More Photos are All You Need: Semi-Supervised Learning for Fine-Grained Sketch Based Image Retrieval

Ayan Kumar Bhunia[1]   Pinaki Nath Chowdhury[1,2]   Aneeshan Sain[1,2]   Yongxin Yang[1,2]
Tao Xiang[1,2]   Yi-Zhe Song[1,2]
[1] SketchX, CVSSP, University of Surrey, United Kingdom
[2] iFlyTek-Surrey Joint Research Centre on Artificial Intelligence.
{a.bhunia, p.chowdhury, a.sain, yongxin.yang, t.xiang, y.song}@surrey.ac.uk

Abstract

A fundamental challenge faced by existing Fine-Grained Sketch-Based Image Retrieval (FG-SBIR) models is the data scarcity – model performances are largely bottlenecked by the lack of sketch-photo pairs. Whilst the number of photos can be easily scaled, each corresponding sketch still needs to be individually produced. In this paper, we aim to mitigate such an upper-bound on sketch data, and study whether unlabelled photos alone (of which they are many) can be cultivated for performances gain. In particular, we introduce a novel semi-supervised framework for cross-modal retrieval that can additionally leverage large-scale unlabelled photos to account for data scarcity. At the centre of our semi-supervision design is a sequential photo-to-sketch generation model that aims to generate paired sketches for unlabelled photos. Importantly, we further introduce a discriminator guided mechanism to guide against unfaithful generation, together with a distillation loss based regularizer to provide tolerance against noisy training samples. Last but not least, we treat generation and retrieval as two conjugate problems, where a joint learning procedure is devised for each module to mutually benefit from each other. Extensive experiments show that our semi-supervised model yields significant performance boost over the state-of-the-art supervised alternatives, as well as existing methods that can exploit unlabelled photos for FG-SBIR.

1. Introduction

With the ever rising popularity of touch screen devices, sketch-based image retrieval (SBIR) has witnessed significant interest within the vision community [5, 36, 37, 54, 11, 49]. Despite starting as a category-level retrieval problem [9, 10, 3, 12], the fine-grained nature of sketches stirred current research focus more towards fine-grained SBIR (FG-SBIR) [5, 30] – which aims to retrieve a *particular* photo based on a query sketch at an intra-category basis.

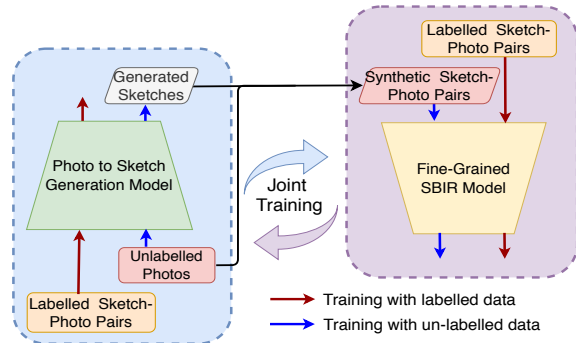Recent FG-SBIR works [52, 28, 5, 30] predominantly

Figure 1. Our proposed method additionally leverages large scale photos without any manually labelled paired sketches to improve FG-SBIR performance. Moreover, we show that the two conjugate process, *photo-to-sketch* generation and *fine-grained SBIR*, could improve each other by joint training.

rely on *fully-supervised* triplet loss-based deep networks to yield retrieval performances of practical value. The underlying assumption is largely inline with the progression of supervised photo-only models – that one can always (relatively easily) obtain additional labelled training data to sustain desired performance gains. This assumption however does not hold for FG-SBIR – sketch-photo pairs can not be easily scaled as per their photo-only counterparts. That is, instead of crawling and then labelling photos, the corresponding sketch for any given photo will need to be separately drawn by hand. As a result, current FG-SBIR datasets still remain in their thousands (6.7K for QMUL-ShoeV2 [52, 43], and 2K for QMUL-ChairV2 [43]), while photo-datasets [34] are available in millions. This data scarcity problem has consequently resulted in very recent attempts that aim at designing generalisable and zero-shot models [28], yet performances of these models remain far away from fully-supervised alternatives.

In this paper, we face the music and make the bold assumption that there will hardly be sufficiently large sketch-photo pairs to train a good model. Instead, we test the hypothesis that – freely-available *unlabelled photo data* would

1

help to mitigate the performance gap imposed by the lack of specifically collected *photo-sketch paired data*. Our utmost contribution is therefore a semi-supervised FG-SBIR framework where unlabelled photo data (i.e., photos without matching sketches) are used alongside photo-sketch pairs for model training. We differ significantly to conventional semi-supervised classification methods [40, 31] – other than learning pseudo photo labels via a learnable classifier, our "label" for a photo is in the form of a visual sketch which needs to be *generated* rather than *classified*. Thus, at the core of our design is a sequential photo-to-sketch *generation model* that outputs pseudo sketches for unlabelled photos. The hope is therefore that such pseudo sketch-photo pairs could augment the training of a *retrieval model*.

Naively cascading a generator with a retrieval model however would not work. This is mainly because off-the-shelf photo-to-sketch generation models [41, 7] could sometimes generate unfaithful sketches that may not resemble their corresponding photos, especially when it comes to fine-grained visual features. The downstream retrieval model would then have no way of knowing which pseudo sketch and photo pairs are worth training with, ultimately resulting in performance degradation. This leads to an important design consideration of ours – we advocate that there is positive complementarity between generation and retrieval that can be explored via *joint learning* (Figure 1). The intuition is simple – pseudo sketches automatically generated from unlabelled photos can help to semi-supervise a better retrieval model, *and vice versa* that better retrieval model can feed back to the generator in producing more faithful sketch-photo pairs.

The key therefore lies with *how* such positive exchange cycle can be facilitated between the generator and retrieval model. To this end, novelty lies in the components introduced in both generator and retrieval model designs, and in how they are jointly trained. *First*, we formulate a novel sequential photo-to-sketch generator with spatial resolution preservation and a cross-modal 2D-attention mechanism. *Second*, a discriminator is formulated in the retrieval model, to quantify the reliability of generated pseudo photo-sketch pairs. Reliability scores are then used for instance-wise weighting of triplet-loss values upon updating the retrieval model. A consistency loss (via distillation) is further introduced to simultaneously suppress the noisy training signal coming from pseudo photo-sketch pairs. *Third*, to enable exchange from retrieval to generation, we rely on the following intuition – good synthetic pairs would trigger a low value on the resulting triplet loss and a higher output of the discriminator. Feeding these training signals back to the generator would however involve passing through a non-differentiable rasterization operation (Figure 2). We thus employ a policy-gradient [44] based reinforcement learning scheme that feeds back these signals as *rewards*.

In summary, our contributions are: (a) For the first time, we propose to solve the data scarcity problem in FG-SBIR by adopting *semi-supervised* approach that additionally leverages large scale unlabelled photos to improve retrieval accuracy. (b) To this end, we couple sequential sketch generation process with fine-grained SBIR model in a joint learning framework based on reinforcement learning. (c) We further propose a novel photo-to-sketch generator and introduce a discriminator guided *instance weighting* along with *consistency loss* to retrieval model training with noisy synthetic photo-sketch pairs. (d) Extensive experiments validate the efficacy of our approach for overcoming data scarcity in FG-SBIR (Figure 4) – we can already reach performances at par with prior arts with just a fraction ($\approx$60%) of the training pairs, and obtain state-of-the-art performances on both QMUL-Shoe and QMUL-Chair with the same training data (by $\approx$6% and $\approx$7% respectively).

2. Related Works

Fine-Grained SBIR: Yu *et al*. [52] introduced the first deep FG-SBIR model which employed a deep triplet network to learn a common embedding space for photo and sketch. Subsequent works have aimed at improving this via attention mechanisms with higher order retrieval loss [43], joint discriminative-generative learning with cross-modal image generation [29], text tags [42], and cross-modal hierarchical co-attention [35]. Cross-category generalisation [28] and on-the-fly retrieval setup [5] are more recent additions to existing FG-SBIR literature. These fully supervised methods suffer from *the* data scarcity, which we aim to address.

Handling *Data-Scarcity* for FG-SBIR: Earlier works have tried resolving the lack of instance-level photo-sketch paired data, by using edge-maps for training [32] or synthetic sketch stroke deformation [53, 52] for data augmentation. Umar *et al*. [33, 26] leveraged reinforcement-learning (RL) in an attempt to augment sketches from edge-maps under the assumption that real sketch-strokes are a subset of edge-maps, which however is negated by the highly abstracted nature of real sketches. Very recently, mixed-modality jigsaw solving [30] has been used as a pre-training task for FG-SBIR to exploit additional photo images and their edge maps for cross-modal matching. Its efficacy remains limited however as edge-maps are not sketches.

Photo-to-Sketch Generation: A plausible solution to data scarcity is synthesising sketches for unlabelled photos to form pseudo photo-sketch pairs. Existing photo-to-sketch generation methods can be classified into two types: the first employs image-to-image translation [25], which however merely works as a contour detection paradigm, thus failing to model the hierarchically abstracted nature of human-drawn sketch. The second follows the seminal work of Sketch-RNN [16], and generates sequential sketch-coordinates given a photo, thus mimicking subjective hu-

man sketching style. The basic design [7], involving a CNN encoder and RNN decoder, has been further augmented with both self-domain and two way cross-modal reconstruction losses [41]. Following this path, we improve sequential sketch-generative process with a 2D attention mechanism to better exploit the spatial-layout of objects in photos.

Semi-supervised Learning: Our learning framework is semi-supervised in the sense that the majority of training data are unlabelled photos without their paired sketches. This is thus very different from most existing semi-supervised learning methods which are designed for classification rather than cross-modal retrieval. This means that these methods, based on either pseudo-labelling [24, 27, 40, 31] or consistency regularisation [2, 1], offer little insight into how our problem can be solved. In contrast, prior works on semi-supervised cross-modal learning such as image captioning [6, 23] are more relevant. However, we uniquely address a cross-modal instance-level retrieval problem, and train the generator jointly with the retrieval model, rather than merely providing model pre-training.

3. Methodology

Overview: For semi-supervised fine-grained SBIR, we consider having access to a limited amount of labelled photo-sketch pairs $\mathcal{D}_{\mathrm{L}} = \{(p_L^i; s_{p,L}^i)\}_i^{N_L}$ and a much bigger set of unlabelled photos $\mathcal{D}_{\mathrm{U}} = \{p_U^i\}_i^{N_U}$, where $N_U \gg N_L$. The key objective is to improve retrieval performance using both $\mathcal{D}_L$ and unlabelled photos $\mathcal{D}_U$ (having no corresponding sketches). More specifically, our framework consists of two models learned jointly: a FG-SBIR model, and a photo-to-sketch generation model. The retrieval model tries to learn an embedding function $\mathcal{F}(\cdot) : \mathbb{R}^{H \times W \times 3} \to \mathbb{R}^d$ mapping any rasterized sketch or photo having height H and width W to a d-dimensional feature.

Instead of image-to-image translation [21, 25], sketch generation process needs to be designed by sequential sketch-coordinate decoding [16] in order to model the hierarchical abstract nature of sketch. In particular, the FG-SBIR model requires rasterized sketch-images to obtain the sketch embedding, as performance can collapse on using sketch-coordinate instead [5, 35]. Thus, the generator learns a function $\mathcal{G}(\cdot) : \mathbb{R}^{H \times W \times 3} \to \mathbb{R}^{T \times 2}$, mapping a photo to equivalent sequential sketch-coordinate points $s_c = \{(x_1, y_1), (x_2, y_2), \cdots, (x_T, y_T)\}$, where T is the number of points. Note that in order to feed the generated sketches to the feature embedding network $\mathcal{F}$ of the retrieval model, a *rasterization* (sketch-image redrawing from coordinates) operation is required which is denoted as $s_p = \phi(s_c) : \mathbb{R}^{T \times 2} \to \mathbb{R}^{H \times W \times 3}$. Finally, we can create synthesised photo-sketch pairs $\mathcal{D}'_{\mathrm{U}} = \{(p_U^i; s_{p,U}^i)\}_i^{N_U}$ for unlabelled photos to train the retrieval model. Once trained, only $\mathcal{F}(\cdot)$ is used for retrieval during inference, while $\mathcal{G}(\cdot)$ augments pseudo/synthetic photo-sketch pairs for training.

3.1. Photo to Sketch Generation Model

The existing sequential photo-to-sketch generation models [41, 7] comprise a convolutional image encoder, followed by an LSTM decoder. This however has two major limitations: Firstly, it reduces the photo representation to a latent vector, leading to significant spatial information loss. Secondly, one fixed global representation is given as input at every time step of the LSTM decoding. To overcome these limitations two novel designs are introduced: (a) keeping the spatial feature map while ignoring global average pooling; (b) looking back at the specific part of photo which it *draws*. Overall, it consists of three major components, a CNN encoding the photo, a 2-D attention module, and an LSTM decoder generating the coordinates sequentially.

Given a photo p, let the extracted convolutional feature map be $\mathcal{B} \in \mathbb{R}^{h' \times w' \times c}$ where h', w' and c denotes the height, width and number of channels, respectively. Next, we perform a global average pooling on $\mathcal{B}$ to obtain a vector of size $\mathbb{R}^c$, and project it as two vectors μ and σ, each having size $\mathbb{R}^{N_z}$. The global embedding of photo is obtained through a reparameterization trick as $z = \mu + \sigma \odot \mathcal{N}(0, 1)$. The initial hidden state h_0 (and optional cell state c_0) of decoder RNN is initialised as $[h_0; c_0] = \tanh(W_z z + b_z)$.

Instead of predicting the absolute coordinates $\{(x_i, y_i)\}_i^T$, we model every point as 5 element vectors $(\Delta x, \Delta y, p_1, p_2, p_3)$ where Δx and Δy represents the off-set distances [16] in the x and y directions from the previous point. The last three elements represent a binary one-hot vector of three pen-state situations: pen touching the paper, pen being lifted and end of drawing. Each offset-position $(\Delta x, \Delta y)$ is modelled using a Gaussian mixture model (GMM) with $M = 20$ bivariate normal distributions [16] given by:

$$p(\Delta x, \Delta y) = \sum_{j=1}^{M} \Pi_j \mathcal{N}(\Delta x, \Delta y \mid \lambda_j); \sum_{j=1}^{M} \Pi_j = 1. \quad (1)$$

Each M bivariate normal distribution has *five* parameters $\lambda = \{\mu_x, \mu_y, \sigma_x, \sigma_x, \rho_{xy}\}$ with mean (μ_x, μ_y), standard deviation (σ_x, σ_y) and correlation (ρ_{xy}). The mixture weights of the GMM is modelled by a categorical distribution of size $\mathbb{R}^M$. Thus every time step's output y_t modelled is of size $\mathbb{R}^{5M+M+3}$, which includes 3 logits for pen-state. At time step t, a recurrent decoder network updates its state $s_t = (h_t, c_t)$ as follows: $s_t = \mathrm{RNN}(s_{t-1}; [g_t, P_{t-1}])$ where g_t is the glimpse vector encoding the information from specific relevant parts of the feature map $\mathcal{B}$ to predict y_t; P_{t-1} is the last predicted point (start-token $P_0 = \{0, 0, 1, 0, 0\}$), $[\cdot]$ signifies a concatenation operation. The glimpse/context vector is obtained by 2D attention as follows:

$$\begin{cases} J = \tanh(W_{\mathcal{B}} \circledast \mathcal{B} + W_S h_{t-1}); \\ \alpha_{i,j} = \mathrm{softmax}(W_a^T J_{i,j}) \\ g_t = \sum_{i,j} \alpha_{i,j} \cdot \mathcal{B}_{i,j}; \ i = [1, ..h'], \ j = [1, ..w'] \end{cases} \quad (2)$$
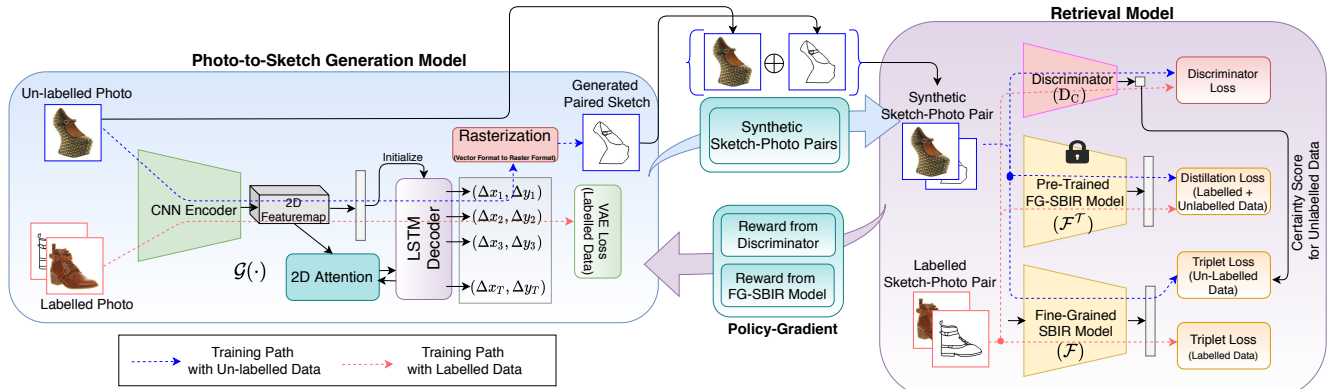
Figure 2. Our framework: a FG-SBIR model ($\mathcal{F}$) leverages large scale unlabelled photos using a *sequential* photo-to-sketch generation model ($\mathcal{G}$) along with labelled pairs. Discriminator (D_C) guided instance-wise weighting and distillation loss are used to guard against the noisy generated data. Simultaneously, $\mathcal{G}$ learns by taking reward from $\mathcal{F}$ and D_C via policy gradient (over both labelled and unlabelled) together with supervised VAE loss over labelled data. Note rasterization (vector to raster format) is a non-differentiable operation.

where W_B, W_S, W_a are the learnable weights. Calculating the attention weight $\alpha_{i,j}$ at every spatial position (i, j), we employ a convolution operation "⊛" with 3×3 kernel W_B to consider the neighbourhood information in the 2D attention module, and g_t is obtained by weighted summation operation at the end. A fully-connected layer over every hidden state outputs $y_t = W_y h_t + b_y$, where $y_t \in \mathbb{R}^{6M+3}$. We refer to [16] for more details. Like the standard VAE, our generator $\mathcal{G}$ is trained from the weighted summation of a reconstruction loss ($L_{\mathcal{G}}^R$) and a KL-divergence loss ($L_{\mathcal{G}}^{kl}$) with unit normal distribution as follows:

$$L_{\mathcal{G}}^{vae} = L_{\mathcal{G}}^R + \omega_{kl} L_{\mathcal{G}}^{kl}, \qquad (3)$$

where $L_{\mathcal{G}}^R$ is composed of the negative log-likelihood loss of the offsets $\Delta z = (\Delta x, \Delta y)$ and the pen states (p_1, p_2, p_3):
$L_{\mathcal{G}}^R = -\frac{1}{T}\Big[\sum_{i=1}^{T} \log p(\Delta z_i \mid \lambda_i) + \hat{p}_i \log(p_i) \Big]$.

3.2. Baseline FG-SBIR Model

For the discriminative retrieval module $\mathcal{F}(\cdot)$, we use the state-of-the-art Siamese network [43, 10, 5] (multi-branch with weight-sharing) with soft spatial attention [47] to focus on salient parts of the feature map. Concretely, given a photo or rasterized sketch image I, we use a pre-trained InceptionV3 model [45] to extract feature map $F' = f_\theta(I)$. This is followed by a residual connection between backbone feature and attention normalised feature to give $F = F' + F' \cdot f_{attn}(F')$, upon which global average pooling is performed to obtain final feature representation of size $\mathbb{R}^d$; and f_{attn} is modelled using 1x1 convolution with softmax across the spatial dimensions. For training, the distance to a sketch anchor (a) from a negative photo (n), denoted as $\beta^- = \|\mathcal{F}(a) - \mathcal{F}(n)\|_2$ should increase while that from the positive photo (p), $\beta^+ = \|\mathcal{F}(a) - \mathcal{F}(p)\|_2$ should decrease. This is brought about by the triplet loss with a margin $\mu > 0$ as a hyperparameter:

$$L_{\mathcal{F}}^{trip} = \max\{0, \mu + \beta^+ - \beta^-\}. \qquad (4)$$

3.3. Semi-Supervised Framework for FG-SBIR

Overview: Firstly, we train the photo-to-sketch generation model and discriminative fine-grained FG-SBIR model independently using the labelled training set $\mathcal{D}_L$. Thereafter, through our semi-supervised learning framework, $\mathcal{F}(\cdot)$ starts exploiting the unlabelled photos to improve its retrieval performance, while enhancing sketch generation quality of $\mathcal{G}(\cdot)$ by using $\mathcal{F}(\cdot)$ as a critic to provide training signal to the sketch-generation model $\mathcal{G}(\cdot)$ using both unlabelled and labelled photos simultaneously. Hence, both $\mathcal{G}(\cdot)$ and $\mathcal{F}(\cdot)$ can now improve itself with the help of each other and by exploiting unlabelled photos (see Figure 2).

Certainty Score for Synthetic Photo-Sketch Pair: The generated photo-sketch pairs $\mathcal{D}'_U$ of unlabelled photos are sometimes noisy compared to real labelled photo-sketch pairs $\mathcal{D}_L$. This is mainly due to large possible output space [41] of sketch drawing even with respect to a particular photo, as well as difficulties in predicting the sketch ending token [16] in the sequential decoding process. Every synthetic photo-sketch instance pair needs to be handled individually based on their quality, thus requiring a specific *certainty score* – signifying the reliability of synthetic photo-sketch pair to train the retrieval model. Existing semi-supervised classification approach usually considers the probability distribution over classes to filter out noisy samples based on a predefined threshold [48], top-K selection [40], or uses entropy-based instance-wise weighting [20] to deal with noisy synthetic labels. A new solution is thus needed to not just measure the quality of generated sketch itself, but to quantify how the generated sketch matches with the particular photo input, in order to help training the retrieval model.

Inspired by the generative adversarial network [8] where the sigmoid normalised output of discriminator shows the probability of being a real vs fake input sample, we use the *discriminator's confidence* to quantify the quality of syn-

4

thetic photo-sketch pairs. Specifically, the discriminator D_C learns to classify between real photo sketch pairs and generated pseudo photo-sketch pairs (concatenated across channels). Thus, the learning objective for D_C is

$$L_{D_C} = -\mathbb{E}_{(p_L;\, s_{p,L})\sim \mathcal{D}_L}\big[\log D_C(p_L, s_{p,L})\big]$$
$$-\mathbb{E}_{(p_U;\, s_{p,U})\sim \mathcal{D}'_U}\big[\log\big(1 - D_C(p_U, s_{p,U})\big)\big]. \quad (5)$$

This objective is computed via a binary cross-entropy loss using label 1 for real pairs, and 0 for synthetic ones. Thus, the discriminator's output $D_C(p_U, s_{p,U}) \in [0,1]$ signifies the extent to which the synthetic photo-sketch pairs match with the distribution of real labelled photo-sketch pairs. Therefore, values closer to 1 indicate better quality synthetic photo-sketch pairs.

Tolerance against Noisy Pseudo-Labelled Data: To further avoid over-fitting to noisy synthetic photo-sketch pairs, we introduce a consistency loss with respect to a pre-trained (on labelled dataset) retrieval model as *weak teacher* [13]. More specifically, once the baseline FG-SBIR model is trained from labelled data, we keep a copy as $\mathcal{F}^T$ with weights frozen. As $\mathcal{F}^T$ has been trained from real clean photo-sketch pairs only, we expect that the feature embedding vector obtained from it would act as an additional supervision via distillation [18] to regularise the main FG-SBIR model ($\mathcal{F}$) which is to be trained from both labelled and synthetic photo-sketch pairs in a semi-supervised manner. This distillation process is expected to improve the tolerance against noisy information of synthetic data. Compared to cross-entropy loss [18] used in distillation of classification network, a naive choice to design distillation for feature embedding network is to minimise the distance between learnable student's embedding and teacher's embedding of a particular photo or sketch image individually. We term it *absolute teacher*. However, instead of considering the actual embedding, we hypothesise that the *relative distance* between paired photo and sketch, minimising which is the major purpose of embedding network, could be a better knowledge to be distilled. We term this as *relative teacher*. Thus, given a photo-sketch image pair (p, s_p) and $d(\cdot, \cdot)$ being a l_2 distance function, the consistency loss for learnable student $\mathcal{F}$ with respect to pre-trained teacher $\mathcal{F}^T$ becomes as follows:

$$L_{\mathcal{F}}^{KD} = \big\| d\big(\mathcal{F}^T(p), \mathcal{F}^T(s_p)\big) - d\big(\mathcal{F}(p), \mathcal{F}(s_p)\big)\big\|_2. \quad (6)$$

3.4. Joint Training

Optimising FG-SBIR Model: We train the fine-grained SBIR model $\mathcal{F}$ using triplet loss over labelled photo-sketch pairs $\mathcal{D}_L$, *instance weighted triplet loss* over generated pseudo photo-sketch pairs $\mathcal{D}_U$, and pre-trained teacher based consistency loss over both $\mathcal{D}_L$ and $\mathcal{D}_U$. Given sampled data $\mathcal{D}_L^i = \{p_L^i, s_{p,L}^i\} \sim \mathcal{D}_L$ and $\mathcal{D}'^{j}_U = \{p_U^j, s_{p,U}^j\} \sim \mathcal{D}'_U$ (on *same* ratio), the instance-wise

weight is calculated as $\omega_j = D_C(\mathcal{D}'^{j}_U)$. The overall semi-supervised loss to train the retrieval model becomes:

$$L_{\mathcal{F}}^{all} = L_{\mathcal{F}}^{trip}(\mathcal{D}_L^i) + \omega_j \cdot L_{\mathcal{F}}^{trip}(\mathcal{D}'^{j}_U) + \lambda_{kd} \cdot L_{\mathcal{F}}^{KD}(\mathcal{D}_L^i, \mathcal{D}'^{j}_U) \quad (7)$$

Optimising Photo-to-Sketch Model: Besides the fully-supervised VAE loss $L_{\mathcal{G}}^R$, during joint training, the photo-to-sketch generation model is also learned considering $\mathcal{F}$ and D_C as critics. In particular, if the generated sketch from $\mathcal{G}$ correctly depicts the corresponding input photo, the triplet loss for that generated photo-sketch pair from retrieval model would be low, signifying a better photo-sketch matching and generated sketch quality. Similarly, the higher the discriminator's output, the better the quality of generated photo-sketch pairs. However, these training signals from $\mathcal{F}$ and D_C cannot be directly back-propagated to $\mathcal{G}$, as there exists a non-differentiable *rasterization* operation s_p before feeding the sketch-image to both retrieval model and discriminator. Hence, we employ reinforcement learning based on policy-gradient [44] with REINFORCE [46] deployed to estimate gradients with respect to parameters $\theta_{\mathcal{G}}$ of $\mathcal{G}$ given some *reward*. As $\mathcal{G}$ aims to lower this triplet loss value $L_{\mathcal{F}}^{trip}$ (Eqn. 4), the reward should be negative of $L_{\mathcal{F}}^{trip}$ that needs to be maximised. Similarly, the discriminator's output quantifying the goodness of photo-sketch pairs needs to be maximised. Thus the weighted joint reward is:

$$R_{\mathcal{G}} = -\lambda_{r1} \cdot L_{\mathcal{F}}^{trip}(\mathcal{D}^i) + \lambda_{r2} \cdot D_C(\mathcal{D}^i) \quad (8)$$

where $\mathcal{D}^i \sim \mathcal{D}_L \cup \mathcal{D}_U$. This reward could be computed for both labelled and unlabelled data as it does not need any ground-truth sketch-coordinates unlike the $L_{\mathcal{G}}^{vae}$ loss (Eqn. 3). Thus two types of gradients are computed to update the parameter $\theta_{\mathcal{G}}$, one using policy gradient [44] based on joint-reward guided by the retrieval model and the discriminator, and the other using back-propagation over only the labelled photos:

$$\nabla_{\theta_{\mathcal{G}}} L(\theta_{\mathcal{G}}) = \underbrace{\nabla_{\theta_{\mathcal{G}}} L_{\mathcal{G}}^{vae}(\theta_{\mathcal{G}})}_{\text{over only labelled data}} \quad (9)$$
$$-\lambda_{\mathcal{G}} \sum_{i=1}^{T} \underbrace{\mathbb{E}_{\substack{p_i\sim p(q_i)\\ \Delta z_i \sim p(\Delta z_i|\lambda_i)}} \nabla_{\theta_{\mathcal{G}}}\Big(\log p(\Delta z_i \mid \lambda_i) + \log p(p_i)\Big) \cdot R_{\mathcal{G}}}_{\text{over both labelled and unlabelled data (via policy gradient)}}$$

In our experiments, we only update the final, fully-connected layer of sketch-decoder (with weights W_y, b_y predicting 6M+3 outputs at every time step), at times using policy gradient, keeping rest of the parameters of $\mathcal{G}$ fixed. We use a single global reward for the whole sketch-coordinate sequence, instead of local reward at every time step, that would otherwise need costly Monte Carlo roll-outs [51]. Note that in our design, $\mathcal{G}$ and D_C are connected in a GAN-like fashion [15] having adversarial objective. Moreover, the retrieval and generative models are trained alternatively improving each other over time (Algorithm 1).

4. Experiments

Datasets: Two publicly available datasets, QMUL-Shoe-V2 [28, 33, 41, 5] and QMUL-Chair-V2 [5, 41] are

Algorithm 1 Training of Semi-Supervised FG-SBIR
___
1: **Input**: Labelled photo-sketch pairs $\mathcal{D}_L$ and Unlabelled photos $\mathcal{D}_U$.
2: **Initialise hyper params**: k_r, k_g.
3: **Pre-training**: $\mathcal{G}$ and $\mathcal{F}$ from $\mathcal{D}_L$ (using $L_{\mathcal{G}}^{\text{vae}}$ & $L_{\mathcal{F}}^{\text{trip}}$).
4: **while** not done training **do**
5:     **for** k_r steps **do**
6:         Sample data $\mathcal{D}_L^i \sim \mathcal{D}_L$ and $\mathcal{D}_U^j \sim \mathcal{D}_U$.
7:         Get synthetic paired images $\mathcal{D'}_U^j$ using $\mathcal{G}(\cdot)$.
8:         TRAIN $\mathcal{F}$ using $\{\mathcal{D}_L^i, \mathcal{D'}_U^j\}$ ▷ Eqn. 7
9:         TRAIN D_C using $\{\mathcal{D}_L^i, \mathcal{D'}_U^j\}$ ▷ Eqn. 5
10:     **end for**
11:     **for** k_g steps **do**
12:         Sample data $\mathcal{D}_L^i \sim \mathcal{D}_L$ and $\mathcal{D}_U^j \sim \mathcal{D}_U$.
13:         Get reward $R_{\mathcal{G}}$ using $\mathcal{F}$ and D_C.
14:         TRAIN $\mathcal{G}$ using $\{\mathcal{D}_L^i, \mathcal{D}_U^j\}$ ▷ Eqn. 9
15:     **end for**
16: **end while**
17: **Output**: Optimised models $\mathcal{F}$, $\mathcal{G}$ and D_C.
___

used, which contain stroke-level coordinate information of sketches in addition to instance-wise paired sketch-photo labels, thus enabling us to train both retrieval and sketch-generative models. Out of the 6,730 sketches and 2,000 photos in Shoe-V2, 6,051 and 1,800 for training respectively, and the rest are for testing [5, 41]. The splits [5, 41] for Chair-V2 dataset are 1,275/725 sketches and 300/100 photos for training/testing respectively. In addition to these labelled training data, we further use all 50,025 UT-Zap50K images [50] as unlabelled photos for shoe retrieval, and 7,800 unlabelled chair photos [30] are collected from shopping websites, including IKEA, Amazon and Taobao. Data, code, and models will be released soon.

Implementation Details: Firstly, for sketch-generation, we use ImageNet pre-trained VGG-16 as encoder, excluding any global average pooling operation. We keep the dimension (N_z) of z as 128, the hidden state of the decoder LSTM as 512, the embedding dimension of the 2D-attention module as 256 respectively. We set the max sequence length to 100, and the generative model is trained with a batch size of 64 with $\omega_{kl} = 1$ using the pre-training strategy from [41]. Secondly, the retrieval model (ImageNet pre-trained Inception-V3 [45]) is trained with a batch-size of 16 with a margin value of 0.3. Finally, after completing individual training from labelled data, we start *joint training* (Section 3.4) by additionally exploiting unlabelled data using k_g and k_r as 5. Architecture of D_C is from [21]. We set $\lambda_{kd} = 0.1$, $\lambda_{r1} = 1$, $\lambda_{r2} = 1$, and $\lambda_{\mathcal{G}} = 10$ respectively. All images are resized to 256×256, with rasterization from sketch-coordinate involving a window of same size having centre scaling as well. We use Adam optimiser for both the generation and retrieval models with a learning rate of 0.0001.

Evaluation Metric: (a) FG-SBIR: Following existing FG-

SBIR works [52, 30], we use Acc@q, i.e. percentage of sketches having true-paired photo appearing in the top-q list. **(b) Sketch Generation:** Following [41, 4], sketch-generation is quantified from three perspectives (i) *Recognition:* Using a ResNet-50 classifier trained on 250-classes from TU-Berlin sketch dataset, a generated sketch getting recognised as the same class as that of corresponding photo signifies category-level realism. (ii) *Retrieval:*[1] To judge whether the generated sketch has object-instance specific agreement, we check the retrieval accuracy Acc@q via a pre-trained FG-SBIR model using the generated sketches to retrieve corresponding photos of the testing set. (iii) *Generation:* Following a recent sketch generation work [4], we further calculate FID-score [17] using a pre-trained sketch-classifier that captures both the quality and diversity of generated data compared to real human sketches.

4.1. Competitors

Sketch Generation: Sketch Generation could be approached in two following ways: (a) *Image-to-image translation* pipeline: **Pix2Pix** [21] could be adapted to perform cross-modal translation in the image space. **PhotoSketch** [25] extends further to handle the one-to-many possible nature of photo-conditioned sketch image generation problem, by calculating a mean loss over multiple sketches corresponding to a particular photo. (b) *Image-to-sequence generation* pipeline: **Pix2Seq** [7] is the ablated version of our model having a convolutional encoder and LSTM decoder, without involving 2D-attention. **L2S** [41] is an extension over [7] that uses two-way cross domain translation with self-domain reconstruction for better regularisation. **Ours-G** is a *supervised* model with 2D-attention, trained independently from labelled data only. **Ours-G-full** is our final sketch-generative model involving joint-training to learn from both labelled and unlabelled data.

Fine-Grained SBIR: We compare with three groups of competitors. (a) *state-of-the-art:* **SN-Triplet** [52] employs triplet ranking loss with Sketch-a-Net as its baseline feature extractor. **SN-HOLEF** [42] is an extension over [52] employing spatial attention along with higher order ranking loss. **SN-RL** [5] is a very recent work employing reinforcement learning based fine-tuning for on-the-fly retrieval. As early retrieval is not our objective, we cite result at sketch-completion point. (b) *Exploiting unlabelled photos:* There has been no prior work addressing semi-supervised learning for FG-SBIR, and model designed for category-level retrieval [22] does not fit here. We thus adopt a few works that could be used to leverage unlabelled photos. **Edgemap-Pretrain** [32] is a naive-approach to use edge-maps of unlabelled photos to pre-train the retrieval model. While edge-maps hardly have any similarity to real free-hand sketches,

___
[1]Note: Retrieval accuracy is used to quantify both FG-SBIR and sketch generation performance. Please refer to [41] for more details.

Table 1. Quantitative results of photo-to-sketch generation

Chair-V2	Recognition (↑)		Retrieval(↑)		FID Score(↓)
	Acc.@1	Acc.@10	Acc.@1	Acc.@10	
Pix2Pix [21]	4.5%	12.1%	2.4%	16.2%	33.4
PhotoSketch [25]	7.1%	14.3%	4.2%	17.9%	25.7
Pix2Seq [7]	5.4%	52.1%	4.0%	31.8%	14.5
L2S [41]	12.3%	53.8%	8.3%	36.7%	12.7
Ours-G (only labelled data)	15.2%	56.9%	13.4%	40.7%	10.1
Ours-G-Full	16.4%	58.6%	14.9%	42.6%	8.9

Shoe-V2	Recognition(↑)		Retrieval(↑)		FID Score (↓)
	Acc.@1	Acc.@10	Acc.@1	Acc.@10	
Pix2Pix [21]	6.2%	14.5%	1.8%	8.4%	31.7%
PhotoSketch [25]	8.9%	17.3%	3.4%	10.2%	24.3%
Pix2Seq [7]	51.3%	86.6%	5.1%	25.8%	11.3
L2S [41]	53.7%	89.7%	6.2%	28.6%	10.7
Ours-G (only labelled data)	56.3%	91.9%	9.7%	33.6%	9.5
Ours-G-Full	58.1%	93.4%	12.3%	35.4%	8.3

Table 2. Quantitative results of fine-grained SBIR

Methods	Chair-V2		Shoe-V2	
	Acc.@1	Acc.@10	Acc.@1	Acc.@10
SN-Triplet [52]	47.4%	84.3%	28.7%	71.6%
SN-HOLEF [42]	50.7%	86.3%	31.2%	74.6%
SN-RL [5]	51.2%	86.9%	30.8%	74.2%
Edgemap-Pretrain [32]	53.9%	87.7%	33.8%	80.9%
Edge2Sketch-Pretrain [33]	54.3%	88.2%	34.2%	81.2%
Jigsaw-Pretrain [30]	56.1%	88.7%	36.5%	85.9%
Ours-F (only labelled data)	53.3%	87.5%	33.4%	80.7%
Vanilla-SSL-F	49.6%	85.6%	30.6%	74.3%
Ours-F-Pix2Pix	53.2%	87.5%	33.2%	80.1%
Ours-F-L2S	57.6%	89.4%	36.6%	84.7%
Ours-F-Full	60.2%	90.8%	39.1%	87.5%

they could be converted to better pseudo-sketches using the work [33] that learns how to abstract sketches based on subset-stroke selection. We term it as **Edge2Sketch** [33]. Recently, **Jigsaw-Pretrain** [30] used jigsaw solving over the mixed patches between a particular photo (unlabelled) and its edge-map, as a pre-text task for self-supervised learning (SSL) to improve FG-SBIR performance. Furthermore, we term our self-implemented *supervised* FG-SBIR model trained only on labelled data as **Ours-F**. **Ours-F-Full** is our final retrieval model employing joint training over both labelled and unlabelled photos. We also replace our 2D-attention based sketch-generation process by baseline sketch-generative model Pix2Pix [21] and L2S [41], and term them as **Ours-F-Pix2Pix** and **Ours-F-L2S** respectively. Finally, we design a naive semi-supervised FG-SBIR baseline (**Vanilla-SSL-F**), where we *blindly* (without instance-weighting and distillation) use the generated sketch to additionally train the retrieval model.

4.2. Performance Analysis

Photo-to-Sketch Generation: From Table 1, we observe: **(i)** *Pix2Pix* and *PhotoSketch* based on cross-modal translation in pixel space perform poorly. They fail to capture the abstraction in human sketching style, where distribution gap with real sketches is reflected in their significantly poor FID scores. **(ii)** *Pix2Seq* and *L2S* outputs vector sketches by sequentially predicting sketch coordinates, thus possessing higher similarity towards human sketches. They however, still lag behind our ablated version *Ours-G* in scores. As both of them reduce the spatial dimension of the convolutional feature-map to a global context vector, spatial information is significantly compromised, with the decoder receiving little guidance from the vector on exact drawing content. In contrast, we retain the spatial dimension of feature-map, and employ 2D-attention to focus on that specific part of the photo it draws at any time step. **(iii)** *L2S* is a notably close competitor to ours in terms of recognition accuracy, but better information passage between every time step of decoder and convolutional encoder deliv-

ers much better sketches with fine-grained details (reflected by retrieval accuracy). Furthermore, our final model *Ours-G-Full* employs joint training with a retrieval model to additionally exploit the unlabelled photos, improving sketch generation performance (retrieval Acc@1) from 9.7% to 12.3% by 2.6% over our baseline *Ours-G* on Shoe-V2, thus justifying the benefits of our semi-supervised learning. Some qualitative results are shown in Figure 3. Blue denotes a supervised baseline, while red is *Ours-G(F)-Full*.

Fine-grained SBIR: From Table 2, we observe: **(i)** Our baseline retrieval model is noticeably better than *SN-Triplet*, and lies at par with recent state-of-the-art FG-SBIR baselines like *SN-HOLEF* and *SN-RL*. **(ii)** With regards to exploiting unlabelled photos, *Edgemap-Pretrain* offers marginal improvement while using it on top of our baseline with ImageNet pretrained weights. Aligning with the intuition, while edge-maps are further augmented with *Edge2Sketch* by a subset of stroke selection to model the abstracted nature of sketch over edge-maps, it increases retrieval performance by a reasonable margin. In context of using edge-maps for pre-training, *Jigsaw-Pretrain* provides maximum benefits, but still lags behind our final model *Ours-F-Full*. **(iii)** While edge-map does not posses sketch abstraction knowledge of human sketching style, our approach of using a sequential photo-to-sketch generation model to generate synthetic photo-sketch pairs for unlabelled photo encodes better knowledge to enhance generalisation. However, it is noteworthy that *Vanilla-SSL-F* blindly using synthetic sketch-photo pairs yields performance lower than the supervised one due to overfitting on noisy information. Overall, for fine-grained SBIR, due to our proposed semi-supervised learning, the retrieval accuracy Acc@1 of *Ours-F-Full* increases from 33.4% to 39.1% by a margin of 5.7% over our baseline *Ours-F* on Shoe-V2. Moreover, replacing our photo-to-sketch generation model by *L2S* and *Pix2Pix* reduces the same by 2.5% and 5.9% respectively, thus justifying the importance of our sketch generative model with 2D attention. **(iv)** Note that policy-gradient based RL scheme could be avoided by using Pix2Pix for sketch generation, and gradient can directly be back-propagated from retrieval to generative model. However, that is still found to be inferior to ours.

Figure 3. Qualitative results on our photo-to-sketch generation process, where sketch is shown with attention-map at progressive instances.

4.3. Ablation Study

A thorough ablative study on Shoe-V2 dataset verifies contributions of individual design components in Table 3. *[i]* **Instance weighting for retrieval:** To simply judge the contribution of discriminator (D_C) guided instance weighting we remove it, and adapt the framework accordingly. Consequently Acc@1 retrieval performance significantly drops to 36.8% with a decrease of 2.3% on Shoe-V2. Due to sigmoid normalisation [8], the output of D_C falls in [0,1]. We quantify it as 10 discrete levels with a step size 0.1. We calculate the average ranking percentile (ARP) of synthetic sketch-photo pairs from testing set which fall under the same discrete level, and plot it against 10 different levels. From Figure 4 (a), it is evident that the synthetic sketch-photo pairs having higher discriminator score (towards 1) tend to have much better ARP [5] values (i.e better quality), while those with lesser ARP values are assigned with lesser (towards 0) certainty score by the discriminator. This observation is consistent with our assumption that D_C should quantify quality of synthetic sketch-photo pairs for instance-wise weighting. *[ii]* **Distillation based noise tolerance for retrieval:** Removing knowledge distillation based regularisation, which additionally tries to provide tolerance against noisy synthetic sketch-photo pairs, Acc@1 is decreased by 1.8% to 37.3% on Shoe-V2 dataset. Our *relative teacher* based distillation process (Section 3.3) for retrieval network surpasses *absolute teacher* alternative by a margin of 0.9% (Acc@1) on Shoe-V2, thus confirming its usefulness. *[iii]* **2D-attention for sketch-generation:** The use of 2D attention significantly improves the sketch generation performance, providing better fine-grained agreement with the input photo. While we employ a 3x3 convolutional kernel to aggregate neighbourhood information, using an 1D attention that treats feature maps as 1D sequence, the retrieval accuracy Acc@1 of generated sketches on Shoe-V2 drops to 8.1% by margin of 4.2%. We conjecture that 2D-spatial attention has higher efficiency in generating fine-grained sequential sketches from input photo than two-way translation based regularisation as done in *L2S* [41]. *[iv]* **Significance of joint-training:** (a) A direct way of judging efficiency of joint training is employing separately trained photo-to-sketch generation model to augment synthetic sketch-photo pairs, and using them *blindly* to train the retrieval model along with labelled data *without* instance weighting or teacher-regularisation. This however lags be-
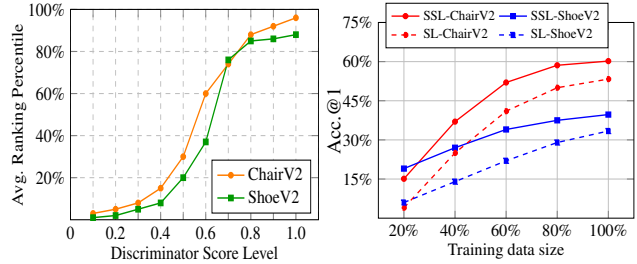

Figure 4. (a): Consistency of discriminator's certainty score. (b): Varying training data size for FG-SBIR - Semi-Supervised Learning (SSL) vs Supervised-Learning (SL).

hind the baseline (supervised) fine-grained SBIR model by 2.8% (i.e. 30.6%), as the model over-fits to the noisy information present in synthetic sketch-photo data. This confirms that naively using sketch generation does not help at all. (b) The retrieval accuracy (Acc@1) from sketch generation performance improves by 1.4% with an additional policy-gradient based training taking reward (Eqn. 8) from the retrieval model and the discriminator D_C as critic. Individually, they help to improve by 0.9% and 0.8%, respectively under the same metric. (c) Furthermore, we compute the performance of our semi-supervised framework at varying training data size for both processes in Figure 4 (b). We notice a significant overhead compared to our supervised baseline model for each dataset individually.

Table 3. Ablative study on Shoe-V2: Instance Weighting (IW), Teacher Regularisation (TR), Attention (AT), Joint-Training (JT).

IW	TR	AT	JT	Fine-Grained SBIR		Sketch Generation	
				Acc.@1	Acc.@10	Recognition Acc.@1	Retrieval Acc.@1
✓	✓	✓	✓	39.1%	87.5%	58.1%	12.3%
✗	✓	✓	✓	36.8%	85.4%	57.3%	11.2%
✓	✗	✓	✓	37.3%	86.1%	57.8%	12.1%
✓	✓	✗	✓	37.6%	86.1%	51.3%	5.1%
✓	✓	✓	✗	37.9%	86.6%	56.3%	9.7%
✗	✗	✗	✗	31.1%	75.4%	51.3%	5.1%

5. Conclusion

We have proposed a semi-supervised fine-grained sketch-based image retrieval framework to solve the data scarcity problem. To this end, we proposed to treat sequential photo-to-sketch generation and fine-grained sketch-based image retrieval as two conjugate problems along with various regularizers to address the intricate issues of reliability and tolerance to noisy synthetic sketch-photo pairs. This leads to substantial improvement on existing baselines in sparse data-scenarios for FG-SBIR.

References

[1] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *ICLR*, 2020. 3

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019. 3

[3] 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*, 2021. 1

[4] Ayan Kumar Bhunia, Ayan Das, Umar Riaz Muhammad, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, Yulia Gryaditskaya, and Yi-Zhe Song. Pixelor: A competitive sketching ai agent. so you think you can beat me? In *Siggraph Asia*, 2020. 6

[5] 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. 1, 2, 3, 4, 5, 6, 7, 8, 11

[6] Wenhu Chen, Aurelien Lucchi, and Thomas Hofmann. A semi-supervised framework for image captioning. *arXiv preprint arXiv:1611.05321*, 2016. 3

[7] Yajing Chen, Shikui Tu, Yuqi Yi, and Lei Xu. Sketchpix2seq: a model to generate sketches of multiple categories. *arXiv:1709.04121*, 2017. 2, 3, 6, 7

[8] LI Chongxuan, Taufik Xu, Jun Zhu, and Bo Zhang. Triple generative adversarial nets. In *NeurIPS*, 2017. 4, 8

[9] John Collomosse, Tu Bui, and Hailin Jin. Livesketch: Query perturbations for guided sketch-based visual search. In *CVPR*, 2019. 1

[10] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and YiZhe Song. Doodle to search: Practical zero-shot sketchbased image retrieval. In *CVPR*, 2019. 1, 4

[11] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *CVPR*, 2019. 1

[12] Titir Dutta, Anurag Singh, and Soma Biswas. Adaptive margin diversity regularizer for handling data imbalance in zeroshot sbir. In *ECCV*, 2020. 1

[13] Tommaso Furlanello, Zachary C Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *ICML*, 2018. 5

[14] Junlong Gao, Shiqi Wang, Shanshe Wang, Siwei Ma, and Wen Gao. Self-critical n-step training for image captioning. In *CVPR*, 2019. 11

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 5

[16] David Ha and Douglas Eck. A neural representation of sketch drawings. In *ICLR*, 2018. 2, 3, 4

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 6

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Workshop*, 2015. 5

[19] Minyoung Huh, Shao-Hua Sun, and Ning Zhang. Feedback adversarial learning: Spatial feedback for improving generative adversarial networks. In *CVPR*, 2019. 11

[20] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *CVPR*, 2019. 4

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 3, 6, 7

[22] Young Kyun Jang and Nam Ik Cho. Generalized product quantization network for semi-supervised image retrieval. In *CVPR*, 2020. 6

[23] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach. In *EMNLP*, 2019. 3

[24] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML*, 2013. 3

[25] Mengtian Li, Zhe Lin, Radomir Mech, Ersin Yumer, and Deva Ramanan. Photo-sketching: Inferring contour drawings from images. In *WACV*, 2019. 2, 3, 6, 7

[26] Umar Riaz Muhammad, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Goal-driven sequential data abstraction. In *ICCV*, 2019. 2

[27] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *NeurIPS*, 2019. 3

[28] 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. 1, 2, 5, 11

[29] Kaiyue Pang, Yi-Zhe Song, Tony Xiang, and Timothy M Hospedales. Cross-domain generative learning for finegrained sketch-based image retrieval. In *BMVC*, 2017. 2

[30] 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. 1, 2, 6, 7

[31] Hieu Pham, Qizhe Xie, Zihang Dai, and Quoc V Le. Meta pseudo labels. *arXiv preprint arXiv:2003.10580*, 2020. 2, 3

[32] Filip Radenovic, Giorgos Tolias, and Ondrej Chum. Deep shape matching. In *ECCV*, 2018. 2, 6, 7

[33] Umar Riaz Muhammad, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Learning deep sketch abstraction. In *CVPR*, 2018. 2, 5, 7, 11

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 1

[35] 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. 2, 3, 11

9

[36] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Stylemeup: Towards style-agnostic sketch-based image retrieval. In *CVPR*, 2021. 1

[37] Leo Sampaio Ferraz Ribeiro, Tu Bui, John Collomosse, and Moacir Ponti. Sketchformer: Transformer-based representation for sketched structure. In *CVPR*, 2020. 1

[38] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM TOG*, 2016. 11

[39] Firas Shama, Roey Mechrez, Alon Shoshan, and Lihi Zelnik-Manor. Adversarial feedback loop. In *ICCV*, 2019. 11

[40] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 2, 3, 4

[41] Jifei Song, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Learning to sketch with shortcut cycle consistency. In *CVPR*, 2018. 2, 3, 4, 5, 6, 7, 8

[42] Jifei Song, Yi-Zhe Song, Tony Xiang, and Timothy M Hospedales. Fine-grained image retrieval: the text/sketch input dilemma. In *BMVC*, 2017. 2, 6, 7

[43] 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. 1, 2, 4

[44] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NeurIPS*, 2000. 2, 5

[45] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 4, 6

[46] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 1992. 5

[47] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 4

[48] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019. 4

[49] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *ECCV*, 2018. 1

[50] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, 2014. 6

[51] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, 2017. 5

[52] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *CVPR*, 2016. 1, 2, 6, 7, 11

[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. 2

[54] Jingyi Zhang, Fumin Shen, Li Liu, Fan Zhu, Mengyang Yu, Ling Shao, Heng Tao Shen, and Luc Van Gool. Generative domain-migration hashing for sketch-to-image retrieval. In *ECCV*, 2018. 1

Supplementary material for
More Photos are All You Need: Semi-Supervised Learning for Fine-Grained Sketch Based Image Retrieval

A. Necessity for rasterization

This is common practice in the FG-SBIR literature. Rasterized sketch-images tend to have better spatial encoding than coordinate-based alternatives [28, 5]. This is also verified in our work – removing rasterization and using sketch-coordinate retrieval reduces acc@1 to only 7.6% on Shoe-V2.

B. Motivation behind using discriminator for certainty score:

We are mostly inspired by recent image generation works [39, 19] that use the discriminator scores to iteratively improve generation quality. We also did an ablative study to investigate further (see L768-785 and Fig. 4(a)). We found that synthetic sketch-photo pairs having higher discriminator score, tend to have much better quality, and vice-versa. We will add some qualitative examples to further illustrate this correlation in supplementary materials. Defining a hard threshold (optimised) to eliminate bad generated sketches is an option – new experiments show acc@1 lags by around 2% compared to ours on Shoe-V2.

C. More details on experimental setup and analysis:

(i) Our self-designed baselines use the same backbone network, while joint-training is employed for Ours-F-Pix2Pix, Ours-F-L2S and Ours-F-Full.

(ii) SOTA data-augmentation strategies are already adopted by existing FG-SBIR works [28, 35]. However, they fail to capture the significant style variations that exist in real human sketches. In fact we already compare with [52] which employed such sketch specific augmentation strategies, and it is found to be much inferior to our semi-supervised framework (see Table. 2).

(iii) Optimising the final layer (using Eq. 9 in our case) is a very common practice during fine-tuning with RL, and is heavily adopted by the image-captioning literature [14], and very recently by on-the-fly FG-SBIR [5].

(iv) Edge-map hardly resembles the highly abstracted and subjective nature of amateur human sketches. For example, sketches do not follow the perfect edge boundary unlike edge-maps, thus model trained on pseudo-sketches via edge2sketch [33] falls short to generalise to real human sketches.

(v) Acc@1 without using RL scheme for Ours-F-Pix2Pix is 34.14%.

(vi) In future, our photo-to-sketch generation model could further be evaluated on Sketchy [38], however, it seems to be comparatively difficult than that of QMUL-ShoeV2 due to more noisy background.