Generalising Fine-Grained Sketch-Based Image Retrieval

Kaiyue Pang[1,2*] Ke Li[1,3*] Yongxin Yang[1] Honggang Zhang[3]
Timothy M. Hospedales[1,4] Tao Xiang[1] Yi-Zhe Song[1]
[1]SketchX, CVSSP, University of Surrey [2]Queen Mary University of London
[3]Beijing University of Posts and Telecommunications [4]The University of Edinburgh
kaiyue.pang@qmul.ac.uk, {yongxin.yang, t.xiang, y.song}@surrey.ac.uk
{like1990, zhhg}@bupt.edu.cn, t.hospedales@ed.ac.uk

Abstract

Fine-grained sketch-based image retrieval (FG-SBIR) addresses matching specific photo instance using free-hand sketch as a query modality. Existing models aim to learn an embedding space in which sketch and photo can be directly compared. While successful, they require instance-level pairing within each coarse-grained category as annotated training data. Since the learned embedding space is domain-specific, these models do not generalise well across categories. This limits the practical applicability of FG-SBIR. In this paper, we identify cross-category generalisation for FG-SBIR as a domain generalisation problem, and propose the first solution. Our key contribution is a novel unsupervised learning approach to model a universal manifold of prototypical visual sketch traits. This manifold can then be used to paramaterise the learning of a sketch/photo representation. Model adaptation to novel categories then becomes automatic via embedding the novel sketch in the manifold and updating the representation and retrieval function accordingly. Experiments on the two largest FG-SBIR datasets, Sketchy and QMUL-Shoe-V2, demonstrate the efficacy of our approach in enabling cross-category generalisation of FG-SBIR.

1. Introduction

Fine-grained sketch-based image retrieval (FG-SBIR) aims to find a specific photo instance given a human free-hand sketch input. This has been actively studied in recent years due to its challenge as a vision problem, and commercial relevance [19, 36, 24, 20, 41]. The key challenge is the sketch/photo domain gap. Photos are perspective projections of visual objects represented as dense pixels, while sketches are subjectively and abstractly rendered iconic line-drawings.

Recent FG-SBIR methods [24, 36, 28, 22] address this issue by learning a deep network embedding of sketch and photo that makes them directly comparable. This embedding is often trained by a triplet ranking loss to ensure that the network embeds positive pairs nearby, and negative pairs farther apart. This line of work has made great progress, with state-of-the-art approaching human performance [22] on the Sketchy benchmark [24].

Nevertheless, existing work has thus far implicitly assumed that instance-level annotations of positive and negative pairs are available for every coarse category to be evaluated. This assumption limits the practical applicability of FG-SBIR. More specifically, as we shall show in this paper, in practice FG-SBIR generalises very poorly if training and testing categories are disjoint. This is of course unsatisfactory for potential users of FG-SBIR such as e-commerce, where it would be desirable to train a FG-SBIR system once on an initial set of product categories, and then have it deployed directly to newly added product categories – without needing to collect and annotate new data and retrain the FG-SBIR model. Compared to other category-level tasks such as object recognition in photo images, this annotation barrier is particularly high for FG-SBIR as instance-specific sketches are expensive and slow to collect.

To understand why the existing FG-SBIR models have limited cross-category generalisation ability, consider that the task of FG-SBIR as essentially binary classification – to differentiate corresponding and non-corresponding sketch-photo tuples. In this sense, a change of *category* is a domain-shift [8] from the perspective of the machine learning model trained to perform matching. For example, a model trained on fine-grained matching of car photos and sketches, would struggle to perform fine-grained matching of bicycle images, due to inexperience with handlebars and saddles. Exposed to such out-of-sample data, the triplet-trained sketch/photo embedding networks may no longer place matching images nearby and vice-versa. Having identified the challenge as one as domain-shift, this suggests two

[*] Equal Contribution

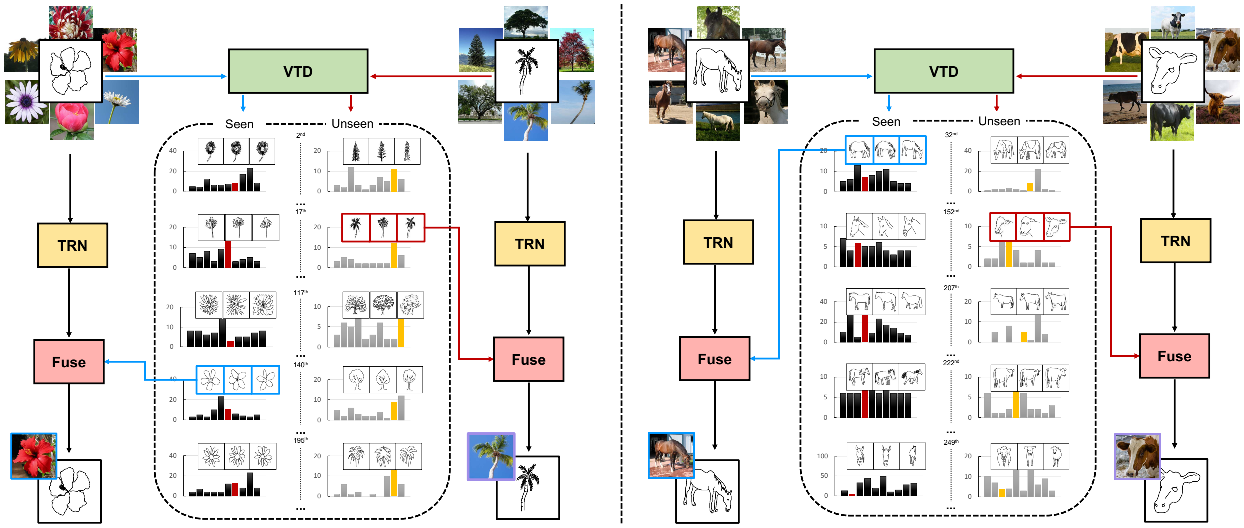


Figure 1: Illustration of our proposed method using four categories, organised into two related pairs. TRN: triplet ranking network. VTD: visual trait descriptor. In each bar-type VTD, we visualise its ten top distributed categories and highlight the specific one along with three belonged representative sketch exemplars. Each sketch is uniquely assigned to one VTD that describes a category-agnostic abstract sketch trait, which is in turn used to dynamically paramaterise the TRN so as to adapt it to the query sketch. See how both training and testing sketches thematically and coherently mapped to some shared VTDs. Best viewed in colour and zoom, more details in text.

categories of approaches to alleviating this issue: (1) Unsupervised domain *adaptation* approaches [8, 34] would use unlabelled target data to adapt the model to better suit the target data; and (2) domain *generalisation* approaches [26] aim to train a model that is robust enough to immediately generalise to the new domain's data off-the-shelf. In this paper we will address the harder domain generalisation setting – due to the practical value of not requiring target domain (category) data collection and model retraining.

To address the identified issue of cross-category FG-SBIR generalisation (CC-FG-SBIR), we propose a new framework that automatically adapts the deep feature extraction to a given query sketch. This ensures a good representation is produced at testing-time, even when dealing with out-of-sample data in the form of sketches and photos from novel categories. The key idea is to learn an auxiliary unsupervised embedding network that maps any given sketch to a universal dictionary of prototypical sketch traits or manifold embeddings. We call this universal because it is a representation that cuts across categories. This network can thus be used to provide a latent visual trait descriptor (VTD) of any sketch (from either a training or novel category). This descriptor is in turn used to paramaterise both photo and sketch feature extractors to adapt them to the current query sketch category. Fig.1 illustrates the unsupervised embedding learned by our auxiliary network via an illustrative five (of 300) learned embeddings (dictionary words). One can see how categories (such as flowers) span multiple embeddings and how individual embeddings group thematically similar sketches. For example descriptor 2 and 140 encompass "complicate-dense" and "simple-sparse" visual patterns for flowers and trees; while descriptor 207 and 249 model "leftwards full-body view" and "frontal face view" respectively for cows and horses. We can also see how both training (left subgroups) and disjoint testing sketch category (right subgroups) are assigned to the same descriptor according to common sketch traits.

The introduction of this auxiliary universal embedding network is inspired by the pioneering *Noise As Targets* (NAT) [3] model. NAT proposes to pre-generate the set of all embeddings randomly – as noise – and then learn a network to map the data to this fixed noise distribution. However NAT approximately solves a cumbersome and costly discrete assignment problem to match images with embeddings at each back-propagation iteration. In contrast, we propose a novel approach to learning an embedding network based on the Gumbel-Softmax [15] reparameterisation trick. As a result, the learning is faster and more stable; and more flexible in that several alternative objectives can be considered in the same formulation. Overall our framework can be considered as a solution to domain generalisation [26] that adapts a model via a domain-descriptor, but where the descriptor is estimated from a single data instance rather than assuming it is given as metadata [32, 33]; and where the perspective on descriptor definition is one of latent-domain discovery [31].

Our contributions are two-fold: (1) For the first time, the cross-category FG-SBIR generalisation (CC-FG-SBIR) problem is identified and tackled. (2) A solution is introduced based on a novel universal prototypical visual sketch trait for instance-specific latent domain discovery. We evaluate our model using the semantic categories in Sketchy [24] and Shoe-V2 [37] – the two largest FG-SBIR datasets to date in terms of the overall and single-category size respectively. In contrast to their original within-category evaluation setup, we establish a new more challenging cross-category-FG-SBIR evaluation protocol that is more in line with real-world requirements. Extensive experiments validate the efficacy of our method compared to a variety of competitors including direct transfer, other approaches to defining instance-embeddings, and state-of-the-art domain generalisation methods.

2. Related Work

Fine-grained SBIR Most earlier SBIR studies [5, 10, 4, 14, 38, 7, 20] focus on category-level cross-domain matching. The finer-granularity retrieval of FG-SBIR recently became topical given the potential for real-world application – users would like to retrieve a specific object (e.g., an e-commerce product photo) given a mental picture. This was first studied in the case of pose [19] using deformable-part models and graph-matching. Subsequent research has focused issues surrounding multi-branch deep learning methods that learn to extract comparable features from these heterogeneous domains [24, 36, 28, 22]. For example heterogeneous vs. Siamese branches [22], instance matching losses (pairwise vs. triplet), attention [28] and improving efficiency via hashing [40]. All of these studies assumed training data was available for the specific categories within which fine-grained retrieval is to be performed. This makes the problem easier (no train-test domain shift), but the models less practically valuable.

Generalisable SBIR Generalising to novel categories beyond the training set is an important capability for computer vision to move out of the lab and impact the real world. This motivates, for example, extensive research in zero-shot object recognition [11, 42, 6]. Nevertheless, in the context of SBIR, only two previous works studied cross-category generalisation. Shen *et al.* focused on a three branch hashing network for efficient SBIR [27]. Yelamarthi *et al.* presents a deep conditional generative model, where a sketch is taken as input, and corresponding photo features are generated. Both studies make use of *category level* features to guide learning: [27] uses word-vectors to form an adjacency matrix to regularise the hidden representation, and [35] extracts ImageNet pre-trained photo features as guidance for sketch-feature regression. Our work differs from these in that (i) we are the first to study cross-category generalisation in FG-SBIR rather than in category-level SBIR as addressed by prior methods; (ii) our unique VTDs are learned to summarise abstract visual traits shared across categories in a data driven way (see Fig.1) rather than steered by category semantics – thus better facilitating their generalisation to novel categories.

Domain Generalisation The CC-FG-SBIR challenge can be seen as a special case of Domain Generalisation (DG) [16, 26, 18]. DG aims to train models that work 'out of the box' on testing data that is out-of-sample with respect to the training data. For example by careful training regularisation [26, 18], or assumptions about how to remove domain-specific biases [16]. A related line of work uses external meta-data about the new domain to synthesise an appropriate model on the fly [32, 13]. In the context of deep networks, such dynamic parameter synthesis has been termed hypernetworks [12] – where one network synthesises the weights of another [12, 2]. Our approach addresses the DG problem in CC-FG-SBIR by embedding the query sketch in our universal embedding space, and using this embedding as the descriptor of the new domain (in place of external descriptors [32, 13]) from which parts of the feature extraction network of both photo and sketch are synthesised (as per hypernetworks [12, 2]).

3. Methodology

Overview Our framework consists of two main components. Firstly, our *unsupervised embedding network* maps any sketch s into one of K unique visual trait descriptors D_s via an encoder-decoder framework $D_s = \phi(s)$. So the full set of M-dimensional trait descriptors defines a matrix $D \in \mathbb{R}^{K \times M}$. This serves to provide the description of any sketch's query domain. Secondly, a *dynamically paramaterised feature extractor with triplet loss* is formulated, which actually performs FG-SBIR by using the generated descriptor to adapt the feature extraction and retrieval to any query sketch. Denoting $\psi(\cdot)$ as Deep CNN feature extractor, FG-SBIR is performed by finding the photo p that minimises the distance $d_{psi(s)}(s,p) = ||\psi_{\phi(s)}(s) - \psi_{\phi(s)}(p)||_2^2$ to query sketch s. The unsupervised embedding network is trained in an unsupervised way on the training sketch categories. And the dynamically paramaterised FG-SBIR model is trained in a supervised way on the training sketch categories. No components touch the held out testing category data until evaluation. In the following two sections we describe each of these components in detail.

3.1. Universal Visual Trait Embedding

The unsupervised embedding network will map any sketch to an entry in a dictionary of descriptors D. Inspired by NAT [3], we pre-generate the descriptor dictionary at random so that each row of D, denoted D_i is sampled from the standard Gaussian and then ℓ_2 normalised. This ensures that the descriptor dictionary spans the available M dimen-

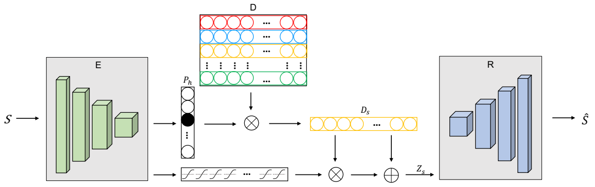


Figure 2: Schematic illustration of our proposed unsupervised encoder-decoder model. See details in text.

sional space well. The network's goal is then to learn to map any sketch onto one of these K (random) dictionary elements so that the representations of the full sketch dataset spread out over the whole embedding space.

Encoder-Decoder We start by feeding an input sketch s into a CNN encoder $E(s)$. We then use one fully-connected (FC) layer to predict a K-dimensional vector of unnormalised probabilities p and select the most probable one as sketch s's descriptor D_s out of the full dictionary D

$$\begin{aligned} p &= W_p E(s) + b_p \\ p_h &= \mathrm{onehot}(\mathrm{argmax}(\mathrm{softmax}(p))) \\ D_s &= p_h D, \quad \hat{s} = R(D_s) \end{aligned} \tag{1}$$

To ensure that each descriptor corresponds to a visually meaningful trait, the assigned descriptor is then decoded by decoder R with de-convolutional layers that reconstruct the input sketch $\hat{s} \approx s$. We denote the extraction of a sketch trait descriptor in this way as $D_s = \phi(s)$.

A Practical Consideration Since the number of descriptors K (300) is much less than sketches (tens of thousands), our approach means that sketches will be coarsely quantised, and reconstruction error will be high. (The clusters do not contain enough information to accurately reconstruct each sketch). Therefore we modify this approach with the following skip connection to improve the decoding via R.

$$\begin{aligned} Z_s &= D_s(1 + \alpha tanh(W_{sk}E(s) + b_{sk})) \\ \hat{s} &= R(Z_s) \end{aligned} \tag{2}$$

where we set $\alpha = 0.02$. This passes through some detailed features of the sketch to augment the coarse dictionary encoding. See Fig. 2 for an intuitive illustration.

Optimisation The method as presented so far is hard to optimise because: (i) The use of argmax is non-differentiable and would naively require Monte Carlo estimates and a REINFORCE-type algorithm [30], which suffers from high variance. (ii) A trivial minimiser of the reconstruction loss is to output one or few constant one-hot vectors p_h. Especially in the early phase of training, this will trap the model in a local minima forever. To alleviate this problem, we employ a low-variance gradient estimated based on a reparameterisation trick.

Hard Assignment via Gumbel-Softmax Applying the Gumbel-Softmax reparameterisation trick [15] and straight-through (ST) gradient estimator, p_h is replaced as

$$\begin{aligned} p_g &= \mathrm{softmax}((p + g)/\tau) \\ p_{hg} &= \mathrm{onehot}(\mathrm{argmax}(p_g)) \end{aligned} \tag{3}$$

where $g \in \mathbb{R}^K$ with $g_1 ... g_k$ are i.i.d samples drawn from $\mathrm{Gumbel}(0,1)$, and τ is the temperature[1]. We further enforce a uniform categorical prior on $p_s = \mathrm{softmax}(p)$ to avoid sketches being assigned to only a subset of dictionary elements, and form a Kullback-Leibler loss as:

$$\begin{aligned} q_y &= [1/K, 1/K, ..., 1/K] \in \mathbb{R}^K \\ D_{\mathrm{KL}}(p_s || q_y) &= \frac{1}{B} \sum_{i=1}^{B} \mathbf{p_s}_{i,:} \log(\mathbf{p_s}_{i,:}/q_y) \end{aligned} \tag{4}$$

where B is the batch size. For simplicity, we use bold $\mathbf{p_s}$ to denote the batch counterpart of p_s, with $\mathbf{p_s}_i$ the i^{th} example and $\mathbf{p_s}_{i,j}$ as its j^{th} element. We will follow this convention for other symbols. This ensures that across the batch as a whole, sketches are assigned to diverse descriptors.

Soft Assignment via Entropy Constraint We also explore an alternative strategy, which is to adopt a soft assignment approach during training. By replacing p_h with p_s, each sketch takes a linear combination of D, rather than selecting a row of D for representation learning. In this soft assignment of sketches to descriptors, we want to motivate sparse probabilities so that each s tends to receive one dominant label assignment. Thus we add a row entropy loss:

$$\mathrm{H}_{\mathrm{row}} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{K} \mathbf{p_s}_{i,j} \log(\mathbf{p_s}_{i,j}) \tag{5}$$

Eq. 5 achieves its minimum 0 only if $\mathbf{p_s}_i$ is an one-hot vector specifying a deterministic distribution. We further encourage equal usage of all $\mathbf{p_s}_{:,j}$ via a column entropy term:

$$\begin{aligned} p_c &= \frac{1}{B} \sum_{i=1}^{B} \mathbf{p_s}_{:,j} \in R^K \\ \mathrm{H}_{\mathrm{col}} &= -\sum_{j=1}^{K} p_{c_j} log(p_{c_j}) \end{aligned} \tag{6}$$

Eq. 6 achieves its maximum 1 only if elements in p_c are uniformly distributed. However, the row entropy constraint is only valid for a large enough minibatch and we empirically find that on average around 30% of p_h are still

[1] Since we are using straight-through argmax, τ will not adaptively adjust the "confidence" of proposed samples during the training process. For the forward pass, p_{hg} is used thus a real one-hot vector is generated, while for the backward pass, it is replaced by p_s to make the (estimated) gradient flows back. In practice, we just assign it a mild value like 1.0.

empty, (no assignments of any sketches). Therefore, we dynamically replace the stale and inactive D_i during training and bring them back in to compete with over-active ones. Specifically, we extract p_h of all training sketches after each epoch, and select the most concentrated D_i. A small random perturbation is then added to define a new centre, i.e., $D_i(1+\beta\mathcal{N}(0,1))$. We find this simple strategy works well[2].

Summary Depending on which assignment strategy we use (Gumbel-Softmax vs. Entropy), and combined with reconstruction loss $L_{\text{rec}} = ||s - \hat{s}||_2$, we obtain our two optimisation objectives:

$$\begin{aligned} &\min \mathbb{E}_{s\sim S}[L_{\text{rec}} + \lambda_{\text{KL}} D_{\text{KL}}(p_s||q_y)] \\ &\min \mathbb{E}_{s\sim S}[L_{\text{rec}} + \lambda_{\text{row}}\text{H}_{\text{row}} - \lambda_{\text{col}}\text{H}_{\text{col}}] \end{aligned} \tag{7}$$

where hyper-parameters $\lambda_{\text{KL}}, \lambda_{\text{row}}, \lambda_{\text{col}}$ control the relative weighting importance. In summary, optimising the unsupervised objective Eq. 7 trains an autoencoder that internally represents sketches in terms of a pre-defined K-element dictionary D. In the following section, we will re-use the sub-network that assigns sketches to dictionary elements $D_s = \phi(s)$ as a descriptor for dynamically paramaterising our FG-SBIR network.

3.2. Dynamic Parameterisation for FG-SBIR

The unsupervised embedding network shown in Fig. 2 extracts a visual trait descriptor (VTD), $\phi(s)$, from each sketch, which is then used to parameterise a triplet ranking network (TRN), $\psi(\cdot)$, for learning domain-generalisable representations for sketch and photo, as illustrated in Fig. 1. Note that sketch and photo feature extractors ψ is Siamese – applied to both sketch and photo for FG-SBIR. Denoting $\psi_{\phi(s)}(\cdot)$ as the feature extractor calibrated to sketch s, and $F(\cdot)$ as a vanilla CNN feature extractor, we have:

$$\psi_{\phi(s)}(\cdot) = \eta(\phi(s)) \odot F(\cdot) + F(\cdot) \tag{8}$$

The above can be interpreted as a small hypernetwork [12], where we generate a sketch-conditional diagonal weight layer to adapt the conventional CNN feature F to the current sketch, along with a residual connection. It can also be interpreted as a generating a sketch-specific soft attention mask on F where η indicates salient dimensions. Using this dynamically paramaterised feature extractor, we finally apply a standard triplet loss to match photos and sketches:

$$\begin{aligned} L_{\text{tri}} = \max(0, \Delta + d(\psi_{\phi(s)}(s), \psi_{\phi(s)}(p^+)) \\ - d(\psi_{\phi(s)}(s), \psi_{\phi(s)}(p^-))) \end{aligned} \tag{9}$$

A Stochastic Paramaterisation A standard solution for the weight generator $\eta(\cdot)$ in Eq. 8 is to transform the input sketch embedding through a few FC layers [12]. However, as the input is a discrete set of descriptor vectors, this causes discontinuity in weight generation. We take inspiration from [39] and mitigate this by introducing layers that predict a Gaussian mean and variance, and then sample these to more smoothly generate the target parameters.

$$\begin{aligned} \mu_s &= W_\mu \phi(s) + b_\mu \\ \sigma_s &= \exp(\frac{W_\sigma \phi(s) + b_\sigma}{2}) \\ \eta(\phi(s)) &= \mu_s + \sigma_s \odot \mathcal{N}(0,1). \end{aligned} \tag{10}$$

Optimisation and Inference Finally, to avoid generative model overfitting [9], we add the commonly applied variational regularisation term, $L_{\text{con}} = D_{\text{KL}}(\eta(\phi(s))||\mathcal{N}(0,I))$, weighted by a small value λ_{con}. Our FG-SBIR objective is:

$$\min \mathbb{E}_{t\sim T}[L_{\text{tri}} + \lambda_{\text{con}} L_{\text{con}}] \tag{11}$$

where t stands for a triplet tuple, consisting of $\{s, p^+, p^-\}$. During testing, for a query sketch s, we sample $\eta(\phi(s))$ ten times to calculate distance for each sketch-photo gallery pair and take the smallest as the final measure.

4. Experiments

4.1. Experimental Settings

Dataset and Pre-processing We use the public Sketchy [24] and QMUL-Shoe-V2 [37] to evaluate our methods. Sketchy contains 125 categories with 100 photos each and at least 5 sketches per photo. We follow the same dataset split as [35] and partition Sketchy into 104 train and 21 test categories to ensure the test ones are not present in 1000 ImageNet Challenge classes [23]. For QMUL-Shoe-V2, we test generalisation by transferring between fine-grained sub-categories and design five groups of such experiments as shown in Table 2. We scale and centre the sketches to 64×64 when training VTD, while for FG-SBIR, the inputs of all three branches are resized to 299×299.

Implementation Details We implement both models in Tensorflow on a single NVIDIA 1080Ti GPU. **Unsupervised Embedding Network**: Our CNN-based encoder-decoder, E and R, contains five stride-2 convolutions and five fractional-convolutions with stride 1/2, with one 1×1 convolution at the end and start of each. BatchNorm-Relu activation is applied to every convolutional layer, except the output of R with Tanh. All hyper-parameters are set to undergo a warm-up phase, so that reconstruction loss dominates the training at the beginning. We train the models for 200 epochs under all settings with λ_{kl}, λ_{row}, λ_{col} linearly increasing from $0, 1, 1$ to $1.5, 2, 10$ respectively. The dictionary D has $M = 256$ dimensions and $K = 300$ elements throughout. We use Adam optimiser with learning rate 0.0002. **FG-SBIR**: we fine-tune ImageNet-pretrained

[2] A side effect is to trade quality with time. We spend almost one-third of the time extracting representations for all training sketches. We set $\beta = 0.05$ throughout the experiments and find it works well empirically.

Competitor	Acc.@ 1	Acc.@ 5	Acc.@ 10	Competitor	Acc.@ 1	Acc.@ 5	Acc.@ 10
Hard-Transfer [36]	16.0%	40.5%	55.2%	Ours-WordVector	18.0%	43.5%	58.7%
CVAE-Regress [35]	2.4%	9.5%	17.7%	Ours-Classify	16.2%	41.4%	57.2%
Reptile [1]	17.5%	42.3%	57.4%	Ours-Full/Edge	16.8%	41.3%	56.2%
CrossGrad [26]	13.4%	34.9%	49.4%	Ours-Full/Hard	20.1%	46.4%	61.7%
Ours-VAE	12.7%	34.5%	49.7%	Ours-Full	**22.6%**	**49.0%**	**63.3%**
Ours-VAE-Kmeans	17.6%	41.9%	56.9%	Upper-Bound	29.9%	65.5%	81.4%

Table 1: Comparative Cross-Category FG-SBIR results on Sketchy [24].

Sub-category	Fine-grained Transfer	No. Train / Test	Competitor	Acc.@ 1	Acc.@ 5	Acc.@ 10
Sandal	Flat → Wedge	560 / 227	Hard-Transfer	9.25%	32.2%	48.0%
			Ours-Sketchy	13.2%	34.3%	50.4%
			Ours-Sketchy-Ft	**15.4%**	**37.9%**	**54.6%**
			Upper-Bound	28.6%	56.8%	72.2%
Toe-shape	Closed → Fish-mouth	400 / 351	Hard-Transfer	14.8%	44.7%	61.5%
			Ours-Sketchy	22.2%	50.4%	65.0%
			Ours-Sketchy-Ft	**24.2%**	**54.5%**	**66.7%**
			Upper-Bound	29.3%	56.7%	71.8%
Shoe-height	Ankle- → Knee-high	2010 / 245	Hard-Transfer	10.6%	32.2%	43.3%
			Ours-Sketchy	14.7%	38.0%	51.0%
			Ours-Sketchy-Ft	**18.4%**	**40.8%**	**55.1%**
			Upper-Bound	25.3%	54.3%	71.8%
Heel-shape	Thick → Thin	828 / 411	Hard-Transfer	12.2%	35.0%	48.7%
			Ours-Sketchy	15.1%	**41.4%**	**59.4%**
			Ours-Sketchy-Ft	**17.3%**	41.1%	57.7%
			Upper-Bound	26.3%	61.8%	80.5%
Topline	Small → Big	5015 / 1543	Hard-Transfer	7.25%	22.9%	34.5%
			Ours-Sketchy	12.2%	28.9%	39.7%
			Ours-Sketchy-Ft	**15.5%**	**31.4%**	**43.8%**
			Upper-Bound	19.6%	44.2%	61.5%

Table 2: Comparative Cross-Category FG-SBIR results on QMUL-Shoe-V2 [37]

Inception-v3 [29] to obtain F with the final classification layer removed. We enforce ℓ_2 normalisation on the output of η to stabilise triplet learning and set hyper-parameters $\Delta = 0.1, \lambda_{con} = 0.004$. We train for 20 epochs on Sketchy, and 10 epochs on QMUL-Shoe-V2 with a learning rate of 0.0001 and Adam optimiser under all settings.

Evaluation Metric We use Acc.@ K to measure the FG-SBIR performance, which is the percentage of sketches whose true-match photos are ranked in the top K.

4.2. Competitors

Sketchy If not otherwise mentioned, all competitors are implemented based on Inception-v3, and our model is trained with soft assignment. **Hard-Transfer** [36] trains a vanilla Siamese triplet ranking model and is directly tested on unseen categories. **CVAE-Regress**[3] [35] is the state-of-the-art zero-shot SBIR method by learning a conditional generative model to regress ImageNet-pretrained photo features to their corresponding sketch features. **Reptile** [1] is a recent meta-learning algorithm that repeatedly samples tasks, trains them, and moves the initialisation towards the trained weights. We integrate it in [36] by each time randomly sampling 52 categories to form two subtasks and train parallelly for 500 iterations. **CrossGrad** [26] is a state-of-the-art domain generalisation method that trains both a label and a domain classifier on examples perturbed by each other's loss gradients. For our task, we regard each of 104 training categories as a unique domain and 100 inter-category photo ids as labels. **Ours-VAE** corresponds to training a conventional variational autoencoder (VAE) [17] without our visual trait descriptor and using the per-instance latent representation as the descriptor ϕ to parameterise the FG-SBIR model. **Ours-VAE-Kmeans** performs K-means clustering in the VAE latent space, to generate a dictionary of sketch descriptors analogous to our approach, but without end-to-end learning. **Ours-WordVector** and **Ours-Classify** replace our descriptor with the category-level semantics driven descriptor either drawn from the class name [21] or extracted from the penultimate feature layer of a sketch classification network. Lastly, we compare our proposed model (**Ours-Full**) with its two ablated versions, including **Ours-Full/Hard** and **Ours-Full/Edge**, which are trained with hard assignment strategy instead of soft, on edgemaps other than human freehand sketches respectively.

QMUL-Shoe-V2 This is a very-fine-grained single category dataset, so we do not have enough data to train a dic-

[3] This method is designed for category-level characterisation, so is expected to perform poorly.

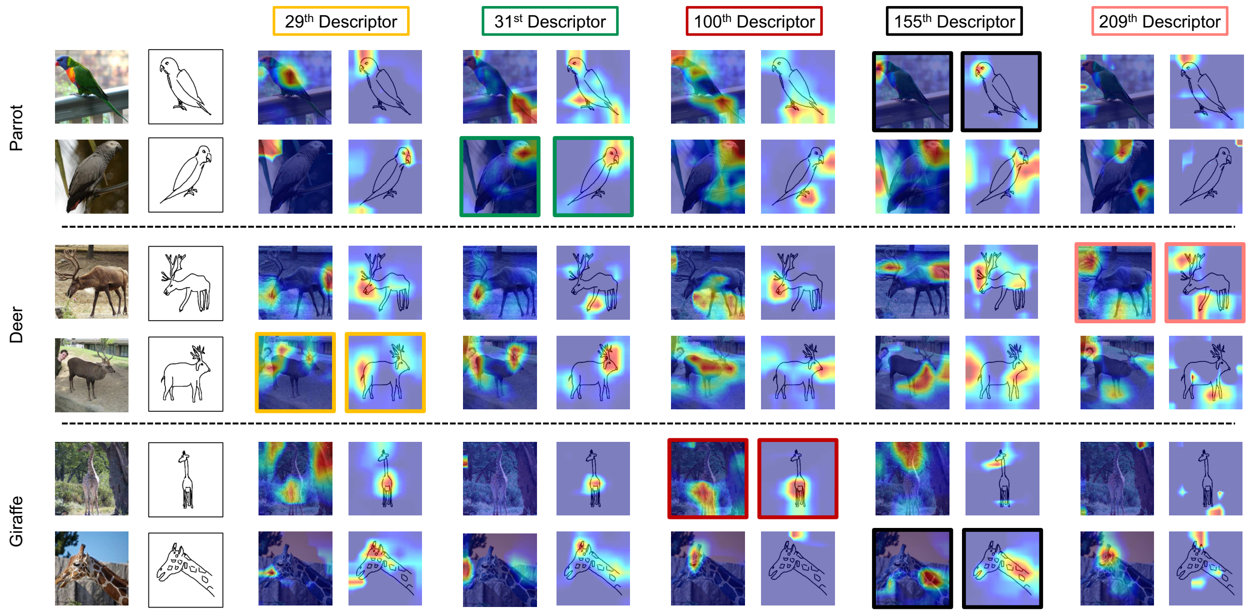
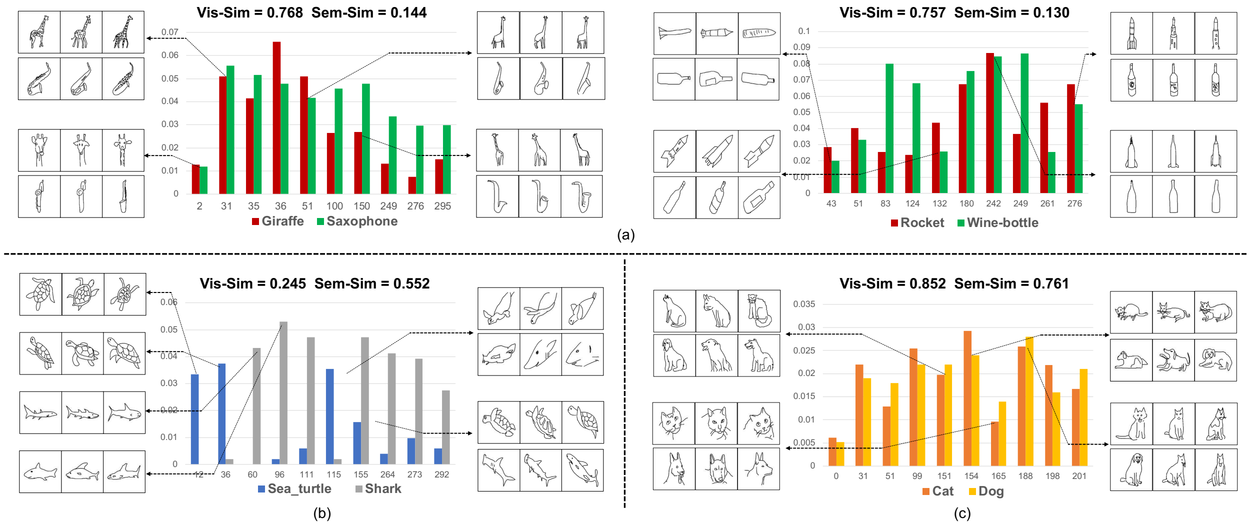


Figure 3: Visualisation of how the VTD adapts the sketch-photo matching process. Coloured image box border indicates when the correct (corresponding to query sketch) descriptor is used to paramaterise the embedding space.



(a)

(b)

(c)

Figure 4: Word-Vector vs. Visual-Semantics. Comparing illustrative category pairs: (a) Visually close but semantically far. (b) Semantically related but visually far. (c) Visually and semantically related. Vis-Sim is cosine distance between the histograms, and Sem-Sim is the cosine distance between word-vectors. Histograms shown here are the ten most similar / dissimilar descriptors jointly shared between two categories. Best viewed in colour and zoom.

tionary D from scratch. Therefore we take the advantage of the best visual trait descriptor trained on Sketchy and introduce two variants **Ours-Sketchy** and **Ours-Sketchy-Ft**. They differ in if we directly use the Sketchy dictionary or further fine-tune it on the seen sub-category of QMUL-Shoe-V2. **Hard-Transfer** is the competitor.

Caveat Since we use all images within one category for constructing a challenging test set. The **Upper-Bound** for both datasets is therefore likely a slight overestimate, as it uses half of these for training before before testing on all.

4.3. Results on Sketchy

We compare the performance of different models in Table 1 and make the observations: (i) The gap between direct transfer (16%) and a model trained using data from the target (unseen) categories (Upper-Bound, 30%) is large, confirming the cross-category generalisation gap. (ii) Our model beats all 10 competitors in bridging this gap. (iii) For DG meta-learning competitors, CrossGrad [26] fails to improve on the direct transfer baseline, but Reptile [1] does improve on it. However both are worse than our full model. (iv) Comparing our two proposed optimisaton methods, soft assignment outperforms hard. We attribute this to the rigid approach of the latter – it enforces a uniform distribution over assignment to descriptors, which may not hold in practice since some will be more common than others. (v) Our visual trait descriptor approach is beneficial as manifested by the dramatic performance gap between ours and the conventional VAE, VAE-Kmeans alternatives in particular. (vi) Using visually abstract but neat human free-hand sketches as source data to train our descriptor is important. Replacing these with the detailed but noisy edgemaps extracted from natural photos hurts the performance. This suggests that the model is able to exploit the clean and iconic free-hand sketches to learn abstract visual traits more effectively.

Qualitative Impact of Descriptors We now qualitatively examine how a visual trait descriptor $D_s = \phi(s)$ impacts sketch photo matching and how retrieval is affected if using another sketch descriptor $D_{\hat{s}}, \hat{s} \neq s$ instead. To achieve this, we select one dimension from $\psi_{\phi(s)}$ that contributes the most to successful matching and use Grad-Cam [25] to propagate gradients back to highlight discriminative image regions. This can be seen as a visualisation of the implicit attention mechanisms that different visual trait descriptors define to adapt the feature extraction. We illustrate this in Fig. 3 across five different D_ss for each of six sketch-photo pairs. It shows that (i) The corresponding D_s helps focus attention on regions with similar spatial support for both s and p^+, while a mismatched $D_{\hat{s}}$ fails to do this; (ii) Individual descriptors D_i are useful for multiple categories, e.g., the 155^{th} descriptor for parrots and giraffes.

How Many Descriptors? We investigate the impact of the descriptor dictionary size K on CC-FG-SBIR performance in Table 3. We can see that our model is not very sensitive to K under either hard and soft assignment strategies, and a few hundred suffices for good performance.

Descriptor-Category Spread We can verify that VTDs cross-cut rather than mirror the category breakdown of sketches. On average, training sketches from each category are assigned to 138 ± 30 unique descriptors. Testing category sketches (upon which the embedding is not trained) are assigned to 129 ± 33 descriptors, indicating that the cross-cutting spread is retained despite the train/test domain-shift.

Word-Vector vs. Visual-Semantics The quantitative results (Table 1) showed that word-vector descriptors do improve performance over hard-transfer, albeit much less than our approach. We can contrast similarity as estimated by word-embeddings, with that of our VTD. Fig. 4(a) shows a pair of categories which are far in semantic word similarity, but near in visual visual trait descriptor similarity. Here *category level* visual similarity is measured by the number of sketches (y-axis) from different categories (bars) co-assigned to a single descriptor (x-axis). In contrast, Fig.4(b) shows semantically related categories that are visually distinct (shark/sea turtle) and Fig.4(c) illustrates categories that are both semantically and visually related (dog/cat).

No.	Hard			Soft		
	Acc.@ 1	Acc.@ 5	Acc.@ 10	Acc.@ 1	Acc.@ 5	Acc.@ 10
20	18.4%	43.3%	58.4%	19.5%	46.0%	60.4%
100	19.6%	45.7%	60.9%	20.7%	47.7%	62.7%
300	**20.1%**	**46.4%**	**61.7%**	**22.6%**	**49.0%**	**63.3%**
1000	17.8%	42.3%	57.6%	18.3%	43.8%	59.0%

Table 3: Effects of the number of descriptors on Cross-Category FG-SBIR performance on Sketchy [24].

4.4. Results on QMUL-Shoe-V2

In this section, we borrow the best VTD dictionary D (Ours-Full-Soft) trained on Sketchy and use it to help transfer between sub-categories in QMUL-Shoe-V2. To test generalisation on this benchmark, we design five groups of experiments, each defining a different type of train/test gap, and with diverse split sizes. We report their performance in Table 2 and find that compared with Hard-Transfer, even when directly applying D to this novel dataset, Ours-Sketchy improves performance in all experiments. This is promising as a Sketchy-trained dictionary is generally applicable and it has potential to benefit other specific FG-SBIR applications. When further fine-tuned on the *train* data split of each experiment, we also usually improve performance (Ours-Sketchy-Ft and Ours-Sketchy).

5. Conclusion

We have for the first time identified the generalisation problem in cross-category FG-SBIR and proposed a novel solution via learning a universal visual trait descriptor embedding. This embedding dictionary is mapped to a set of latent domains that cross-cut sketch categories, and enable a retrieval network to be suitably paramaterised given a query sketch – by mapping query sketches to the corresponding descriptor in the dictionary. Extensive experiments on Sketchy and QMUL-Shoe-V2 demonstrate the superiority of our proposed method for cross-category FG-SBIR.

Acknowledgements: We thank the anonymous reviewers for their valuable comments. This work was supported in part by National Science Foundation of China (NSFC) under joint grant # 61701032 and # 61806184. We especially would like to thank the China Scholarship Council (CSC) for funding the first two authors to conduct the entirety of this project at the SketchX research lab.

References

[1] Nichol Alex and Schulman Johnn. Reptile: A scalable meta-learning algorithm. https://blog.openai.com/reptile/, 2018. 6, 8

[2] Luca Bertinetto, João F Henriques, Jack Valmadre, Philip Torr, and Andrea Vedaldi. Learning feed-forward one-shot learners. In *NIPS*, 2016. 3

[3] Piotr Bojanowski and Armand Joulin. Unsupervised learning by predicting noise. In *ICML*, 2017. 2, 3

[4] Yang Cao, Changhu Wang, Liqing Zhang, and Lei Zhang. Edgel index for large-scale sketch-based image search. In *CVPR*, 2011. 3

[5] Yang Cao, Hai Wang, Changhu Wang, Zhiwei Li, Liqing Zhang, and Lei Zhang. Mindfinder: interactive sketch-based image search on millions of images. In *ACM MM*, 2010. 3

[6] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, 2016. 3

[7] John Collomosse, Tu Bui, Michael J Wilber, Chen Fang, and Hailin Jin. Sketching with style: Visual search with sketches and aesthetic context. In *ICCV*, 2017. 3

[8] Gabriela Csurka. *Domain Adaptation in Computer Vision Applications*. Springer, 2017. 1, 2

[9] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016. 5

[10] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *TVCG*, 2011. 3

[11] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013. 3

[12] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. In *NIPS*, 2017. 3, 5

[13] Judy Hoffman, Trevor Darrell, and Kate Saenko. Continuous manifold based adaptation for evolving visual domains. In *CVPR*, 2014. 3

[14] Rui Hu and John Collomosse. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *CVIU*, 2013. 3

[15] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017. 2, 4

[16] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *ECCV*, 2012. 3

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 6

[18] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018. 3

[19] Yi Li, Timothy M. Hospedales, Yi-Zhe Song, and Shaogang Gong. Fine-grained sketch-based image retrieval by matching deformable part models. In *BMVC*, 2014. 1, 3

[20] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, 2017. 1, 3

[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 6

[22] Kaiyue Pang, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. Cross-domain generative learning for fine-grained sketch-based image retrieval. In *BMVC*, 2017. 1, 3

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 5

[24] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: Learning to retrieve badly drawn bunnies. *SIGGRAPH*, 2016. 1, 3, 5, 6, 8

[25] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 8

[26] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *ICLR*, 2018. 2, 3, 6, 8

[27] Yuming Shen, Li Liu, Fumin Shen, and Ling Shao. Zero-shot sketch-image hashing. In *CVPR*, 2018. 3

[28] 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, 3

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

[30] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 1992. 4

[31] Zheng Xu, Wen Li, Li Niu, and Dong Xu. Exploiting low-rank structure from latent domains for domain generalization. In *ECCV*, 2014. 2

[32] Yongxin Yang and Timothy M Hospedales. A unified perspective on multi-domain and multi-task learning. In *ICLR*, 2015. 2, 3

[33] Yongxin Yang and Timothy M. Hospedales. Multivariate regression on the grassmannian for predicting novel domains. In *CVPR*, 2016. 2

[34] Hana Ajakan Pascal Germain Hugo Larochelle Franois Laviolette Mario Marchand Victor Lempitsky Yaroslav Ganin, Evgeniya Ustinova. Domain-adversarial training of neural networks. In *JMLR*, 2016. 2

[35] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *ECCV*, 2018. 3, 5, 6

[36] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen Change Loy. Sketch me that shoe. In *CVPR*, 2016. 1, 3, 6

[37] Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. SketchX! - Shoe/Chair fine-grained SBIR dataset. http://sketchx.eecs.qmul.ac.uk, 2017. 3, 5, 6

[38] Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao. Sketchnet: Sketch classification with web images. In *CVPR*, 2016. 3

[39] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 5

[40] 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. 3

[41] 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

[42] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via joint latent similarity embedding. In *CVPR*, 2016. 3