Exploiting Unlabelled Photos for Stronger Fine-Grained SBIR

Aneeshan Sain\textsuperscript{1,2} Ayan Kumar Bhunia\textsuperscript{1} Subhadeep Koley\textsuperscript{1,2} Pinaki Nath Chowdhury\textsuperscript{1,2} Soumitri Chattopadhyay\textsuperscript{*} Tao Xiang\textsuperscript{1,2} Yi-Zhe Song\textsuperscript{1,2}
\textsuperscript{1}SketchX, CVSSP, University of Surrey, United Kingdom.
\textsuperscript{2}iFlyTek-Surrey Joint Research Centre on Artificial Intelligence.
\{a.sain, a.bhunia, p.chowdhury, t.xiang, y.song\}@surrey.ac.uk

Abstract

This paper advances the fine-grained sketch-based image retrieval (FG-SBIR) literature by putting forward a strong baseline that overshoots prior state-of-the-arts by \( \approx 11\% \). This is not via complicated design though, but by addressing two critical issues facing the community (i) the gold standard triplet loss does not enforce holistic latent space geometry, and (ii) there are never enough sketches to train a high accuracy model. For the former, we propose a simple modification to the standard triplet loss, that explicitly enforces separation amongst photos/sketch instances. For the latter, we put forward a novel knowledge distillation module can leverage photo data for model training. Both modules are then plugged into a novel plug-n-playable training paradigm that allows for more stable training. More specifically, for (i) we employ an intra-modal triplet loss amongst sketches to bring sketches of the same instance closer from others, and one more amongst photos to push away different photo instances while bringing closer a structurally augmented version of the same photo (offering a gain of \( \approx 4-6\% \)). To tackle (ii), we first pre-train a teacher on the large set of unlabelled photos over the aforementioned intra-modal photo triplet loss. Then we distill the contextual similarity present amongst the instances in the teacher’s embedding space to that in the student’s embedding space, by matching the distribution over inter-feature distances of respective samples in both embedding spaces (delivering a further gain of \( \approx 4-5\% \)). Apart from outperforming prior arts significantly, our model also yields satisfactory results on generalising to new classes. Project page: \url{https://aneeshan95.github.io/Sketch}

Figure 1. (left) A strong FG-SBIR baseline with strong PVT \cite{74} backbone trained on intra-modal triplet loss, stabilised by EMA. (right) It additionally leverages unlabelled photos to enrich its latent space by distilling instance-wise discriminative knowledge of a teacher pre-trained on unlabelled photos.

is on fully utilising the faithful nature of sketches to conduct instance-level retrieval \cite{70,81}.

Despite great strides made in the field, without exception, all existing FG-SBIR models \cite{60,82} work around a cross-modal triplet objective \cite{81} to learn a discriminative embedding to conduct retrieval. The general intuition has always been to make a sketch sit closer to its paired photo while pushing away non-matching ones (Fig. 1). However, a conventional triplet setup does not enforce sufficient separation amongst different photos or sketch instances – largely because the conventional objective fails to retain the holistic latent space geometry, being overly restrictive on learning within-triplet feature separation.

For that and as our first contribution, we utilise an intra-modal triplet objective in both modalities, in addition to the regular cross-modal \cite{81} triplet objective. For sketch, a query-sketch (anchor), is brought closer to another sketch of the same target-photo (positive) while distancing it from a sketch of any non-matching photo (negative). For the photo modality, as there is only one uniquely paired photo (anchor), we treat a morphologically augmented version of the photo as positive, and a non-matching one as negative. Importantly, we show that the best morphological operations are those that operate on visual attributes bearing relevance with information exclusive to sketches (e.g., shape) Fig. 1.

1. Introduction

Sketch \cite{7,17,39} has long established itself as a worthy query modality that is complementary to text \cite{16,23,26}. Conceived as a category-level retrieval task \cite{10,18}, sketch-based image retrieval (SBIR) has recently taken a turn to a “fine-grained” setting (i.e., FG-SBIR), where the emphasis

*Interned with SketchX
community is that of sketch data scarcity – there are just not enough sketches for the extra performance gain [5, 6]. Instead of going for the more explicit route of photo to pseudo sketch synthesis [4], we vote for the simple and more implicit route of leveraging unlabelled photos. For that, we adapt a knowledge distillation setup – we first train a model on unlabelled photos only via an intra-modal triplet loss, then distill its instance-wise discriminative knowledge to a FG-SBIR model. This apparently simple aim is however non-trivial to train owing to our cross-modal setup as naively using standard knowledge distillation paradigms involving logit distillation [33] from teacher’s embedding space to the student’s would be infeasible. For a cross-modal problem like ours, we need to preserve the instance-wise separation amongst photos, as well as corresponding sketches. For that, we propose to preserve the contextual similarity between instances and their nearest neighbours, modelled as a distribution over pairwise distances, from the pre-trained teacher (photo model), to that of the student (FG-SBIR). Inspired from recent literature, we introduce a novel distillation token into our PVT-backed student, that is dedicated towards distilling contextual similarity.

However fitting this token into the existing PVT-architecture is non-trivial. Unlike other vision transformers like ViT [25], PVT employs a reshape operation [74], to reshape the resultant individual patch tokens at one level back to the feature map, for input to the next level. It follows that adding a token here naively would break this operation. We thus use our distillation token only during input to the transformer layer at each level [74], and set the modified token aside before the reshaping operation. A residual connection [70] thereafter connects this modified token as input to the transformer layer at the next level. Engaging the distillation token at every level like ours, helps it imbibe the inductive bias modelled by the pyramidal structure of PVT, thus facilitating better distillation.

Finally, on the back of a pilot study (Sec. 3), we uncover a widespread problem with standard triplet loss training – that training is highly unstable, reflected in the highly oscillating evaluation accuracies noted at every 100th training iteration. For that, we take inspiration from literature on stabilising GAN training [78] on Exponential Moving Average [78] – a strategy that employs a moving average on model parameters iteratively, with a higher priority (mathematically exponential) to recent iterations over earlier ones.

To sum up: (i) We propose a strong baseline for FG-SBIR, that overshoots prior arts by ≈10% (ii) We achieve this by putting forward two simple designs each tackling a key problem facing the community: inadequate latent space separation, and sketch data scarcity. (iii) We introduce a simple modification to the standard triplet loss to explicitly enforces separation amongst photos/sketch instances. (iv) We devise a knowledge distillation token in PVT [74] that facilitates better knowledge distillation in training from unlabelled data. Finally, apart from surpassing prior arts significantly, our model also shows encouraging results on generalising to new classes, without any paired sketches.

2. Related Works

Fine-grained SBIR (FG-SBIR): FG-SBIR aims at retrieving one particular photo from a gallery of specific-category corresponding to a query-sketch. Initially proposed as a deep triplet-ranking based siamese network [81], FG-SBIR has progressively improved via attention-based modules with a higher order retrieval loss [70], textual tags [16, 68], self-supervised pre-training [50], hierarchical co-attention [61], cross-category generalisation [49], and reinforcement learning [9]. Overall, they aim to learn a joint embedding space so as to reduce the cross-modal gap [29, 61], typically using a triplet ranking loss [81] that aims to bring a sketch closer to its paired photo while distancing it from others. Some have optimised target-photo rank [9], fused classification loss with ranking [33], used multi-task attribute loss [69], or other loss modules in self-supervised pretext tasks [5, 50] for efficiency. Others addressed sketch-specific traits like style-diversity [62], data-scarcity [4] and redundancy of sketch-strokes [6] in favor of better retrieval. Arguing against such complex frameworks, we aim for a simpler yet stronger FG-SBIR pipeline, that can generalise well across categories even in low-data regime.

Transformers in Computer Vision: Transformers [73] are end-to-end neural networks that leverage self-attention mechanism [14] for modelling sequential data. Vision transformers [24, 43, 74] win over traditional CNNs [66] in their ability to model long-range dependencies in sequential data (here, visual patches) thus learning a stronger feature representation than CNNs [24]. Since the seminal work of ViT [24] that introduced image patches as input to transformer layers, it has been improved further via convolution components [76], attention-guided knowledge distillation for data-efficiency [72] and a feature pyramid [74, 75] modeling inductive bias. Common applications include object detection [12], image super-resolution [77], image synthesis [36], etc. Despite their recent usage in fine-grained image recognition [35, 42], only few have employed transformers in encoding sketches [22, 63] Such transformers however have not been used for FG-SBIR. In this paper, we adapt transformers for the first time in FG-SBIR to propose a strong baseline that would outperform existing state-of-the-arts and we hope research would progress further considering our baseline as a standard baseline.

Training from Unlabelled Data: Mainly two frontiers of research have emerged in taking advantage of unlabelled data – self-supervised and semi-supervised learning. The former spans over a large literature [37], including generative models [27] like VAEs [38], contrastive learning ap-
proaches [15, 30], clustering [13], etc. Furthermore, several pre-text self-supervised tasks have been explored like image colorization [85], in-painting [54], etc. Recent applications in sketch include a pre-text task of solving jigsaw puzzles [50], and learning from the dual representative nature of sketches [4, 5]. Semi-supervised learning aims at exploiting large unlabeled data together with sparsely labeled data to improve model performance. Common approaches include entropy minimization [28], pseudo-labeling [67], or consistency regularization [47]. While pseudo-labeling employs confident prediction [67] from trained classifiers to create artificial labels [40] for unlabeled data, consistency regularization learns a classifier by promoting consistency in predictions between different views of unlabeled data, either via soft [48] or hard [67] pseudo-labels. Recently, data scarcity in SBIR was handled, by generating more sketches for unlabeled photos [4] in a semi-supervised setup.

We focus on one of its related paradigm of knowledge distillation [2] that aims at transferring knowledge of a pre-trained teacher network to a student. While some leverage output logits [2] others focus on hidden layers [59] or attention-maps [84] of pre-trained teachers for the same. Improvising further, self-distillation [3] employed the same network for both student and teacher models, whereas a multi-exit strategy [55] optimised compute via multiple exits at different depths for adaptive inference. Common applications include object detection [21], semantic segmentation [32], depth-estimation [56], etc. Contrarily, FG-SBIR demands: (a) instance level discrimination (b) cross-modal one-to-one correspondence. As collecting sketch-photo labelled pairs is costly, we train a teacher from abundant unlabeled photos. The instance-discrimination thus learned in photo space, is distilled to a student FG-SBIR model to make it more discriminative.

3. Pilot Study: Delving deeper into FG-SBIR

Training stability: Being an instance-level cross-modal retrieval problem, training an FG-SBIR [61] model is often found to be unstable, as evaluation accuracy oscillates significantly during training. Hence, it is evaluated frequently to capture the best accuracy. Being the first to consider this, we first analyse by plotting (Fig. 2) test-set evaluation at every 100 training-steps of an existing baseline FG-SBIR model with VGG-16 backbone, trained over standard triplet loss [81]. Towards reducing such instability, we design a stronger baseline model, which we hope will be a standard baseline for future research.

Variable dataset size: Existing literature shows FG-SBIR model performance to suffer from scarcity of training sketch-data [4]. This leads us to wonder if unlabeled photos can be leveraged to train an FG-SBIR model to reduce the data annotation bottleneck on sketches for training. Consequently, we conduct a study to analyse how an FG-SBIR model performs when training data-size is varied. Accordingly we compute and plot the performances of existing and our stronger baselines (Sec. 4) on varying dataset-size in Fig. 2. As both baselines, perform poorly on decreasing labelled training data, we explore if unlabelled photos can boost performance in such a scenario. We thus design a distillation strategy from unlabelled photos, following which, our method at ∼50% training data, matches the accuracy (32.03%) of an existing baseline at 100% training data (red square in Fig. 2), thereby justifying our paradigm.

Classes without any sketch-photo pairs: Besides training data scarcity, generalisation to unseen classes for multiclass category FG-SBIR [8], is also one of the challenging issues in FG-SBIR. Considering a realistic scenario, sketch-photo pairs might not always be available for all classes, however photos pertaining to such classes can be curated fairly easily. Therefore, given a model trained on classes having sketch-photo pairs, we aim to discover how well can it perform on classes lacking paired sketches. We thus conduct a small pilot study by taking the first 25 classes (alphabetically) from Sketchy [64] – 15 seen training (sketch-photo pairs available) and 10 testing (only photos available), in an FG-SBIR setting. Fig. 2 shows baseline model to perform significantly better on classes with access to sketch-photo pairs but fails poorly on classes lacking paired sketches. The main challenge therefore is to leverage the knowledge of classes having unlabelled photos, to preserve the model’s accuracy. In doing so, our method maintains a stable performance via its knowledge distillation paradigm.

4. A Stronger FG-SBIR baseline

Overview: Unlike CNN-based existing FG-SBIR baselines [70, 81] that learn a strong sketch-embedding function over a cross-modal triplet loss, we enhance the paradigm with three distinct modifications: (i) Employ vision transformer,
particularly PVT [74] as our backbone after thorough analysis (Sec. 6), as unlike CNNs having local receptive fields, vision transformers ensure a global receptive field, modelling better feature representations. (ii) Besides cross-modal triplet loss [81], we formulate an intra-modal triplet loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning. (iii) Towards increasing training stability we follow GAN-literature where loss that helps in discriminative learning.

5. Training with Unlabelled Photos via KD

Once we have a stronger baseline, we aim to harness the learning potential of unlabeled photos, towards training a
strong FG-SBIR model [4]. We thus employ a knowledge distillation paradigm [33], that transfers the photo instance discrimination potential from a pre-trained ‘teacher’ (Ω) trained on unlabelled photos to our ‘student’ (ϕ) FG-SBIR model for cross-modal retrieval. Given a limited amount of labelled sketch-photo pairs \( \mathcal{D}_L = \{ (p^l_i, s^l_i) \}_{i=1}^{N_L} \), for our student FG-SBIR model, and a larger set of unlabelled photos \( \mathcal{D}_U = \{ p^u_i \}_{i=1}^{N_U} \), for pre-training our teacher (\( N_U \gg N_L \)), our goal is to improve retrieval accuracy of our student FG-SBIR model, using both \( \mathcal{D}_L \) and \( \mathcal{D}_U \) (no paired sketches).

However applying this KD paradigm is non-trivial to our setting. While conventional KD usually transfers knowledge via logit-distillation [33] for classification problem, ours is a cross-modal retrieval setup where the output is a continuous d-dimensional feature in a joint-embedding space. Moreover, if one naively regresses between features from a teacher and student for sketch and photo branches, it might suffer from incompatibility as the two embedding spaces are different. Also, if teacher [2] and student embedding dimensions are different, an additional feature transformation layer [59] is needed to match them. While distilling through pair-wise distance [4,65] might be an option, it fails to transfer the structural knowledge of the teacher’s entire latent space to the student, as it focuses on only one pair at a time. We thus aim to design a distillation framework that considers the contextual similarity amongst samples in the teacher’s embedding space and preserves its structural knowledge while distilling to the student.

**Modified PVT for Distillation:** Concretely, our network involves two models - a teacher \( \Omega(\cdot) \) pre-trained large-scale photo, and a learnable FG-SBIR student \( \phi(\cdot) \), both of which uses PVT [74] as backbone feature extractor. While the teacher’s backbone PVT remains unchanged, we follow recent transformer literature [72] to introduce a learnable distillation token \( \Delta \in \mathbb{R}^d \), that allows our model to learn from the output of the teacher, while remaining complementary to the feature extracted from PVT [74]. However, naively concatenating a token is infeasible, as unlike other vision transformers [72], PVT involves reshaping (for down-scaling) a fixed number of individual patch tokens to the subsequent feature map, which would be imbalanced on adding one extra token. Therefore, we focus on the second step of PVT-block [74] involving transformer layer (Sec. 4), which can accommodate a variable number of tokens. Accordingly, at every level, \( \Delta \) is fed to the transformer layer concatenated with rest of the N image patch tokens \( \{ p^u_i \}_{i=1}^{N+1} \in \mathbb{R}^{(N+1) \times d} \), to obtain the resultant set of tokens as \( \{ p^u_i \}_{i=1}^{N+1} \in \mathbb{R}^{(N+1) \times d} \). Before reshaping, \( \Delta \) is excluded to prevent dimensional mismatch, and fed again similarly to the transformer layer at the next level via a residual connection (Fig. 3). Being processed at every level, \( \Delta \) not only accommodates the output knowledge of the teacher network but also, imibes the inductive bias [74] contributed by the pyramidal structure of PVT. From the final layer, the student outputs two features: a discriminative \( f \in \mathbb{R}^d \), and a distillation feature \( \mu \in \mathbb{R}^d \).

**Pre-training Teacher:** The teacher \( \Omega(\cdot) \), is trained on photos of both labelled and unlabelled sets \( \mathcal{G} = \mathcal{D}_U \cup \mathcal{D}_L^p \), over an intra-modal triplet loss following Eqn. 2.

**Training FG-SBIR Student:** The student’s discriminative feature is used to train over a combination of cross-modal and intra-modal triplet losses on labelled data \( \mathcal{D}_L \) following Eqn. 4. Intra-modal triplet loss over unlabelled photos is leveraged to harness the potential of unlabelled photos \( \mathcal{D}_U \) with a weighting hyperparameter \( \lambda_3 \) as,

\[
\mathcal{L}_\text{Disc} = \mathcal{L}_\text{Trn} + \lambda_3 \mathcal{L}_\text{Tri}^U \tag{5}
\]

In order to transfer [59] the instance-discriminative knowledge of photo domain from the pre-trained teacher \( \Omega(\cdot) \) to improve the cross-modal instance level retrieval of student FG-SBIR model \( \phi(\cdot) \), we leverage both unlabelled photos and sketch/photo pairs from labelled data during knowledge distillation. First, we pre-compute the features of all the unlabelled photos \( \mathcal{D}_U \) using the frozen teacher model as \( \mathcal{F}_U^\Omega = \{ f^\Omega_i \}_{i=1}^{N_U} \) where \( f^\Omega_i = \Omega(p^u_i) \in \mathbb{R}^d \).

Now a very naive way of distillation [33] would be to pass \( p^u_i \) through student model \( \phi(\cdot) \) to obtain the distillation feature \( \mu^\phi_i = \phi(p^u_i) \), and directly regress it against \( f^\Omega_i \) considering it as ground-truth. Alternatively, one may regress the feature-wise distance between two photos \( (p^u_i, p^u_j) \) in a similar fashion from teacher to student model [33]. However, we focus on preserving the structure of the embedding space of the teacher while distillation to the student. We thus calculate the pairwise distance of \( f^\Omega_i \) from its K nearest neighbours \( \{ f^\Omega_{k1}, \cdots, f^\Omega_{kK} \} \), as \( \mathcal{D}^\Omega_i = \{ \delta(f^\Omega_i, f^\Omega_j) \}_{j=1}^{K} \). Equivalently, we pass \( p^u_i \) and its K-nearest neighbours \( \{ p^u_{k1}, \cdots, p^u_{kK} \} \) via the student model to obtain corresponding distillation features \( \mu^\phi_i \) and \( \{ \mu^\phi_{k1}, \cdots, \mu^\phi_{kK} \} \) respectively, thus calculating \( \mathcal{D}^\phi_i = \{ \delta(\mu^\phi_i, \mu^\phi_{jk}) \}_{j=1}^{K} \). Similarly, one may calculate a regression loss between \( \mathcal{D}^\Omega_i \) and \( \mathcal{D}^\phi_i \), for better stability we model them as probability distribution of pairwise-similarity amongst K-nearest neighbours in the teacher’s embedding space. As pairwise similarity is negative of pairwise distances, we calculate the temperature (\( \tau \)) normalised softmax (S) [61] probability as \( S_r(\Delta \mathcal{D}^\phi_i) \) where,

\[
S_r(\Delta \mathcal{D}^\phi_i)_{rj} = \frac{\exp(-\delta(f^\Omega_i, f^\Omega_j)/\tau)}{\sum_{k=1}^{K} \exp(-\delta(f^\Omega_i, f^\Omega_{jk})/\tau)} \tag{6}
\]

Similarly obtained \( S_r(\Delta \mathcal{D}^\phi_i) \), and \( S_r(\Delta \mathcal{D}^\Omega_i) \) represent the structural knowledge of embedding spaces of student \( \phi(\cdot) \) and teacher \( \Omega(\cdot) \) respectively. The consistency constraint can therefore be defined as the Kullback-Leibler (KL) divergence [46] between them as,

\[
\mathcal{L}_{\text{KL}} = KL(S_r(\Delta \mathcal{D}^\Omega_i) \parallel S_r(\Delta \mathcal{D}^\phi_i)) \tag{7}
\]
On the other hand, labelled dataset comes with sketch and photo, where, using the photos we can calculate $L_{\text{Dist}}^{\phi}$ in a similar fashion. Given the cross-modal nature of student’s embedding space [81], we also need to align the sketches with their paired photos while preserving the existing contextual similarity of those photos with their neighbours. Being trained on photos alone, extracting sketch features via the teacher $\Omega(\cdot)$ to extract sketch-features, would be a flawed design choice. Considering a sketch-photo pair $(s_i, p_j)$, we first obtain $D_{p_i}^{\phi}$ using $p_j$. However this time we use $s_i$ to calculate $D_{s_i}^{\phi} = \{\delta(p_j^\phi, \mu_{p_i}^\phi)\}_{K=1}^K$. Now, calculating $L_{KL}^T$ between these two resultant probabilities implies maintaining the same contextual similarity for sketches as that for their paired photos, as guided by the teacher model. With $\lambda_{1,5}$ as respective weighting hyperparameters, the total distillation loss becomes,

$$L_{\text{Disc}}^{\phi} = L_{KL}^{\phi} + \lambda_4 L_{KL}^{DL} + \lambda_5 L_{KL}^{PU} \quad (8)$$

Summing up, our student model is trained from a weighted combination of two losses as:

$$L_{\text{Un}}^{\phi} = L_{\text{Disc}}^{\phi} + \lambda_6 L_{\text{Dist}}^{\phi} \quad (9)$$

6. Experiments

Datasets: We use two publicly available datasets, QMUL-Chair-V2 and QMUL-Shoe-V2 [61,81]. They contain 2000 (400) and 6730 (2000) sketches (photos) respectively with fine-grained sketch-photo associations. We keep 1275 (300) and 6051 (1800) sketches (photos) from QMUL-Chair-V2 and QMUL-Shoe-V2 respectively for training while the rest is used for testing. We also use Sketchy [64] which contains 125 categories with 100 photos each, having at least 5 sketches per photo with fine-grained associations. While, training uses a standard (90:10) train-test split [64], during inference we construct a challenging gallery using photos across one category for retrieval. Besides such labelled training data, we use all 50,025 photos of UT-Zap50K [80] and 7,800 photos [50] collected from shopping websites, including IKEA, Amazon and Taobao, as unlabelled photos for shoe and chair retrieval, respectively. For Sketchy we use its extended version with 60,502 additional photos [41] introduced later for training from unlabelled data.

Implementation Details: ImageNet [20] pre-trained PVT-Large [74] model extracts features from $224 \times 224$ resized images, keeping patch-size ($p_i$) of 4 at each level and 1, 2, 5, and 8 spatial-reduction-attention heads [74] in 4 successive levels, with the final feature ($\mu$ and $f$) having size 512. Implemented via PyTorch [53], our model is trained using Adam-W optimiser [45] with momentum of 0.9 and weight decay of 5e-2, batch size of 16, for 200 epochs, on a 11 GB Nvidia RTX 2080-Ti GPU. Initial learning rate is set to 1e-3 and decreased as per cosine scheduling [44]. Determined empirically, $m_{\text{CM}}, m_{\text{LM}}, m_{\text{PM}}$ and $\tau$ are set to 0.5, 0.2, 0.3 and 0.01, while $\lambda_{1-6}$ to 0.8, 0.2, 0.4, 0.4, 0.7 and 0.5 respectively. Following [61] we use Acc.@q, i.e. percentage of sketches having true matched photo in the top-q list.

6.1. Competitors


<table>
<thead>
<tr>
<th>Methods</th>
<th>Chair-V2 (%)</th>
<th>Shoe-V2 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1</td>
<td>Top-10</td>
</tr>
<tr>
<td>SOTA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Triplet-SN</td>
<td>53.48</td>
<td>89.91</td>
</tr>
<tr>
<td>HOLEF-SN</td>
<td>55.23</td>
<td>88.61</td>
</tr>
<tr>
<td>Jigsaw-SN</td>
<td>58.51</td>
<td>88.78</td>
</tr>
<tr>
<td>OnTheFly</td>
<td>59.18</td>
<td>89.35</td>
</tr>
<tr>
<td>StyleMeUp</td>
<td>65.85</td>
<td>90.84</td>
</tr>
<tr>
<td>Semi-sup-SN</td>
<td>66.86</td>
<td>91.12</td>
</tr>
<tr>
<td>Ours-Strong</td>
<td>71.22</td>
<td>92.18</td>
</tr>
<tr>
<td>B-ResNet-18</td>
<td>48.42</td>
<td>85.62</td>
</tr>
<tr>
<td>B-ResNet-50</td>
<td>47.78</td>
<td>82.34</td>
</tr>
<tr>
<td>B-InceptionV3</td>
<td>55.41</td>
<td>88.21</td>
</tr>
<tr>
<td>B-VGG-16</td>
<td>58.23</td>
<td>88.78</td>
</tr>
<tr>
<td>B-VGG-19</td>
<td>61.46</td>
<td>89.16</td>
</tr>
<tr>
<td>B-ViT</td>
<td>38.71</td>
<td>72.65</td>
</tr>
<tr>
<td>B-DeiT</td>
<td>56.25</td>
<td>87.72</td>
</tr>
<tr>
<td>B-SWIN</td>
<td>66.34</td>
<td>91.03</td>
</tr>
<tr>
<td>B-CVt</td>
<td>68.42</td>
<td>92.11</td>
</tr>
<tr>
<td>B-CoAtNet</td>
<td>69.68</td>
<td>91.78</td>
</tr>
<tr>
<td>Ours-Strong</td>
<td>71.22</td>
<td>92.18</td>
</tr>
<tr>
<td>B-Edge-Pretain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B-Edge2Sketch</td>
<td>71.58</td>
<td>90.87</td>
</tr>
<tr>
<td>B-Regress</td>
<td>72.65</td>
<td>91.32</td>
</tr>
<tr>
<td>B-RKD</td>
<td>73.02</td>
<td>91.78</td>
</tr>
<tr>
<td>B-PKT</td>
<td>72.16</td>
<td>91.16</td>
</tr>
<tr>
<td>Ours-Full</td>
<td>74.68</td>
<td>92.79</td>
</tr>
</tbody>
</table>

(ii) SOTA-Augmentation (SOTA++): To judge how generic our paradigm is compared to existing frameworks, we augment the mentioned state-of-the-arts by introducing our intra-modal triplet objective (Sec. 4) with Exponential Moving Average in their respective training paradigms.

(iii) Architectural Variants: Using popular CNN architectures like InceptionV3 [71], ResNet-18,50 [31] and VGG-16,19 [66] as backbone feature extractors, we explore their potential in FG-SBIR against our strong baseline. Similarly, we also explore a few existing vision transformers as backbone feature extractors for our training paradigm, namely,
B-ViT [24] (ViT-B16 variant), B-DeiT [72] (DeiT-B variant), B-SWIN [43] (SWIN-B variant), B-CvT [76] (CvT21 variant), and B-CoAtNet [19] (CoAtNet3 with 384-dim).

(iv) On training from unlabelled data: Following [4] we curate a few more baselines that could be used to leverage unlabelled photos for training. Keeping rest of the paradigm same, B-Edge-Pretrain [57] naively uses edge-maps of unlabelled photos to pre-train the retrieval model. Although edge-maps bear little resemblance to sketches, we similarly design a baseline B-Edge2Sketch that follows [58] in converting edge-maps of photos to pseudo-sketches as labels to harness potential of unlabelled photos. As there are no previous works employing KD for FG-SBIR, we curate a few baselines offering alternatives on the paradigm of KD. These methods although repurposed for a cross-modal retrieval setting, use our network architecture (\(\Omega, \phi\)) during training. B-Regress - directly regresses between features computed by teacher and student for both sketches and images, after matching corresponding feature dimensions via additional feature transformation layer, over a \(l_2\) regression loss. B-RKD follows [51] to distill the knowledge of relative pairwise feature-distances of unlabelled photos from teacher’s embedding space to that of the student, over a distance-wise distillation loss. Following [52] off-the-shelf, for unlabelled photos, B-PKT computes the conditional probability density of any two points in teacher’s embedding space [52], which models the probability of any two samples being close together. Taking ‘N’ such samples, it obtains a probability distribution over pairwise interactions in that space. Obtaining a similar distribution over the same samples in student’s embedding space, it minimises their divergence over a KL-divergence loss.

6.2. Performance Analysis

FG-SBIR pipelines: Tables 1 and 2 compare our methods against state-of-the-arts and curated baselines. Triplet-SN [81] and HOLEF-SN [70] score low due to weaker backbones of Sketch-A-Net [83]. Jigsaw-SN [49] on the other hand improves performance, owing to self-supervised Mixed-modal jigsaw solving strategy learning structural information better. Enhanced by its RL-optimised reward function OnTheFly [9] surpasses them but fails to exceed StyleMeUp [62] (2.37\%↑ top1 on ShoeV2), thanks to its complex meta-learned disentanglement module addressing style-diversity. Unlike others that are restricted to paired training data, Semi-sup-SN [4] harnesses knowledge of unlabelled data with its additionally generated pseudo-sketch labels in achieving comparatively higher performance. However, being dependent on the usually unreliable quality of generated sketch and the instability of reinforcement learning involved, it lags behind our relatively simpler yet robust knowledge distillation paradigm, boosted with our better transformer-based feature extractor. Importantly, when augmenting our ‘strong-baseline’ paradigm to existing SOTAs, we observe a relative rise in top-1 performance of all methods by \(\approx 4 - 6\%\) overall in SOTA++ section (Table 1). Despite costing a minimal memory overhead (training only), these objectives reward a considerably high accuracy boost which verifies our method to be an easy-fit and quite generic to serve as a strong FG-SBIR baseline.

Table 2. Quantitative comparison of pipelines on Sketch [64].

<table>
<thead>
<tr>
<th>Methods</th>
<th>Sketchy (%)</th>
<th>Top-1</th>
<th>Top-5</th>
<th>Sketchy (%)</th>
<th>Top-1</th>
<th>Top-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triplet-SN-ours</td>
<td>19.48</td>
<td>37.91</td>
<td>B-Edge-Pretrain 34.98 61.32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOLEF-SN-ours</td>
<td>20.23</td>
<td>38.61</td>
<td>B-Edge2Sketch 35.81 61.74</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jigsaw-SN-ours</td>
<td>21.45</td>
<td>39.56</td>
<td>B-Regress 36.35 62.31</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OnTheFly-ours</td>
<td>07.58</td>
<td>12.14</td>
<td>B-RKD 37.02 63.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>StyleMeUp-ours</td>
<td>22.95</td>
<td>45.84</td>
<td>B-PKT 38.62 63.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours-Strong</td>
<td>34.72</td>
<td>65.10</td>
<td>Ours-Full 38.54 71.52</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Backbone architectures: Comparing efficiency of CNNs as backbone feature extractors (B-CNNs) for FG-SBIR, we find B-VGG19 to perform best, being slightly better than B-VGG16 (by 3.23\%) at an extra memory cost (21mb), and much better than B-ResNet18 (by 10.67\%) mainly due to the latter’s thinner conv-layers in Top-1 accuracy on ShoeV2. Among transformers, B-CoAtNet, B-CvT and B-SWIN perform much better than B-ViT or B-DeiT thanks to their hierarchica design and mechanisms like shifting window or convolutional tokens capturing fine-grained sparse details of a sketch via local patch-level interaction, thus representing sketches better unlike other two. Surpassing all others, with its unique pyramidal structure imbibing inductive bias, PVT [74] (Ours-Strong,Full) fits optimally to our method.

Training from Unlabelled Data: While B-Edge-Pretrain offers little gain over our stronger baseline (Ours-Strong), augmenting edge-maps with B-Edge2Sketch by a selected subset of strokes, to imibe the abstractness of sketch, increases accuracy reasonably. Lower scores of B-Regress is largely due to its metric-based regression across two different embedding spaces causing misalignment, whereas attending to one pair at a time, B-RKD [51] fails to preserve the structural knowledge of the embedding space, thus scoring below B-PKT (by 0.48 top1 on ShoeV2). Although B-PKT [52] is slightly similar to our paradigm of preserving the latent space, lacking cross-modal discrimination objective [81] and therefore being trained on unlabelled data alone, performs lower than our method which additionally leverages the potential of labelled sketch-photo pairs.

6.3. Ablation Study

Importance of loss objectives: To justify each loss in our network of a stronger baseline, we evaluate them in a stripdown fashion (Table 3). Performance against cross-modal objective alone (Type-I), increase on aid from intra-modal objectives (Type-II), as the latter encourages distancing of
multiple photos visually close to the sketch, based on finer intra-modal relative discrimination. Adding EMA (Ours) optimisation [78] ensures smoother convergence of above objectives with a slight increase (by 2.96%) in accuracy. In similar spirit we perform experiments on losses of distillation paradigm (Eqn. 8). Performance using unlabelled data only (Type-III) rises with added knowledge from labelled photos (Type-IV). However, without aligning sketches using $L_{KL}^{EM}$ it lags behind Ours-Full by 1.14%.

### Further Analysis

(i) Although ours is a larger model ($≈20-25\%$ gain while taking similar inference time (0.37ms/ShipV2).
(ii) Our method takes 44.1 ms per training step compared to 56.83 ms of SOTA StyleMeUp [62], proving itself as a strong and efficient method.
(iii) Although our method takes more hyperparameters than the simpler Triplet-SN [81] SOTA, they are quick to tune, and justifies itself with a boost of ≈25% Acc @1 on Chair-V2 [81].

### Augmentation Strategy

We design a few experiments to explore other photo augmentation techniques [1] for $L_{Tri}^{IM}$ (Eqn. 2) like, colour distortion, partial blurring and random shift in sharpness. Outperforming their respective scores of 46.14%, 47.32% and 47.54% (Acc@1-ShoeV2) our methods confirms our intuition that morphological augmentations would better direct the loss objective, as it instills discrimination based on information exclusive to a sketch.

### Influence of Distillation Token ($\Delta$):

To justify its importance we explore two more designs. (A) Without any explicit distillation token we perform knowledge distillation using the same feature used for discriminative learning. (B) Following [72] we append a learnable token to the input of the last stage, which after processing acts as the distillation feature. Although case-B (45.31%) surpasses case-A (44.71%) by 1.6% in Acc@1 on QMUL-ShoeV2, confirming the need of a dedicated distillation token, it lags behind by 3.04% to ours. We argue that our design of engaging the distillation token at every level via residual connection instills the inductive bias modelled by the pyramidal downsampling, into the distillation token via transformer layers, thus creating a better representation.

### Further Analysis

(i) Although ours is a larger model than the earlier SOTA of Triplet-SN [81], it offers ≈20-25% gain while taking similar inference time (0.37ms/ShipV2).
(ii) Our method takes 44.1 ms per training step compared to 56.83 ms of SOTA StyleMeUp [62], proving itself as a strong and efficient method.
(iii) Although our method takes more hyperparameters than the simpler Triplet-SN [81] SOTA, they are quick to tune, and justifies itself with a boost of ≈25% Acc@1 on Chair-V2 [81].

### 6.4. Multi-category FG-SBIR via Unlabelled Photos

Our pilot study (Sec. 3) reveals FG-SBIR models to perform poorly on classes lacking sketch-photo pairs. This has been explored in a few recent FG-SBIR works. Apart from Jigsaw-SN [50] (Sec. 6.1), (CC-Gen) [49] takes a cross-category (CC) domain-generalisation approach, modelling a universal manifold of prototypical visual sketch traits that dynamically embeds sketch and photo, to generalise on unseen categories. Recently, [8] uses a dedicated meta-learning framework that adapts a trained model to new classes using a few corresponding sketch-photo pairs as support. However, both methods have access to sufficient [50] or few [8] sketch-photo pairs of novel classes, unlike our setup of no paired sketches. This goes beyond the dependency on paired sketches offering a realistic mid-ground between standard [81] and zero-shot [23] inference setup, where not all classes have paired sketches, but their photos can be collected easily. Consequently, following [79] we split Sketchy [64] as 21 unseen-test classes, and 104 training classes with a 70:30 training:validation split in the latter. Unlike existing methods, our student additionally trains via distillation from photos of 21 unseen classes for better retrieval accuracy. For retrieval a separate gallery is maintained per category, and average top-q accuracy across 21 classes is reported (Table 4). We compare with a few methods that leverage data of novel classes, besides labelled data for training: Jigsaw-SN (extending [50]) trains via an auxiliary jigsaw-solving task on unlabelled photos without any fine-tuning (our setup), Adaptive-SN [8] trains on a few sketch-photo pairs of new classes. B-EdgePretrain and B-Edge2Sketch trains on edge-maps and synthetic sketches via [58] from edge-maps of unlabelled photos respectively. CC-Gen [49] evaluates in zero-shot [23] setup with no data from novel classes. Consequently CC-Gen scores lower (Table 4), compared to Jigsaw-SN with its auxiliary jigsaw-solving task, whereas baselines using edge-maps of unlabelled images as pseudo-sketches score relatively better. Even without using real sketch-photo pairs of test-set for quick-adaptation like Adaptive-SN [8], we achieve a competitive retrieval accuracy (2.47% Acc@1).

### 7. Conclusion

In this paper we put forth a strong baseline for FG-SBIR with PVT [74]-backbone, and offer a novel paradigm that at its core aims at learning from unlabelled data in FG-SBIR by distilling knowledge from unlabelled photos. While our proposed intra-modal triplet loss increases feature separation in model’s latent space, an EMA paradigm stabilises its training. Importantly, we for the first time introduce a distillation token in PVT architecture that explicitly caters to knowledge distillation. Extensive experiments against existing frameworks and various baselines show our method to outperform them, thus proving its significance.
References


Fei Huang, Yong Cheng, Cheng Jin, Yuejie Zhang, and Tao Zhang. Deep multimodal embedding model for fine-grained sketch-based image retrieval. In ACM SIGIR, 2017. 2

Ruyi Ji, Longyin Wen, Libo Zhang, Dawei Du, Yanjun Wu, Chen Zhao, Xianglong Liu, and Feiyue Huang. Attention convolutional binary neural tree for fine-grained visual categorization. In CVPR, 2020. 2

Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. In NeurIPS, 2021. 2


Xinda Liu, Lili Wang, and Xiaoguang Han. Transformer with peak suppression and knowledge guidance for fine-grained image recognition. Neurocomputing, 2022. 2


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

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. 2, 7, 8

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. 2, 3, 6, 7, 8


Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In ECCV, 2018. 7


Mary Phuong and Christoph H Lampert. Distillation-based training for multi-exit architectures. In ICCV, 2019. 3


Umar Riaz Muhammad, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Learning deep sketch abstraction. In CVPR, 2018. 7, 8


Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. CLIP for All Things Zero-Shot Sketch-Based Image Retrieval, Fine-Grained or Not. In CVPR, 2023. 1

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, 5, 6

Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Stylemeup: Towards style-agnostic sketch-based image retrieval. In CVPR, 2021. 2, 6, 7, 8


Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. ACM TOG, 2016. 3, 6, 7, 8

Muhammad Risqi U Saputra, Pedro PB De Gusmao, Yasin Almalioglu, Andrew Markham, and Niki Trigoni. Distilling knowledge from a deep pose regressor network. In ICCV, 2019. 5


Jifei Song, Yi-Zhe Song, Tony Xiang, and Timothy M Hospedales. Fine-grained image retrieval: the text/sketch input dilemma. In BMVC, 2017. 2
[69] Jifei Song, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Xiang Ruan. Deep multi-task attribute-driven ranking for fine-grained sketch-based image retrieval. In BMVC, 2016. 2

[70] 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, 3, 6, 7

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


[74] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In ICCV, 2021. 1, 2, 4, 5, 6, 7, 8


[81] 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, 3, 4, 6, 7, 8

[82] Qian Yu, Jifei Song, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Fine-grained instance-level sketch-based image retrieval. IJCV, 2021. 1

[83] 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. 6, 7
