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

Aneeshan Sain[1,2]
a.sain@surrey.ac.uk
Ayan Kumar Bhunia[1]
a.bhunia@surrey.ac.uk
Yongxin Yang[1,2]
yongxin.yang@surrey.ac.uk
Tao Xiang[1,2]
t.xiang@surrey.ac.uk
Yi-Zhe Song[1,2]
y.song@surrey.ac.uk

[1] SketchX, CVSSP
University of Surrey, UK

[2] iFlyTek-Surrey Joint Research Centre
on Artificial Intelligence

Abstract

Sketch as an image search query is an ideal alternative to text in capturing the fine-grained visual details. Prior successes on fine-grained sketch-based image retrieval (FG-SBIR) have demonstrated the importance of tackling the unique traits of sketches as opposed to photos, e.g., temporal vs. static, strokes vs. pixels, and abstract vs. pixel-perfect. In this paper, we study a further trait of sketches that has been overlooked to date, that is, they are hierarchical in terms of the levels of detail – a person typically sketches up to various extents of detail to depict an object. This hierarchical structure is often visually distinct. In this paper, we design a novel network that is capable of cultivating sketch-specific hierarchies and exploiting them to match sketch with photo at corresponding hierarchical levels. In particular, features from a sketch and a photo are enriched using cross-modal co-attention, coupled with hierarchical node fusion at every level to form a better embedding space to conduct retrieval. Experiments on common benchmarks show our method to outperform state-of-the-arts by a significant margin.

1 Introduction

Great strides have been made towards cross-modal image retrieval [16, 34], predominantly using text as a query [25, 53]. Sketches as an alternative query mainly focussed on category-level retrieval at first [2, 7, 8, 52]. It was not until recently when the fine-grained characteristic of sketches has been noted, which then triggered the study on fine-grained image retrieval [36, 46, 58]. At its inception, *fine-grained* sketch-based image retrieval (FG-SBIR) focussed on retrieving photos of commercial objects [1, 36, 58]. This had very recently been extended to the problem of graphical user interface (GUI) retrieval [19], where GUI sketches are used to conduct fine-grained retrieval of mobile app screenshots.

It has been shown that by specifically targeting the unique traits of sketches, the sketch-photo domain gap can be effectively narrowed. Exemplar works include those that address

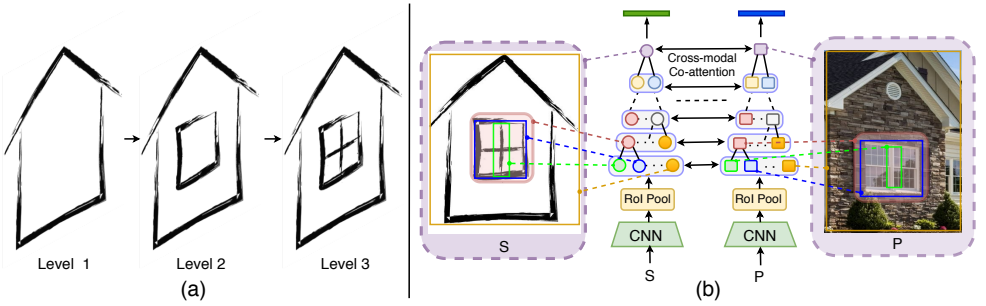

Figure 1: (a) shows different hierarchical levels of detail in a free-hand sketch (illustrative purpose only), while (b) illustrates our overall idea of cross-modal hierarchical modelling.

the sequential [17], abstract [32], and stroke-wise [1] characteristics of sketches, either separately or in combination [40, 56]. However, all previous works predominately treat sketches as a single flat structure, without recognising the inherent hierarchical structure within. Hierarchies in a sketch importantly underpin its flexibility as a search query – when an user sketches, the *extent of details being sketched* can vary from coarse to fine. As Figure 1(a) shows, a house may be drawn as a simple outline (Level 1), with further strokes denoting the window (Level 2), or with even finer strokes to depict window pattern (Level 3). Understanding sketching hierarchies is crucial in handling the uncertainty brought by the varying levels of sketch details. Importantly, it enables the learning of discriminative features for sketch-to-photo matching at different levels of detail.

Devising a hierarchy amongst sketching elements is non-trivial as they do not follow any predefined composition rules. Rather, sketches exhibit variable details and hierarchical compositions, due to differences in subjective perception and sketching skill. Furthermore, in the context of sketch-photo retrieval, it is crucial for both modalities to withhold cross-modal semantic relevance at corresponding hierarchical levels. In this paper, we propose a novel FG-SBIR framework that discovers the hierarchical structure within each modality, where the discovery is aided by exchanging cross-modal information at each level. This cross-modal hierarchy building has two vital benefits: (i) it helps with the discovery of underlying semantic hierarchy, especially for sketches which exhibit highly sparse visual cues compared to photos, and (ii) it encourages the learning of a better cross-modal embedding by exchanging information across two domains at each level and propagating across different levels to form the final embedding vectors.

To establish a hierarchy in each modality, we mimic an agglomerative merging scheme where the model chooses two nodes to be merged at every level to form the next. Albeit desirable, merging more than two nodes simultaneously would make the problem computationally intractable [5]. The key question therefore comes down to *choosing* which two nodes to merge. Such a choice involves a *discrete* decision, thus invoking non-differentiability into the network. As our first contribution, we model this choice via a straight-through Gumbel-Softmax [22] operation that approximates one-hot vectors sampled from a distribution by making them continuous (§3.1). This helps with calculating gradients of discrete decisions thus allowing backpropagation, and making the merging scheme end-to-end trainable. Upon discovery, two nodes are merged over a learnable layer forming a higher order semantic.

In order to encourage the learning of a better cross-modal embedding, we further constrain node merging via a cross-modal interaction module, which computes feature similarity across the two modalities at every hierarchy level. This is fundamentally different to prior works [46, 58] that independently embed sketches and photos into a common space with-

out any cross-modal interaction. More specifically, we introduce a cross-modal co-attention module that attends to salient regions in a sketch and its photo counterpart, thus ensuring mutual awareness of both modalities. Information aggregated from the photo branch (based on attentive cues from the sketch branch), is integrated with the sketch branch via a gating mechanism; and vice-versa. The gating mechanism adaptively controls the fusion intensity thus filtering out negative effects of mismatched parts.

In summary, our contributions are: (i) an end-to-end trainable architecture that enables the discovery of the underlying hierarchy of a sketch, (ii) a co-attention module to facilitate cross-modal hierarchy construction, and (iii) an unique perspective of utilising hierarchies for the problem of FG-SBIR. Extensive ablative studies and evaluations against state-of-the-arts on three standard FG-SBIR and GUI retrieval datasets, show our method to outperform most existing methods by a significant margin.

2 Related Work

Category-level SBIR: Most category-level SBIR approaches fall into two categories: (i) handcrafted descriptors [39, 49] which constructs global [38] or local [18] photo-sketch joint representations. (ii) deep learning methods [28, 46, 58], where classical ranking losses, like contrastive or triplet loss have been used. Other related problems such as zero-shot SBIR [11, 13, 29] and sketch-photo hashing [28, 42] are also studied.

Fine-grained SBIR: Unlike category-level SBIR, fine-grained SBIR aims at instance-level matching. This meticulous task relies on learning the unique traits of a sketch that lie in its fine-grained details. Being fairly new, FG-SBIR is less studied in comparison to category-level SBIR tasks. Starting from the study of deformable-part models [26], recent methods have learned to extract comparable features from heterogeneous domains [35, 46, 58]. Yu *et al*. [58] proposed a deep triplet-ranking model for instance-level FG-SBIR, which was enhanced via hybrid generative-discriminative cross-domain image generation [35] and attention based techniques, in addition to advanced triplet loss formulations [46]. Analogous to 'zero-shot' SBIR tasks, Pang *et al*. [36] worked on cross-category FG-SBIR. Recently, Huang *et al*. [19] applied a triplet network to retrieve GUI photos from sketch [19] and contributed the first GUI sketch dataset SWIRE [19]. These methods however, do not involve any cross-modal correlation in the context of FG-SBIR. Furthermore the notion of hierarchy has not yet been applied in this context.

Cross-modal Interaction: Existing FG-SBIR works [46, 58] *independently* embed sketches and photos into a common space, disregarding cross-modal interaction. This leads to sub-optimal features as the query representation is unaware of its paired photo during training. Others in the non-sketch literature had otherwise successfully investigated the benefits of cross-modal interaction. Varior *et al*. [51] employed a gating mechanism for person re-identification, comparing mid-level network features across both modalities. While they focus on global feature maps, Wang *et al*. [53] involved local image sub-regions, capturing cross-modal correlation better. Recently, ViLBERT [31] leveraged the BERT framework to interact separately processed visual and textual inputs via co-attentional transformer layers. In this paper, we introduce cross-modal feature matching between sketches and photos, and show how it can be embedded during hierarchy construction.

Hierarchical Modelling: Several works have emerged modelling latent textual hierarchy using Recursive Neural Networks (RvNN) [44]. Tree-RNNs [43], enhanced by RNTN networks [45] surpassed vanilla LSTMs in synthesising sentence meaning from word-vectors. Stacking tree-structured vanilla RvNNs [21] improved sentiment classification. While gated RvNNs [4] controlled children-to-parent information flow, Neural tree indexer (NTI) [33]

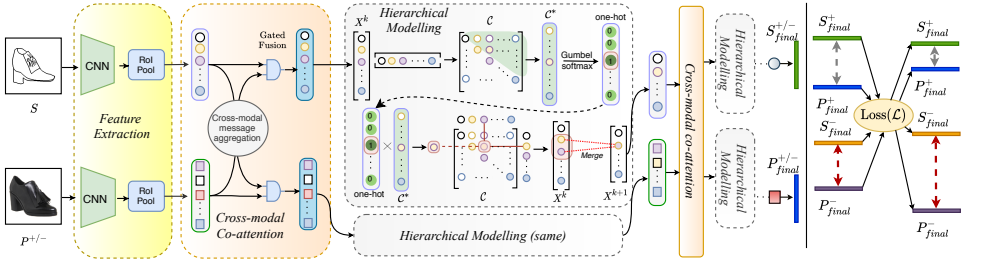

Figure 2: Our framework. After extracting region-wise features from a sketch (S) and a photo ($P^{+/-}$), they are enriched via cross-modal attention ($\mathcal{M}$), followed by hierarchical parsing operation (h_ϕ). The last two steps are repeated consecutively, until a final representation from each branch ($S^{+/-}_{final}$,$P^{+/-}_{final}$) is obtained. A loss ($\mathcal{L}$) brings the matching sketch-photo pair (S^{+}_{final},P^{+}_{final}) closer (grey) while distancing (red) the unmatched one (S^{-}_{final},P^{-}_{final}).

used soft hierarchical structures to employ Tree-LSTMs. Methods using discrete levels of contextual information [50], recursive context propagation networks [41], shape-parsing strategies [30] or scene-graphs [23, 55] have aimed at encoding regions with semantic relations. Graph neural networks (GNN) [3, 37] being well-suited to structured data offer viability towards implementing hierarchy. DIFFPOOL [57] learns a differentiable soft cluster assignment to map nodes into clusters, which are pooled as input for the next GNN layer. Some works aim at capturing contextual structure [23, 55] by building a graph between scene objects and texts. Recent works like SAGPool [24] and gPool [12] focussed on top-K node selection to formulate a subgraph for the next GNN layer. While EdgePool [17] reduces almost half its nodes per GNN layer by contracting its edges, EigenPool [10] controls pooling ratio based on graph Fourier transform. Nevertheless, all such approaches usually leverage a pre-defined structure. Sketches however, are devoid of such strict composition rules, but harbour an implicit hierarchy which our model learns to discover in this work.

3 Methodology

Overview: We aim to embed sketches and photos into a common space such that, (a) the underlying hierarchy of a sketch is modelled, and (b) feature learning in either modality is reinforced via cross-modal interaction at every hierarchical level. Formally, we learn an embedding function $F : I \rightarrow \mathbb{R}^d$ that maps a photo or rasterised sketch to a d dimensional feature, which we use to retrieve against a distance metric. We employ three modules:

Backbone Feature Extractor: Let f_θ be a backbone feature extractor initialised from pre-trained InceptionV3 [1] weights. Using input I, we obtain a feature map $\mathcal{F}_I = f_\theta(I) \in \mathbb{R}^{H\times W\times C}$ where I denotes a sketch S or RGB photo P, and H, W, C signifies height, width and number of channels respectively. To extract region-wise features, we treat every sketch-stroke as an individual region, and calculate bounding boxes enclosing every stroke. Meanwhile, we use the unsupervised selective search algorithm [50] to discover sub-regions from P. Let S and P have N_S strokes and N_P discovered sub-regions with bounding box sets of $B_S \in \{b^s_1, b^s_2, ..., b^s_{N_S}\}$ and $B_P \in \{b^p_1, b^p_2, ..., b^p_{N_P}\}$ respectively. We perform ROI pooling on feature maps $\mathcal{F}_S$ and $\mathcal{F}_P$ using B_S and B_P, to obtain sets of region-wise latent feature vectors $S_r = \{s_1, s_2, ..., s_{N_S}\}$ and $P_r = \{p_1, p_2, ..., p_{N_P}\}$; where $S_r \in \mathbb{R}^{N_S\times C}$, $P_r \in \mathbb{R}^{N_P\times C}$ (as matrices). A linear projection layer then reduces each vector of P_r and S_r to a d-dimensional space.

Hierarchical Parsing Module: This module (denoted as h_ϕ) takes S_r or P_r as input and con-

structs a hierarchy of multiple levels. At any hierarchical level k, h_ϕ inputs $X^k = \{x_j\}_{j=1}^{N^k}, x_j \in \mathbb{R}^d$ consisting of a set of N^k, d-dimensional vectors (nodes). It starts with S_r or P_r and progressively identifies and merges two nodes at a time to form a higher order semantic, representing the next hierarchical level: $X^{k+1} = \{x_j\}_{j=1}^{N^{k+1}}$, $x_j \in \mathbb{R}^d$, $N^{k+1} = N^k - 1$. This node-merging process continues until a final single d-dimensional vector ($\mathbb{R}^d$) remains per branch (see Sec. 3.1 for details and Figure 2 for an illustration).

Cross-modal Co-attention: While h_ϕ is able to discover the underlying hierarchy of a sketch, we additionally stress on cross-modal co-attention between sketch and photo branches, at every hierarchical level. More specifically, we design a *cross-modal co-attention module* $\mathcal{M}$ which takes both P_r^k and S_r^k at the k^{th} hierarchical level, aggregates sketch branch information onto the photo branch and vice-versa, and finally outputs the modified vectors $\tilde{P}_r^k$ and $\tilde{S}_r^k$. Overall, $\mathcal{M}$ enforces cross-modal interaction on the sets of vectors from photo branch P_r^k and sketch branch S_r^k having region-wise features at the k^{th} hierarchical level (see Sec 3.2). Following such cross-modal information exchange, h_ϕ is applied to merge two nodes in each branch at each level. These two steps occur consecutively at every hierarchical level, until one node remains in each branch.

3.1 Hierarchical Modelling

Given a set of region-wise features $X^k = \{x_i\}_{i=1}^{N^k}$, $x_i \in \mathbb{R}^d$ at the k^{th} hierarchical level, h_ϕ aims to measure the affinity between all possible unique pairs of nodes and merge the two having the highest validity score to form a higher order semantic. We approach this by projecting the node features to a low d_h-dimensional space and formulating an intra-regional compatibility matrix $\mathcal{C} = (X^k \cdot \mathbf{W}_\phi^\mathcal{C}) \cdot (X^k \cdot \mathbf{W}_\phi^\mathcal{C})^\mathrm{T}$, $\mathcal{C} \in \mathbb{R}^{N^k \times N^k}$ where $\mathbf{W}_\phi^\mathcal{C} \in \mathbb{R}^{d \times d_h}$ is a projection matrix. $\mathcal{C}_{i,j}$ represents the validity score of each pair of nodes. Since $\mathcal{C}$ is symmetric, we consider only its upper triangular elements, excluding the principle diagonal: $\mathcal{C}^* = \texttt{flatten}(\texttt{UpTri}(\mathcal{C}))$, $\mathcal{C}^* \in \mathbb{R}^{H^k}$ where $H^k = \frac{N^k(N^k-1)}{2}$, $\texttt{UpTri}()$ extracts upper-triangular elements of a matrix and $\texttt{flatten}()$ compiles them into a vector. The pair having the highest compatibility score shall intuitively suggest the highest probability of merging.

Using *argmax* here is non-differentiable and might naively need Monte Carlo estimates with a REINFORCE-type algorithm [54], which would typically suffer from high variance [22]. We thus apply a low-variance gradient estimate via Gumbel-softmax re-parameterisation trick and Straight-Through (ST) gradient estimator [22] to $\mathcal{C}^*$. Gumble-softmax approximates one-hot vectors sampled from a distribution by introducing a *Gumbel noise* [22] in every sampling step. This replaces the discontinuous *arg-max* function by a differentiable softmax function. Given a H^k-dimensional categorical distribution across every node-pair, with $\mathcal{C}^* = (c_1, c_2, ...c_{H^k})$, where $c_i = \log(\pi_i)$ and π_i is an unnormalised log probability, we draw a sample $\mathbf{q} = (q_1, q_2, ...q_{H^k})$ from Gumbel-softmax distribution as:

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

where, g_i represents *Gumbel-noise* and τ is the temperature parameter [6]. As τ tends to 0, $\mathbf{q}$ resembles a one-hot sample. In the forward pass it discretises a continuous probability vector $\mathbf{q}$ sampled from Gumble-Softmax distribution into a one-hot vector $\mathbf{q}^{ST} = (q_1^{ST}, q_2^{ST}, ...q_{H^k}^{ST})$ where $\mathbf{q}_i^{ST} = \mathbb{1}_{[i = \mathrm{argmax}_j(q_j)]}$. During backward pass, it uses the continuous $\mathbf{q}$, thus allowing backpropagation. If $\mathcal{C}_{a,b}$ is the element for which q_i^{ST} equals 1, we fuse nodes x_a and x_b as: $\hat{x}_{a,b} = \texttt{ReLU}(\mathbf{W}_\phi^F \cdot [x_a, x_b])$, $\mathbf{W}_\phi^F \in \mathbb{R}^{d \times 2d}$ leading to the next hierarchical level output:

$$X^{k+1} := X^k - \{x_a, x_b\} + \{\hat{x}_{a,b}\} \;; \qquad x_a, x_b \in X^k. \tag{2}$$

Note that our formulation considers fusion of distant nodes in addition to adjacent ones [5]. Most importantly, the merging scheme does not require low-level nodes to be merged with low-level ones only – this fusion is guided entirely by the loss (Equation 5) alone, with no explicit control. This makes the merging scheme completely data-driven and is thus learned jointly with the rest of the FG-SBIR model. Hierarchical modelling is executed in parallel in both sketch and photo branches with shared weights, so that fusion in one branch is aware of that in the other, thus ensuring a better feature matching. Formally, $P_r^{k+1} \leftarrow h_\phi(P_r^k)$, $S_r^{k+1} \leftarrow h_\phi(S_r^k)$ and $|P_r^{k+1}| = |P_r^k| - 1$, $|S_r^{k+1}| = |S_r^k| - 1$, where $|\cdot|$ denotes cardinality of a set. In case of unequal nodes in two branches, if one reduces to $\mathbb{R}^d$ first, the other continues with hierarchical fusion until that reduces to $\mathbb{R}^d$ as well, with no new fusion on the former branch. Thus we have $|S_r^{k_{final}}|$, $|P_r^{k_{final}}| = 1$, k_{final} being the final hierarchical level.

3.2 Cross-modal Co-attention

Given a set of region-wise latent vectors $S_r^k \in \mathbb{R}^{N_s^k \times d}$ and $P_r^k \in \mathbb{R}^{N_p^k \times d}$ from photo and sketch branches at the k^{th} hierarchical level respectively, we aim to enrich S_r^k and P_r^k by passing fine-grained information between them. Formally: $[\tilde{S}_r^k, \tilde{P}_r^k] \leftarrow \mathcal{M}(S_r^k, P_r^k)$. Here '$k$' is dropped for notational brevity. Photo branch features (P_r) are aggregated w.r.t every vector from sketch feature set S_r. Finally the aggregated features (P_r^S) are combined with the sketch branch updating it as, $\tilde{S}_r = \texttt{fused}(S_r, P_r^S)$. Similarly we obtain $\tilde{P}_r = \texttt{fused}(P_r, S_r^P)$.

Towards this goal we calculate a stroke-region affinity matrix $\mathbf{A} = (S_r \cdot \mathbf{W}_\psi^S) \cdot (P_r \cdot \mathbf{W}_\psi^P)^{\mathrm{T}}$, $\mathbf{A} \in \mathbb{R}^{N_S \times N_P}$, where every element $\mathbf{A}_{i,j}$ represents the affinity between i^{th} sketch-stroke and j^{th} image-region respectively. $\{\mathbf{W}_\psi^P, \mathbf{W}_\psi^S\} \in \mathbb{R}^{d \times d_h}$ are corresponding projection matrices. Attending to every stroke feature of S_r with respect to every feature from P_r, $\mathbf{A}$ is normalised across sketch-dimension, providing a sketch-specific stroke-region affinity matrix $\mathbf{A}_S^* = \texttt{softmax}(\mathbf{A}^{\mathrm{T}}/\sqrt{d_h})$. Now we accumulate all sketch branch features S_r w.r.t. each photo-region on $\mathbf{A}_S^*$ to finally obtain $S_r^P = \mathbf{A}_S^* \cdot S_r$, $S_r^P \in \mathbb{R}^{N_P \times d}$, via dot-product based feature aggregation. On similar notion, with an attention on P_r w.r.t features from S_r we obtain, $\mathbf{A}_P^* = \texttt{softmax}(\mathbf{A}/\sqrt{d_h})$ and consequently $P_r^S = \mathbf{A}_P^* \cdot P_r$, $P_r^S \in \mathbb{R}^{N_S \times d}$. P_r^S is the aggregated feature from photo branch to be fused with the sketch branch, and vice-versa for S_r^P.

A sketch query is not only compared with its positive photo pair, but also with other negative photos which are to be filtered out during retrieval. We thus need to suppress passing of unrelated information, as well as adjust the extent of fusing aggregated features from other modality with the original one. Consequently, we design a learnable gate, $\mathcal{G}^S = \texttt{sigmoid}([S_r, P_r^S] \cdot \mathbf{W}_\mathcal{G}^S)$, $\mathbf{W}_\mathcal{G}^S \in \mathbb{R}^{2d \times d}$ where every element is normalised between 0 (no fusion) and 1 (complete fusion), adaptively controlling the extent of fusion. Multiplying $\mathcal{G}^S \in \mathbb{R}^{N_S \times d}$, with a combination of S_r and P_r^S, signifies that the greater the *correlation* of a region with a stroke, the further shall be the impact of fusion. Finally, to preserve the original data of region features which should not have been fused intensively, the original features are further added to the fused ones over a residual connection, giving $\tilde{S}_r$ as,

$$\tilde{S}_r = \mathcal{Z}_S(\mathcal{G}^S \odot (S_r \oplus P_r^S)) \oplus S_r \,, \quad \tilde{S}_r \in \mathbb{R}^{N_S \times d}, \tag{3}$$

where, $\mathcal{Z}_S$ is a transformation having a ReLU activation on a linear layer; $\odot$ denotes Hadamard product and $\oplus$ is element-wise summation. Similarly we get $\tilde{P}_r$ from S_r^P and P_r^S as:

$$\mathcal{G}^P = \texttt{sigmoid}([P_r, S_r^P] \cdot \mathbf{W}_\mathcal{G}^P) \;; \quad \tilde{P}_r = \mathcal{Z}_P(\mathcal{G}^P \odot (P_r \oplus S_r^P)) \oplus P_r \,, \quad \tilde{P}_r \in \mathbb{R}^{N_P \times d}. \tag{4}$$

3.3 Learning Objective

Taking independent embedding of a sketch (S) as an anchor, Triplet loss [53] aims to minimise its distance from a true-match photo (P^+) while maximising that from a non-matching photo (P^-) in a joint embedding space. We however have two representations of one sketch due to cross-modal pairwise embedding w.r.t photos namely S^+_{final} (paired with P^+_{final}) and S^-_{final} (paired with P^-_{final}). On similar motivation, our loss aims at decreasing distance between S^+_{final} and P^+_{final}, while increasing it between S^-_{final} and P^-_{final}, as:

$$\mathcal{L}(S^+_{final}, S^-_{final}, P^+_{final}, P^-_{final}) = max\{0,\ \Delta + \mathcal{D}(S^+_{final}, P^+_{final}) - \mathcal{D}(S^-_{final}, P^-_{final})\} \quad (5)$$

where, $\mathcal{D}(a,b) = ||a-b||^2$ and Δ is the margin value. This loss trains our network in an end-to-end manner, guiding to learn the hierarchical representation of sketch for better retrieval.

4 Experiments

4.1 Experimental Settings

Datasets: We evaluate our method on FG-SBIR specific datasets of QMUL-Chair-V2 [47] and QMUL-Shoe-V2 [36, 47], along with SWIRE [19] dataset which was curated using GUI examples from the RICO [9] dataset. Out of 6,730 sketches and 2,000 photos in QMUL-Shoe-V2, 6,051 sketches and 1,800 photos were used for training, while the rest for testing. A training/testing split of 1,275/725 sketches and 300/100 photos respectively, has been used for QMUL-Chair-V2 dataset. RICO [9] contains 72,000 examples of graphical user interfaces (GUI) from 9,700 free Android apps, out of which Huang *et al.* [19] used a subset of 167 apps to curate 3,802 sketches of 2,201 GUI examples. For a fairer evaluation, we do not repeat interfaces from the same apps between training and testing sets.

Implementation Details: ImageNet pre-trained InceptionV3 network (excluding auxiliary branch) is used as a backbone feature extractor on 299×299 resized images. Based on selective search [50], top 16 regions are considered empirically for photo branch. As coordinate information is absent in sketches from SWIRE [19], individual connected components (treated as strokes) are used, to obtain their bounding box information. For QMUL datasets, bounding boxes are calculated from available coordinate stroke information. Our model is implemented in PyTorch, taking around 2.3 mil. parameters which is roughly 5% more than a Siamese network baseline model (§4.2 B-Siamese). It is trained with Adam optimiser, using learning rate of 0.0001, batch size of 16, a margin (Δ) of 0.5, temperature (τ) of 1.0 and embedding size $d = 512$, for upto 200 epochs on a TitanX 12-GB GPU. Performance is evaluated using percentage of sketches having true-match photos appearing in the top-1 (acc.@1) and top-10 (acc.@10) lists.

4.2 Competitors

We evaluate our model against three groups of competitors: *(i)* Evaluation against existing *state-of-the-arts* (SoA): **Triplet-SN** [53] uses Sketch-a-Net [59] as a baseline feature extractor trained using triplet ranking loss. **Triplet-Attn-SN** [46] further extended Triplet-SN employing spatial attention along with higher order ranking loss. **SWIRE** [19] uses basic VGG-A network trained with typical triplet loss for the recently introduced GUI retrieval. *(ii)* As paired-embedding is ignored in SoAs, shadowing cross-modal retrieval literature [15, 20, 51] we design a few Baselines (B) *employing paired-embeddings*: **B-Siamese** is a naive baseline built similar to TripletSN but replaces Sketch-a-Net with more advanced

Table 1: Quantitative comparisons against various methods.

Methods		Chair-V2		Shoe-V2		SWIRE	
		acc.@1	acc.@10	acc.@1	acc.@10	acc.@1	acc.@10
State-of-the-arts	Triplet-SN [58]	45.65	84.24	28.71	71.56	-	-
	Triplet-Attn-SN [46]	56.54	88.15	31.74	74.78	-	-
	SWIRE [19]	-	-	-	-	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**

InceptionV3 as backbone feature extractor. **B-Gated-Siamese** involves paired embedding by employing a *matching gate* [51] between spatial feature-maps from intermediate layers of photo and sketch networks. **B-Localised-Coattn** harbours paired embeddings by employing co-attention between local photo-sketch sub-regions [53], without any hierarchical modelling. **B-Graph-Hierarchy** models a graph-based method. While typical graph convolutional networks (GCNs) are inherently flat with no hierarchical mechanism, we employ DIFFPOOL [57] on local region-features from each branch along with cross-modal co-attention, thus reducing the number of nodes every time by 1, until a final d-dimensional vector is obtained. *(iii)* In context of FG-SBIR, *we verify the potential of recent sketch-embedding techniques* under category-level SBIR. **SketchBERT-Variant** uses a transformer encoder as a sketch feature extractor on the five-point sketch coordinate representation inspired from a recent sketch classification/retrieval work [27]. **SketchFormer-Variant** [40] focusses on jointly embedding coordinate (via Transformer) and raster image (f_θ) representations, by concatenating and passing them via a two layer MLP, as a sketch query. Similar sketch-embedding topology has been recently used in SketchMate [56] and LiveSketch [8], where RNN embeds coordinate representation. For both these variants, the transformer architecture is designed following SketchBERT. We take the input via a normal linear layer and final feature is max-pooled across time, for classification.

4.3 Performance Analysis

Comparative performance results are shown in Table 1. **(i)** The inferior results of *Triplet-SN* and *Triplet-Attn-SN* are partially due to their apparently weaker backbone feature extractor of Sketch-A-Net. As for *SWIRE* [19] we suspect this is due to its inefficient training strategy. Using a stronger InceptionV3 backbone in *B-Siamese* does not add much either as it uses independent sketch-photo embeddings for retrieval. **(ii)** *B-Gated-Siamese* boosts scores as it introduces cross-modal feature matching at *spatial feature level*, thus creating more robust feature representation for retrieval while *B-Localised-Coattn* increases it further owing to *localised region-wise co-attention* based feature aggregation, from both modalities. **(iii)** *B-Graph-Hierarchy* represents the strongest alternative to our method, however it performs consistently worse than ours. This is because in our model, at every hierarchical level, only the two nodes selected to be merged are updated, while all others remain the same. This importantly preserves the integrity of the untouched nodes, so that they can better take part in cross-modal interaction at higher levels. On the contrary for graph-based methods, all nodes are updated via information passing at every level, which dilutes node features for consequent cross-modal interaction. Furthermore, graph-based approaches dictate the provision of a predefined adjacency matrix. As no such predefined rules are available for our problem,

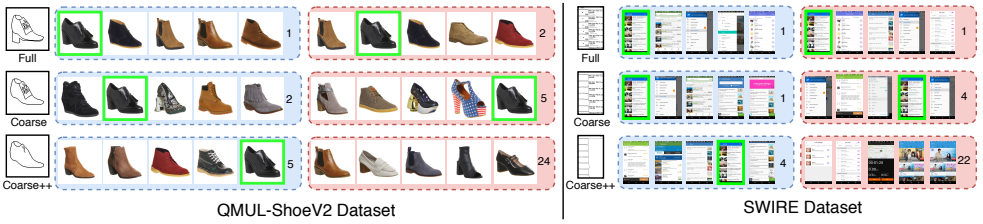

Figure 3: Our method's (blue) efficiency over *B-Siamese* (red) at varying extent (*coarse, coarse++*) of sketch *details* is shown (§4.4). Numbers denote rank of the matching photo.

we simply assume a heuristic of all nodes being fully connected. Our model on the other hand can discover the hierarchical structure from the node features themselves, without the need for such adjacency matrices. **(iv)** Both *Sketchformer-Variant* and *SketchBERT-Variant* use independent embedding of sketches and photos, exploring the coordinate information from a sketch using recent Transformer based frameworks. As GUI sketches do not contain temporal stroke information, this baseline was not performed on them. Despite their recent success in category-level SBIR [40, 56], results on QMUL datasets show a drop in performance even after collating the rasterised sketch-image with coordinate information in *Sketchformer-Variant*. This implies that coordinate representation of a sketch is not ideally suited for the cross-modal FG-SBIR tasks.

4.4 Ablation Study

Is hierarchy useful for FG-SBIR? To answer this we focus on GUIs [19], since they follow a more explicit hierarchy compared to free-hand sketches. Design elements in GUIs and their sketches exhibit a hierarchical relationship defined by containment, e.g., larger rectangular boxes representing window panes encompassing smaller ones symbolising buttons. This hierarchical information is distinctly defined in the meta-data released with the RICO [9] dataset (an example shown in Figure. 4). This means both sketch and photo branches already hold an *explicit* hierarchy, bypassing the need for explicit hierarchy discovery. If hierarchies are at all helpful for FG-SBIR, then using explicit hierarchies (with only the hierarchy discovery module (Equation 1) removed), would provide a weak upper-bound towards sketch-GUI retrieval. Such explicit hierarchical modelling would just remove the number of nodes at every level by one, until a final d-dimensional vector is obtained. The result obtained using an *Explicit hierarchy* (71.54%) is more than that obtained using an *Implicit* one (67.23%). This justifies that devising an optimal hierarchy would lead to higher retrieval performance, and shows that our implicit hierarchy discovery module can reach a performance close to its upper bound. As sketches from QMUL-ShoeV2 and QMUL-ChairV2 lack predefined hierarchical structure, implicit hierarchy discovery becomes the only option.

Further Analysis: (i) Had we skipped cross-modal feature matching (**w/o Localised-Coattn**) at every hierarchical level, the performance would drop by 10.6% (4.45%) for QMUL-ChairV2 (QMUL-ShoeV2) respectively (Table 2); however it has a lower time cost. On the other hand, eliminating hierarchical modelling (**w/o Hierarchy**) drops the performance by 7.21% (3.06%). This justifies the contribution of both these modules in our method. (ii) Instead of a distance based ranking in the shared embedding space between query and target photo, one alternative could be to predict their *scalar similarity value* by passing the concatenated features through a two-layer MLP followed by a sigmoid as inspired from [51, 53]. Empirically, performance drops to 53.32% (32.65%) justifying the superiority of distance based ranking. (iii) It has been observed in QMUL datasets that the first few contour strokes

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

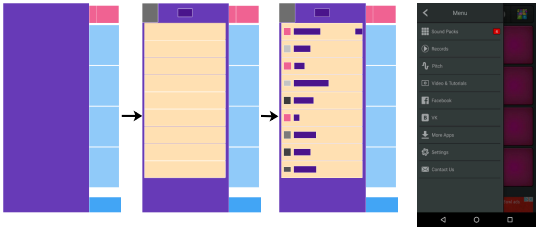

Figure 4: *Pre-defined* layout-order in a GUI

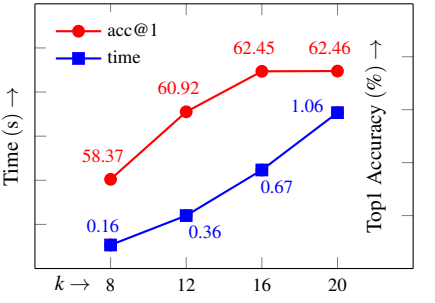

Figure 5: Study on number of regions (k)

are usually long [28], constituting majority ($\approx$ 50%) of the whole sketch (pixel-wise) thus denoting a coarse version of the sketch, while later strokes denote fine-grained details [28]. To justify our claim of learning a robust sketch-representation against varying levels of detail, we test our model on sketches where out of all strokes remaining after 50% pixel-wise sketch completion; (a) half (*Sketch-coarse*) and (b) all (*Sketch-coarse++*); are dropped (Figure 3). For SWIRE, we drop inner connected components to do the same. It may be noted that these two settings would yield different, yet similar number of hierarchical levels during training. More specifically, given a sketch-photo pair, the sketch, being typically incomplete, will result in its corresponding hierarchy to reduce to one node sooner compared to the photo, whereas the hierarchical fusion will continue in the photo branch, till it reduces to one node as well. With retrieval being the objective, we achieve relatively stable performance scoring acc.@10 of 87.58% (77.23%) and 85.64% (75.91%) with *Sketch-coarse* and *Sketch-coarse++* respectively, on QMUL-ChairV2 (QMUL-ShoeV2) datasets. The same for *B-Siamese* however falls to 75.32% (62.68%) and 65.31% (54.32%). Please refer to Table 1 for comparison against original acc.@10 values. Table 2 shows that even acc.@1 for our method does not drop much in such settings. Qualitative comparisons are given in Figure 3. Increasing number of regions (k) chosen for the photo branch (Figure. 5) elevated time cost, while decreasing so chipped at accuracy, proving 16 to be optimal.

5 Conclusion

In this paper, we studied a further intrinsic trait of sketches – that they are hierarchical in nature. The benefit of modelling and utilising hierarchies is demonstrated for the problem of FG-SBIR. Our model learns to discover a hierarchy that is implicit to a sketch. This ensures that no matter what level of detail an user sketches to, the model shall be able to retrieve the same/similar image fairly accurately. Unlike earlier approaches, our network enforces a cross-modal co-attention between a sketch and a photo, so as to arrive at a better cross-modal embedding for retrieval. Extensive experiments show our model to outperform most existing approaches by a significant margin, on both image and GUI retrieval.

References

[1] 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.

[2] Tu Bui, Leonardo Ribeiro, Moacir Ponti, and John Collomosse. Deep manifold alignment for mid-grain sketch based image retrieval. In *ACCV*, 2018.

[3] Xinchi Chen, Xipeng Qiu, Chenxi Zhu, and Xuan-Jing Huang. Gated recursive neural network for chinese word segmentation. In *ACL-IJCNLP*, 2015.

[4] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *SSST*, 2014.

[5] Jihun Choi, Kang Min Yoo, and Sang-goo Lee. Learning to compose task-specific tree structures. In *AAAI*, 2018.

[6] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *NeurIPS*, 2015.

[7] John Collomosse, Tu Bui, Michael J Wilber, Chen Fang, and Hailin Jin. Sketching with style: Visual search with sketches and aesthetic context. In *ICCV*, 2017.

[8] John Collomosse, Tu Bui, and Hailin Jin. Livesketch: Query perturbations for guided sketch-based visual search. In *CVPR*, 2019.

[9] Biplab Deka, Zifeng Huang, Chad Franzen, Joshua Hibschman, Daniel Afergan, Yang Li, Jeffrey Nichols, and Ranjitha Kumar. Rico: A mobile app dataset for building data-driven design applications. In *UIST*, 2017.

[10] Tyler Derr, Yao Ma, and Jiliang Tang. Signed graph convolutional networks. In *ICDM*, 2018.

[11] 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.

[12] Frederik Diehl. Edge contraction pooling for graph neural networks. *CoRR*, 2019.

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

[14] Hongyang Gao and Shuiwang Ji. Graph u-nets. In *ICML*, 2019.

[15] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 2014.

[16] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *CVPR*, 2018.

[17] David Ha and Douglas Eck. A neural representation of sketch drawings. In *ICLR*, 2018.

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

[19] Forrest Huang, John F Canny, and Jeffrey Nichols. Swire: Sketch-based user interface retrieval. In *CHI*, 2019.

[20] Sung Ju Hwang and Kristen Grauman. Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *IJCV*, 2012.

[21] Ozan Irsoy and Claire Cardie. Deep recursive neural networks for compositionality in language. In *NeurIPS*, 2014.

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

[23] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *CVPR*, 2015.

[24] Junhyun Lee, Inyeop Lee, and Jaewoo Kang. Self-attention graph pooling. In *ICML*, 2019.

[25] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, 2018.

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

[27] 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.

[28] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, 2017.

[29] Qing Liu, Lingxi Xie, Huiyu Wang, and Alan Yuille. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *ICCV*, 2019.

[30] Xian-Ming Liu, Rongrong Ji, Changhu Wang, Wei Liu, Bineng Zhong, and Thomas S Huang. Understanding image structure via hierarchical shape parsing. In *CVPR*, 2015.

[31] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.

[32] Umar Riaz Muhammad, Yongxin Yang, Timothy Hospedales, Tao Xiang, and Yi-Zhe Song. Goal-driven sequential data abstraction. In *ICCV*, 2019.

[33] Tsendsuren Munkhdalai and Hong Yu. Neural tree indexers for text understanding. In *ACL*, 2017.

[34] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *CVPR*, 2017.

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

[36] 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.

[37] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018.

[38] Yonggang Qi, Yi-Zhe Song, Tao Xiang, Honggang Zhang, Timothy Hospedales, Yi Li, and Jun Guo. Making better use of edges via perceptual grouping. In *CVPR*, 2015.

[39] Jose M Saavedra, Juan Manuel Barrios, and S Orand. Sketch based image retrieval using learned keyshapes (lks). In *BMVC*, 2015.

[40] Leo Sampaio Ferraz Ribeiro, Tu Bui, John Collomosse, and Moacir Ponti. Sketchformer: Transformer-based representation for sketched structure. In *CVPR*, 2020.

[41] Abhishek Sharma, Oncel Tuzel, and Ming-Yu Liu. Recursive context propagation network for semantic scene labeling. In *NeurIPS*, 2014.

[42] Yuming Shen, Li Liu, Fumin Shen, and Ling Shao. Zero-shot sketch-image hashing. In *CVPR*, 2018.

[43] Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. Parsing natural scenes and natural language with recursive neural networks. In *ICML*, 2011.

[44] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. In *NeurIPS*, 2013.

[45] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, 2013.

[46] 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.

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

[48] Damien Teney, Lingqiao Liu, and Anton van Den Hengel. Graph-structured representations for visual question answering. In *CVPR*, 2017.

[49] Giorgos Tolias and Ondrej Chum. Asymmetric feature maps with application to sketch based retrieval. In *CVPR*, 2017.

[50] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *IJCV*, 2013.

[51] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016.

[52] Fang Wang, Le Kang, and Yi Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *CVPR*, 2015.

[53] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *ICCV*, 2019.

[54] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 1992.

[55] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017.

[56] Peng Xu, Yongye Huang, Tongtong Yuan, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, Zhanyu Ma, and Jun Guo. Sketchmate: Deep hashing for million-scale human sketch retrieval. In *CVPR*, 2018.

[57] Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. In *NeurIPS*, 2018.

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

[59] 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.

[60] Long Zhu, Yuanhao Chen, Yuan Lin, Chenxi Lin, and Alan Yuille. Recursive segmentation and recognition templates for image parsing. *TPAMI*, 2011.