponding branches of the teacher. We formulate this process as a composition of contrastive and relational cross-modal distillation.

Since the uni-modal students only ever encounter information from a single modality, the modality-fused XMA representations obtained from the teacher act as oracles, directing the optimizers of the uni-modal students to converge to a locality in the representation space that captures information from both the modalities, while being also instance-discriminative.

Contrastive Cross-Modal Distillation: We introduce an objective that treats cross-modal distillation as a contrastive learning problem, aiming to pull teacher and student representations for the same input close together, while pushing those for different inputs farther apart. More specifically, we leverage the contrastive nature of the proposed XAQC loss (Eq. (1)) and use it as a representation alignment criterion for training the students, as we detail below.

Given a corresponding photo-sketch pair (p_i, s_i), we formulate our contrastive cross-modal distillation objective so as to learn photo and sketch encoders ξ_{photo} and ξ_{sketch} that

[1] In implementation, the queue update is done at the level of a mini-batch rather than a sample.

bring $\xi_{\text{photo}}(p_i)$ and $\xi_{\text{sketch}}(s_i)$ close to $\mathbf{x}_{p_i}$ and $\mathbf{x}_{s_i}$ (teacher embeddings obtained from Γ) respectively, and push apart the representations $\mathbf{x}_{p_j}$ and $\mathbf{x}_{s_j}$, where $i \neq j$. For this purpose, we use the **XAQC** loss as:

$$\mathcal{L}_{\text{XAQC}}^{\text{sketch-student}} = \textbf{XAQC}(\xi_{\text{sketch}}(s_i), \mathbf{x}_{s_i}, \mathcal{Q}_{\mathcal{S}}) \tag{2}$$

where $\xi_{\text{sketch}} : \mathcal{S} \rightarrow \mathbb{X}$ is an encoder mapping sketches to the cross-modal-attention space, s_i is a sketch of the instance i, and $\mathbf{x}_s^i$, the XMA-sketch with its corresponding photo. $\mathcal{Q}_{\mathcal{S}}$ is a fixed size dynamic queue containing XMA-sketch representations of samples from previous mini-batches, and which is enqueued with $\mathbf{x}_s^i$ after computing Equation (2). The learning objective for the photo student, *i.e.*, $\mathcal{L}_{\text{XAQC}}^{\text{photo-student}}$, is also symmetrically formulated.

Transfer of Cross-Modal Attention Geometry: We also introduce the additional constraint of preserving certain semantically meaningful geometric properties of the teacher's cross-modal attention space, in the students' representation space. Such constraints have been found to benefit knowledge distillation in the uni-modal scenario [26]. To fulfill this goal, we model the relationships between the embeddings of arbitrary k-tuples of datapoints in terms of distance and angular relationships.

Consider photo-sketch pairs $(p_1, s_1), (p_2, s_2)$ and (p_3, s_3). Let the XMA and the student embeddings be given by $\mathbf{x}_{p_i}, \mathbf{x}_{s_i} = \Gamma(p_i, s_i)$ and $\mathbf{z}_{p_i} = \xi_{\text{photo}}(p_i)$, $\mathbf{z}_{s_i} = \xi_{\text{sketch}}(s_i)$, respectively. Let m be an abstract notation for a modality, i.e., $m \in \{p, s\}$, generically standing for both the photo and the sketch modalities. The computation of the mutual relational potentials among the teacher and the student embeddings can then be expressed as $\pi_m^{\text{teacher}} = \psi_1(\mathbf{x}_{m_1}, \mathbf{x}_{m_2}) + \psi_2(\mathbf{x}_{m_1}, \mathbf{x}_{m_2}, \mathbf{x}_{m_3})$ and $\pi_m^{\text{student}} = \psi_1(\mathbf{z}_{m_1}, \mathbf{z}_{m_2}) + \psi_2(\mathbf{z}_{m_1}, \mathbf{z}_{m_2}, \mathbf{z}_{m_3})$, respectively. ψ_1 and ψ_2 are distance and angle based relation potential functions respectively, defined as:

$$\psi_1(x,y) = \frac{1}{\mu}||x-y||_2; \; \psi_2(x,y,z) = \frac{x-y}{||x-y||_2} \cdot \frac{z-y}{||z-y||_2}$$

where μ is a normalization factor equal to the average distance among all (x,y) pairs in a mini-batch. Finally, the Cross-Modal Relational Distillation (XMRD) loss between the teacher and the student is obtained as follows:

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

where $\delta(\cdot)$ is the Huber loss, given by $\delta(a,b) = \begin{cases} \frac{1}{2}(a-b)^2 & \text{for } |a-b| \leq 1 \\ |a-b| - \frac{1}{2}, & \text{otherwise.} \end{cases}$

Learning Objective: The final objective for training the photo and the sketch students after combining the contrastive and the cross-modal relational distillation losses is as follows:

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

where λ is a hyperparameter for balancing the contrastive and the relational components.

4 Experiments

In this section, we perform an extensive experimental evaluation of our model on several FG-SBIR benchmarks and ablate our model components.

Datasets: We consider both single and multi-class standard FG-SBIR datasets for experimental evaluation, where the former contain photos and sketches of a large number of instances of a single class, and the latter contain a large number of classes, with relatively fewer instances per class. The single-class datasets comprise of the QMUL-Shoe-V2 and QMUL-Chair-V2 [52], while for multi-class evaluation, we use the Sketchy database [35].

The total number of photos and corresponding sketches present in each of the datasets along with their fraction used for testing are listed in the supplementary. We follow the dataset splitting convention as well as the performance evaluation metric of recent works [2, 3, 33, 34, 37] for the QMUL FG-SBIR datasets and [23, 35] for Sketchy. We use the acc@K evaluation metric [38] (with $K = 1$ and 10) which computes the proportion of query sketches for which the correct photo was present in the top-K returned results [3, 34, 50].

Implementation Details: We use ViT base (ViT-B) models pre-trained on ImageNet [30] from [45] as homogeneous encoders of the XMA teacher. We use 12 layers of cross-modal attention, which produce an output embedding of 768 dimensions. For the students, we primarily report and suggest the usage of ViT small (ViT-S) networks as backbones, since they provide the most optimal size-to-accuracy ratio. However, depending on requirements, one may choose to use other ViT architectures or even CNNs for the purpose of knowledge distillation. The size of the XAQC queues were computed as $2^{\lfloor \log N \rfloor}$, where N is the number of datapoints. The teacher network was optimized with a learning rate of 3×10^{-6} under a cosine-annealed schedule and a weight decay of 10^{-4} using the Adam optimizer [17]. The student networks were optimized using an initial learning rate of 10^{-5} with an exponential decay and a weight decay of 10^{-5}. The details of our hardware and software platforms are provided in the supplementary.

4.1 Ablation Studies

Model Components and Losses: Table 1 reports our ablation studies, where XMA stands for cross-modal attention, and its presence means that the XMA operator is applied to the outputs of the homogeneous encoders. Otherwise, the outputs of the homogeneous encoders serve as the final representation. XAQC-R and XAQC-D refer to the usage of the XAQC loss for ranking the teacher's representation space and as a contrastive knowledge distillation criterion for the students respectively. Upon ablation, the triplet loss serves as the alternative in both cases. XMRD indicates the presence or absence of the cross-modal relational distillation criterion while training the students.

ID	Teacher		Student		Shoe-V2	Chair-V2
	XMA	XAQC-R	XAQC-D	XMRD	acc@1	acc@1
1	–	–	–	–	30.83	49.66
2	✓				34.31	52.35
	✓	✓			41.33	57.65
	✓		✓		35.62	53.40
	✓			✓	35.45	52.40
3	✓	✓	✓		43.21	62.70
	✓	✓		✓	42.73	60.55
	✓		✓	✓	35.70	54.98
		✓	–	–	36.50	55.22
4	✓	✓	✓	✓	**45.05**	**63.48**

Table 1: Results of ablating the core components of our model; grouped so as to provide a view of how each of the components contribute individually and in conjunction with each other. '–' indicates that the component is irrelevant for that particular setting.

We group our results into 4 categories (indicated by ID). ID-1 establishes a baseline performance by following a classic deep learning based SBIR framework [35] with ViTs as encoders for the individual modalities, with the triplet loss as the only optimization objective. The improvements over this model could be used to demonstrate the algorithmic contributions of our work. ID-2 illustrates the contribution of the cross-modal attention operation,

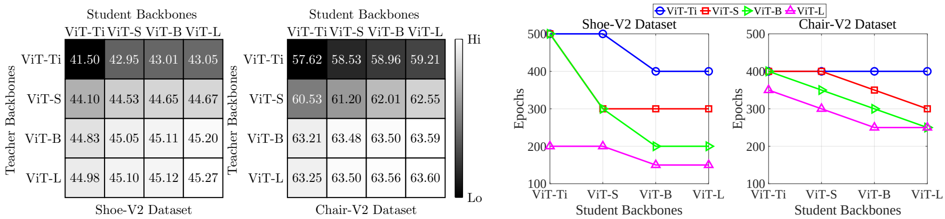

Figure 3: (Left) Variation in acc@1 and (right) student convergence time (in number of epochs) with teacher/student encoder size.

and how each of the other components of the network enhances the overall performance by appropriately utilizing the representation learned by it. ID-3 provides an estimate of how individually suppressing each of the components of the framework affects the overall performance. ID-4 reports the performance of the complete model. The first row in ID-2 indicates that the XMA in isolation provides over 2.6% gain in acc@1. The results also indicate that the contrastive loss used for ranking the teacher's representation space is of pivotal importance. Using the XAQC loss to train the teacher instead of triplet can improve the performance by up to 7%. Thus, it is only when the teacher learns robust enough representations that the students are able to distill out the relevant knowledge. The results in ID-3 also echo the findings of ID-2; suppressing the cross-modal attention and the XAQC loss in the teacher have the most degrading effects. The last row in ID-3 is equivalent to the setup in ID-1 trained with the XAQC-R loss. With no modality fusion, the encoders are already independent, and hence, do not require the distillation step.

Encoder Backbones: We considered the Tiny (Ti), Small (S), Base (B) and Large (L) versions of the ViT and used them as backbones in both the teacher and the students. Figure 3 summarizes the effects of pairing backbones of different sizes. We observe that, with the Tiny teacher, the performance significantly improves as we increase the size of the student. However, this trend stabilizes as we increase the size of the teacher. However, with the student fixed, as we increase the size of the teacher, the convergence times reduce. This drop in convergence time continues as we increase the size of the student. Based on the above observations, we postulate that, larger teachers, by the virtue of their higher entropic capacity and output dimensionality, are able to learn more expressive representations, thereby presenting a simpler target hypothesis for the downstream students to approximate. Smaller teachers are forced to obtain a more compressed representation, thereby limiting their amount of expressivity. This results in a more complex target hypothesis and thus, can be better learned by larger students (searching through a larger hypothesis space). We also experimented with CNNs as student backbones and achieved similar results (details in the supplementary).

4.2 Comparison with State-of-the-Art

QMUL FG-SBIR: The quantitative comparison of our method with the QMUL FG-SBIR datasets (Shoe-V2 and Chair-V2) is given in Table 2. *Triplet-SN* [52] uses triplet loss to train a Sketch-a-Net [53] baseline. *Triplet-Attn* [36] is a spatial attention based extension of [52]. *Triplet-RL* performs on-the-fly FG-SBIR by a reinforcement learning based fine-tuning. *CC-Gen* [24] models a universal manifold of prototypical cross-category sketch traits. *TVAE* [16] employs a VAE with single modality translation, while *DVML* [19] disentangles sketch fea-

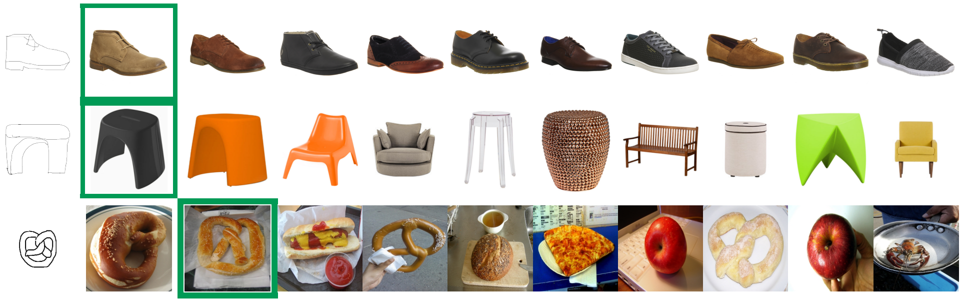

Figure 4: Qualitative fine-grained SBIR results of our method on the Shoe-V2 (row-1), Chair-V2 (row-2) and Sketchy (row-3) datasets. More qualitative results can be found in the supplementary.

tures into variant and invariant components. Strong performance on Shoe-V2 was achieved by *ReinfGen* [3] via a sketch generative framework based on reinforcement learning using additional unpaired training photos. Promising results were reported by StyleVAE [34] and *SketchAA* [50] via sketch disentanglement into independent style and content components, and sketch abstraction through a graph convolutional network respectively. *Partial-SBIR* [7] and *NT-SBIR* [5] respectively proposed dealing with partial information and noise in sketches via optimal transport and reinforcement learning based approaches. *NT-SBIR* is currently the SOTA on acc@1 for Chair-V2. With our novel XModalViT framework, we were able to beat the acc@1 and acc@10 SOTA on Shoe-V2 and acc@10 SOTA on Chair-V2 by 1.35%, 2.73%, and 0.14% respectively.

Method	Shoe-V2		Chair-V2	
	acc@1	acc@10	acc@1	acc@10
Triplet-SN [52]	28.71	71.56	47.65	84.24
Triplet-Attn [36]	31.74	75.78	53.41	87.56
Triplet-RL [2]	34.10	78.82	56.54	89.61
CC-Gen [24]	33.80	77.86	54.21	88.23
TVAE [16]	27.62	70.32	49.37	81.63
DVML [19]	32.07	76.23	52.78	85.24
SketchAA [50]	32.33	79.63	52.89	94.88
StyleVAE [34]	36.47	81.83	62.86	91.14
ReinfGen [3]	39.10	87.50	62.20	90.80
Partial-SBIR [7]	39.90	82.90	-	-
NT-SBIR [5]	43.70	-	**64.80**	-
Ours (XModalViT)	**45.05**	**90.23**	63.48	**95.02**

Table 2: Quantitative comparison (in %) with state-of-the-art for fine-grained SBIR on the QMUL datasets.

Sketchy: The quantitative comparison of our method on the Sketchy dataset is given in Table 3. *GN-Siamese* [35] and *GN-Triplet* [35] train a GoogleNet [39] with the siamese and triplet losses respectively in a contrastive manner, while *SAN-Triplet* [52] applies the triplet loss on Sketch-a-Net [53]. *XDGen* [23] introduces the task of cross-domain image synthesis, achieving the current SOTA on acc@1. By additionally leveraging textual descriptions and cascaded coarse-to-fine instance-level features, *DCCRM* (S+I+D) [42] is the current SOTA on acc@10. With our novel XModalViT framework, we were able to beat both the acc@1 and acc@10 SOTA on Sketchy by 6.01% and 0.37% respectively, without requiring hard to obtain textual annotations as in DCCRM (S+I+D). We were also able to surpass the average human-level acc@1 performance on Sketchy, as reported in [35].

Method	Sketchy	
	acc@1	acc@10
Human [35]	54.27	–
SAN-Triplet [52]	25.87	–
GN-Siamese [35]	27.36	–
GN-Triplet [35]	37.10	–
XDGen [23]	50.14	–
DCCRM (S+I) [42]	40.16	92.00
DCCRM (S+I+D) [42]	46.20	96.49
Ours (XModalViT)	**56.15**	**96.86**

Table 3: Quantitative comparison (in %) with state-of-the-art for fine-grained SBIR on Sketchy dataset.

4.3 Qualitative Results

Figure 4 depicts examples of retrieval by our model, with additional results in the supplementary. If the target photo has a significant number of discriminative local features compared to others in the gallery, it can be seen to always appear in the top-1. However, correct instances that get demoted in rank are preceded by ones that bear significant resemblance to the fine-grained local features of the query. For instance, the photo ranked first in the last row can be seen to bear significant resemblance to the query (considering the fact that it belongs to the same class as the ground-truth, *i.e.*, ‘pretzel’, and the curvature pattern of the object). These results provide evidence that our framework learns to draw fine-grained associations of the underlying concept across the two modalities.

In Figure 5, we depict examples where modality-specific features uniquely identify an instance. Sketchers use techniques like scribbles to illustrate color/texture differences (Shoe) and grid/mesh-like structures or primitive shapes like squares/triangles as approximations for complex patterns (Chair and a Turtle’s shell). As an example (row 1), while a large variety of shoes would have the same kind of front, which modality-shared models focus on (right), our model (left), by the virtue of modality fusion (XMA), is capable of attending to more discriminative features like contrasting color/texture depicted by scribbles, or partition denoted by a line, while also capturing shared features like local geometry.

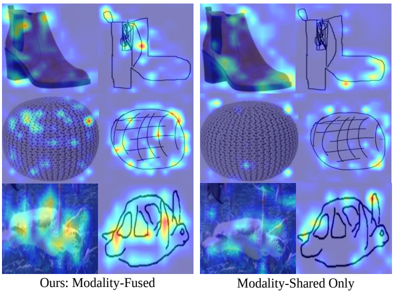

Figure 5: Comparison of attention maps obtained from our modality-fused representations and the conventional modality-shared-only approaches.

5 Conclusion

We approached the problem of cross-modal representation learning for the task of fine-grained sketch-based image retrieval (FG-SBIR) by taking a detour from the conventional objective for the task. We posited that sketches and photos are instances of an underlying singular abstract concept, and both modalities contain complementary information for constructing a representation of that abstraction. This motivated us to frame our representation learning objective so as to fuse information from local correspondences across both the modalities via the cross-modal attention operation. We then formulated a technique to decouple the modality-fused representations into independent modality-specific encoders via a contrastive learning objective that would direct the modality specific encoders to converge towards the unified, fused representations, while preserving the geometry of the cross-modal attention space. We empirically validated the capability of our method to learn expressive representations by achieving state-of-the-art results on FG-SBIR benchmark datasets.

Acknowledgements

This work has been partially supported by the ERC 853489–DEXIM, by the DFG–EXC number 2064/1–Project number 390727645, and as part of the Excellence Strategy of the German Federal and State Governments.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv*, 2016.

[2] 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.

[3] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval. In *CVPR*, 2021.

[4] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Yongxin Yang, Timothy Hospedales, Tao Xiang, and Yi-Zhe Song. Vectorization and rasterization: Self-supervised learning for sketch and handwriting. In *CVPR*, June 2021.

[5] Ayan Kumar Bhunia, Subhadeep Koley, Abdullah Faiz Ur Rahman Khilji, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketching without worrying: Noise-tolerant sketch-based image retrieval. In *CVPR*, June 2022.

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

[7] Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Viswanatha Reddy Gajjala, Aneeshan Sain, Tao Xiang, and Yi-Zhe Song. Partially does it: Towards scene-level fg-sbir with partial input. In *CVPR*, 2022.

[8] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle to Search: Practical Zero-Shot Sketch-Based Image Retrieval. In *CVPR*, 2019.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[10] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *CVPR*, 2019.

[11] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for any-shot sketch-based image retrieval. *IJCV*, 2020.

[12] Xinqian Gu, Bingpeng Ma, Hong Chang, S. Shan, and Xilin Chen. Temporal knowledge propagation for image-to-video person re-identification. *ICCV*, 2019.

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *CVPR*, 2020.

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPSW*, 2015.

[15] Rui Hu and John Collomosse. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *CVIU*, 2013.

[16] Haque Ishfaq, Assaf Hoogi, and Daniel Rubin. Tvae: Triplet-based variational autoencoder using metric learning. In *ICLRW*, 2018.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[18] Yi Li, Timothy Hospedales, Yi-Zhe Song, and Shaogang Gong. Fine-grained sketch-based image retrieval by matching deformable part models. In *BMVC*, 2014.

[19] Xudong Lin, Yueqi Duan, Qiyuan Dong, Jiwen Lu, and Jie Zhou. Deep variational metric learning. In *ECCV*, 2018.

[20] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *CVPR*, 2021.

[21] Satyam Mohla, Shivam Pande, Biplab Banerjee, and Subhasis Chaudhuri. Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification. In *CVPRW*, 2020.

[22] Krishna D N and Ankita Patil. Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks. In *INTERSPEECH*, 2020.

[23] Kaiyue Pang, Yi zhe Song, Tony Xiang, and Timothy Hospedales. Cross-domain generative learning for fine-grained sketch-based image retrieval. In *BMVC*, 2017.

[24] Kaiyue Pang, Ke Li, Yongxin Yang, Honggang Zhang, Timothy M. Hospedales, Tao Xiang, and Yi-Zhe Song. Generalising fine-grained sketch-based image retrieval. In *CVPR*, 2019.

[25] Kaiyue Pang, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval. In *CVPR*, 2020.

[26] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, 2019.

[27] Angelo Porrello, Luca Bergamini, and Simone Calderara. Robust re-identification by multiple views knowledge distillation. In *ECCV*, 2020.

[28] F. Radenovic, G. Tolias, and O. Chum. Deep shape matching. In *ECCV*, 2018.

[29] Karsten Roth, Timo Milbich, Bjorn Ommer, Joseph Paul Cohen, and Marzyeh Ghassemi. Simultaneous similarity-based self-distillation for deep metric learning. In *ICML*, 2021.

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[31] Jose M Saavedra. Sketch based image retrieval using a soft computation of the histogram of edge local orientations (s-helo). In *ICIP*, 2014.

[32] Jose M Saavedra, Juan Manuel Barrios, and S Orand. Sketch based Image Retrieval using Learned KeyShapes (LKS). In *BMVC*, 2015.

[33] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Cross-modal hierarchical modelling for fine-grained sketch based image retrieval. In *BMVC*, 2020.

[34] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Stylemeup: Towards style-agnostic sketch-based image retrieval. In *CVPR*, 2021.

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

[36] 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.

[37] Jifei Song, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. Learning to sketch with shortcut cycle consistency. In *CVPR*, 2018.

[38] Wanhua Su, Yan Yuan, and Mu Zhu. A relationship between the average precision and the area under the roc curve. In *ICTIR*, 2015.

[39] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. doi: 10.1109/CVPR.2015.7298594.

[40] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2020.

[41] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, 2018.

[42] Yanfei Wang, Fei Huang, Yuejie Zhang, Rui Feng, Tao Zhang, and Weiguo Fan. Deep cascaded cross-modal correlation learning for fine-grained sketch-based image retrieval. *PR*, 2020.

[43] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *CVPR*, 2020.

[44] Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *PAMI*, 2021.

[45] Ross Wightman. PyTorch Image Models. https://github.com/rwightman/pytorch-image-models, 2019.

[46] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

[47] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *CVPR*, 2020.

[48] Jiaqing Xu, Haifeng Sun, Qi Qi, Jingyu Wang, Ce Ge, Lejian Zhang, and Jianxin Liao. Dla-net for fg-sbir: Dynamic local aligned network for fine-grained sketch-based image retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[49] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *CVPR*, 2022.

[50] Lan Yang, Kaiyue Pang, Honggang Zhang, and Yi-Zhe Song. Sketchaa: Abstract representation for abstract sketches. In *ICCV*, 2021.

[51] Lu Yu, Vacit Oguz Yazici, Xialei Liu, Joost van de Weijer, Yongmei Cheng, and Arnau Ramisa. Learning metrics from teachers: Compact networks for image embedding. In *CVPR*, 2019.

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

[53] Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. Sketch-a-net: A deep neural network that beats humans. *IJCV*, 2017.