Cross-domain Generative Learning for Fine-Grained Sketch-Based Image Retrieval

Kaiyue Pang[1]
kaiyue.pang@qmul.ac.uk
Yi-Zhe Song[1]
yizhe.song@qmul.ac.uk
Tao Xiang[1]
t.xiang@qmul.ac.uk
Timothy M. Hospedales[1, 2]
t.hospedales@ed.ac.uk

[1] SketchX Research Lab
Queen Mary University of London
London, UK

[2] School of Informatics
University of Edinburgh
Edinburgh, UK

Abstract

The key challenge for learning a fine-grained sketch-based image retrieval (FG-SBIR) model is to bridge the domain gap between photo and sketch. Existing models learn a deep joint embedding space with discriminative losses where a photo and a sketch can be compared. In this paper, we propose a novel discriminative-generative hybrid model by introducing a generative task of cross-domain image synthesis. This task enforces the learned embedding space to preserve all the domain invariant information that is useful for cross-domain reconstruction, thus explicitly reducing the domain gap as opposed to existing models. Extensive experiments on the largest FG-SBIR dataset Sketchy [19] show that the proposed model significantly outperforms state-of-the-art discriminative FG-SBIR models.

1 Introduction

Fine-grained sketch-based image retrieval (FG-SBIR) addresses the problem of finding a specific photo containing the same instance as in an input sketch. It has received increasing research attention recently [12, 14, 19, 22, 23, 27], due to its potential in commercial applications such as searching online product catalogues for shoes, furniture, and handbags by finger-sketching on a touchscreen device.

FG-SBIR is a very challenging problem and remains unsolved due to the facts that: (i) Sketch and photo are two distinctive image domains – the former is characterised by sparse black line drawings with white background whilst the latter consists of dense colour pixels. Such a large domain gap underpins the main challenge of the retrieval task. (ii) Sketches often exhibit a varied level of abstraction and sophistication, especially when drawn based on a mental picture of an object without a reference photo. With variable levels of human drawing skill, there could be large discrepancies in shape and spatial misalignments both globally and locally between a matching pair of photo and sketch, further enlarging the domain gap. Such discrepancies and misalignments also vary across different classes as

humans are often good at drawing certain classes but not others. (iii) Collecting sufficient photo-sketch instance pairs are tedious and expensive. As a result, even the largest FG-SBIR dataset [19] contains limited data. The primary objective of learning a FG-SBIR model is thus to reduce the domain gap with limited training data so that images from the two domain become comparable.

The state-of-the-art FG-SBIR models [19, 27] are deep models that aim to close the domain gap by learning a joint feature embedding for the two domains. Concretely, multi-branch deep convolutional neural networks (CNNs) are employed where each branch corresponds to one domain and the final shared layer defines the embedding space which is subject to various discriminative losses such as pairwise contrastive loss or triplet ranking loss. These losses are designed to pull matching pairs of photos and sketches close and push mis-matched pairs away. These models thus indirectly align the two domains. However, with limited training data and by focusing only on discriminative losses, these models struggle to capture all the domain-invariant information and thus generalise poorly to test data where the domain discrepancies and misalignments could be different from those in the training data.

In this paper, we propose a novel discriminative-generative hybrid deep neural network for FG-SBIR. Our model also aims to learn a joint embedding space. The key difference to the existing models is that we introduce a generative task of cross-domain image synthesis. Concretely when an input photo is embedded in the joint space, the embedding vector is used as input to a generative model to synthesise the corresponding sketch. By doing so, we explicitly enforce the model to preserve all the domain-invariant information in the embedding space. This richer representation thus enables the model to generalise better to unseen test data. More specifically, the proposed model is a multi-branch cross-domain deep encoder-decoder model (see Figure 1). The encoder in each branch is a deep CNN that takes an image as input and outputs a feature embedding vector. This vector is then used as input to a deep transposed-convolutional (deconvolutional) network [28] regularised by the reconstruction loss to reconstruct the corresponding sketch. It is a discriminative-generative hybrid model because both discriminative and generative losses are used for learning the embedding, corresponding to the photo-sketch matching discriminative task and the cross-domain image synthesis generative task respectively.

The contributions of this work are as follows: (1) For the first time, problem of FG-SBIR is solved with a discriminative-generative hybrid model which explicitly aligns the sketch and photo domains to improve the model generalisation ability. (2) A multi-branch cross-domain deep encoder-decoder model is formulated; and in-depth analysis is provided on the model architectural design. Extensive experiments on the largest FG-SBIR dataset Sketchy [19] show that our model yields state-of-the-art performance and validate our claim that the performance gain is contributed by introducing the additional generative learning task.

2 Related Work

Fine-grained SBIR Most existing SBIR works [2, 3, 7, 13, 25, 29] focus on category-level sketch-to-photo retrieval. The FG-SBIR problem was first tackled in [12] which employed a deformable part-based model (DPM) representation followed by a graph matching strategy for cross-domain pose correspondence. The similarity measure is based on only pose and ignores other fine-grained details. The two most recent FG-SBIR models [19, 27] are all based on deep learning. Both models [19, 27] are multi-branch CNNs designed to learn a joint embedding space for the two domains. The popular pairwise contrastive loss and triplet

ranking loss are evaluated by both works and the latter is shown to be better in both models. The two models differ mainly in whether the photo and sketch CNN branches are Siamese (i.e., with weight sharing) or heterogeneous (i.e., without sharing). Our network is similar to the model in [27] in that it is a Siamese two-branch CNN, but differs from both in that it has an additional generative decoder. We show that our model outperforms both state-of-the-art models [19, 27] significantly.

Deep Generative Models Remarkable progress has been made on deep generative models in the past 2-3 years, which can be broadly categorised into Variational Autoencoder (VAE) [10], Autoregressive Model [24] and Generative Adversarial Network (GAN) [5]. These advances have been actively applied to various practical applications including single-image super resolution [11], video frame prediction [15] , image to image translation [8], text to image synthesis [17] and artistic style transfer [4]. Among the existing deep generative models, the most relevant ones are the deep encoder-decoder models. Such a model takes an image as input and produces a bottleneck latent code embedding via the encoder, which is then used by the decoder as input to generate an image that shares the same identity or semantic information. Comparing to the existing deep encoder-decoder models [8, 20, 26, 30] our model has a vital difference: it is a hybrid model and the main objective is to learn a discriminative feature representation with image synthesis used only as an auxiliary task designed to serve the main objective. In other words, we do not care about image synthesis quality which is the sole purpose of those models.

Deep discriminative-generative hybrid models Early works [1, 6] have exploited the general idea of using autoencoder reconstruction as an auxiliary generative task for dimension reduction and unsupervised feature learning. Recently, deep discriminative-generative hybrid models [16, 31] have been proposed, which inspired this work. Note that in these models, although the generative task is integrated with the discriminative task, it is still fundamentally different in that it does single domain rather than cross-domain reconstruction. In addition, these models use lateral connections between the corresponding convolution/deconvolution layers that relies heavily on the reconstruction loss (unsupervised generative counterparts) to build better hierarchical representation. In contrast, we mainly use the generative decoder as an auxiliary task to learn a more generalisable cross-domain representation. This fundamental difference is reflected by our asymmetrical design choice, i.e., the decoder's architecture does not mirror that of the encoder and a different learning strategy to favour the learning of the encoder over the decoder.

3 Methodology

3.1 Network Architecture

Overview The overall network architecture of the proposed discriminative-generative hybrid FG-SBIR model is illustrated in Figure 1. It consists of four sub-networks: (1) a three-branch Siamese encoder subnet E that aims to learn a joint embedding space for matching input sketch-photo pairs, (2) a Siamese decoder subnet D that takes an embedding vector and reconstruct a target sketch, (3) a classification subnet C to make the embedding vector class-discriminative and (4) a triplet ranking subnet T to make the vector instance-discriminative. Each encoder branch has the same base network and share their parameters, hence the name Siamese; so does each decoder branch. The four subnets are connected by the joint embedding layer: it is the output of E and input of D, C and T.

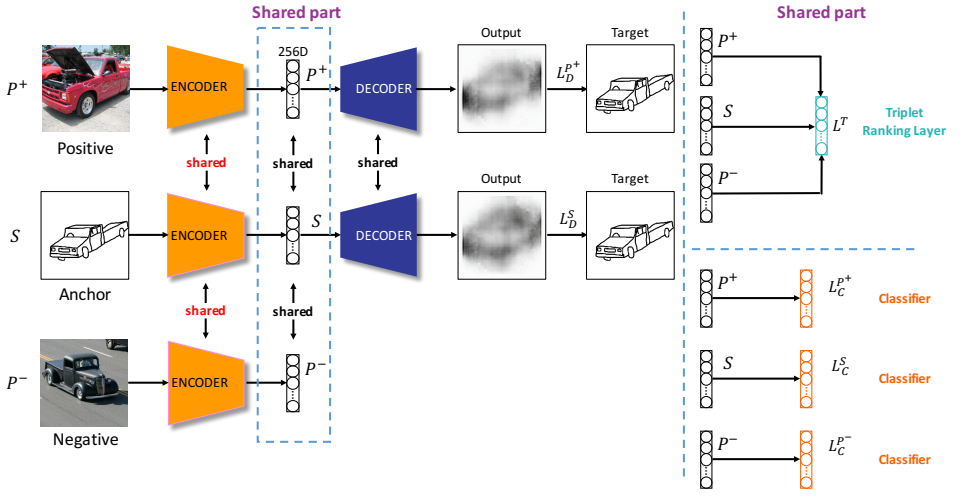

Figure 1: Architecture of the proposed deep cross-domain encoder-decoder FG-SBIR model.

Encoder The encoder architecture is based on that of VGGNet [21], which has been widely used as the base network in many vision applications. The final classification layer of the network pretrained on classifying the 1000 ImageNet classes is dropped and an additional shared 256-D fully-connected (FC) layer is added after the 4096-D penultimate FC layer of VGGNet. The ℓ_2 normalised 256-D output of the encoder is the joint embedding layer and once learned shall be used as the feature representation for both domains for retrieval.

Classification Subnet Although FC layers can be added in the classification subnet, in this work, the classifier directly feeds the latent code to a softmax layer with classification loss L_C being the cross-entropy loss, and the number of output nodes equalling the number of object categories. The classification loss makes sure that the learned embedding space preserves class-discriminative information.

Triplet Ranking Subnet Similar to the classification subnet, we directly add the triplet ranking layer after the shared 256-D embedding layer. In this subnet, each instance tuple $\{s, p^+, p^-\}$ contains an anchor sketch s, a positive photo p^+ containing the same object instance and a negative photo p^-. The subnet has three branches and the goal is to learn a instance-discriminative embedding space where the positive photo p^+ is ranked above the negative photo p^- in terms of its distance to the query sketch s. Note our model is flexible in that any instance-discriminative loss can be used. But as in [19, 27], we found that the triplet ranking loss alone works the best.

Decoder The decoder network consists of five upsampling blocks and one final convolution block with a filter size of 4×4 (see Table 1 for details). Each upsampling block has the structure of Deconvolution-BatchNorm(BN)-ReLU, except the final layer which uses Deconvolution-Tanh for generating the final output. Compared with the encoder-decoder architectures in existing deep generative models [8, 20, 26, 30], ours differs in that: (i) The decoder is not architecturally symmetric with the encoder. (ii) The decoder is much shallower than the encoder. This design is due to the factor that with the limited training sketch-photo pairs, a deeper decoder network would be prone to overfitting which can make the training process unstable. Furthermore, rather than producing a loyal reconstruction output, the sole

Input Size	Filters	Stride	BN	Activation
7 x 7 x 512	512	2 x 2	Yes	ReLU
14 x 14x 512	256	2 x 2	Yes	ReLU
28 x 28 x 256	128	2 x 2	Yes	ReLU
56 x 56 x 128	64	2 x 2	Yes	ReLU
112 x 112 x 64	32	2 x 2	Yes	ReLU
224 x 224 x 32	3	1 x 1	No	Tanh

Table 1: Detailed Architecture of the decoder subnet.

objective of this decoder is to help the encoder to learn a richer representation in the embedding layer which is domain-invariant. (iii) The generative process is also asymmetric: We use the embedding vector of the anchor sketch to reconstruct itself, and the positive photo to also reconstruct the anchor sketch. The opposites are not attempted, i.e., sketch-to-photo and photo-to-photo reconstructions. The reason is simple: to compare a photo with a sketch, the additional colour and texture information in the photo domain has to be removed in the embedding layer, so any effort to recover that in the decoder would be futile.

3.2 Model Learning and Deployment

Learning Objectives Suppose the encoder, classification and decoder subnets are denoted as ϕ_E, ϕ_C, and ϕ_D, where they are parametrised by θ_E, θ_C and θ_D respectively. Given N sketch-photo triplets $\mathcal{X} = \{\mathbf{x}_i^s, \mathbf{x}_i^{p^+}, \mathbf{x}_i^{p^-}\}_{i=1}^N$ within a training batch, our learning objective is:

$$\arg\min_{\theta_E,\theta_C,\theta_D} \mathcal{L} = \mathcal{L}_\mathcal{C} + \lambda_D \mathcal{L}_\mathcal{D} + \lambda_T \mathcal{L}_\mathcal{T}, \tag{1}$$

where $\mathcal{L}_\mathcal{C}$ is the cross-entropy softmax loss for classification:

$$L_C = -\sum_{i=1}^{N} (\hat{p}_i{}^s \log p_i{}^s + \hat{p}_i{}^{p^+} \log p_i^{p^+} + \hat{p}_i{}^{p^-} \log p_i^{p^-}), \tag{2}$$

$$p_i^{\{s,p^+,p^-\}} = \frac{exp(\phi_C(\phi_E(\mathbf{x}_i^{\{s,p^+,p^-\}})))}{\sum_{j=1}^N exp(\phi_C(\phi_E(\mathbf{x}_j^{\{s,p^+,p^-\}})))}, \tag{3}$$

$\mathcal{L}_\mathcal{D}$ is the pixel-wise ℓ_2 reconstruction loss that takes either the input sketch or photo from a ground-truth pair as input and synthesises the input sketch[1] as

$$L_D = \sum_{i=1}^{N} ||\mathbf{x}_i^s - \phi_D(\phi_E(\mathbf{x}_i^s))||_2 + ||\mathbf{x}_i^s - \phi_D(\phi_E(\mathbf{x}_i^{p^+}))||_2, \tag{4}$$

and $\mathcal{L}_\mathcal{T}$ is the triplet ranking loss:

$$L_T = \sum_{i=1}^{N} \max(0, \Delta + ||\phi_E(\mathbf{x}_i^s) - \phi_E(\mathbf{x}_i^{p^+})||_2 - ||\phi_E(\mathbf{x}_i^s) - \phi_E(\mathbf{x}_i^{p^-})||_2). \tag{5}$$

[1] We have also experimented adding the popular adversarial loss [8] and found that the decoder would suffer from significant mode collapsing problems due to the visual sparsity of sketches, which is commonly observed in generative adversarially trained nets [18].

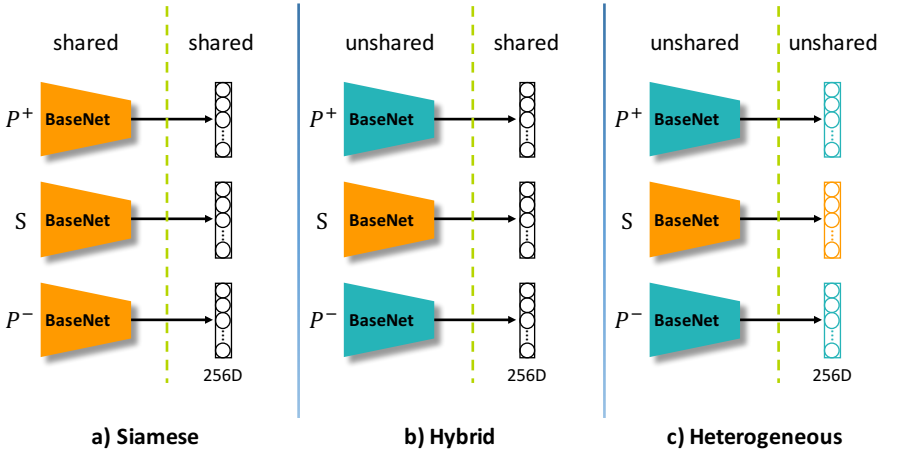

Figure 2: Three different weight sharing strategies for fine-grained SBIR task.

λ_D and λ_T weight the three losses.

Model training strategy The most straightforward way for training a deep model with multiple losses is to update all the parameters together; however the disadvantage of this strategy is that it could lead to detrimental competition between the downstream and upstream tasks. For example, when two sketches belong to different categories but exhibit similar structural and visual cues, θ_E may be sacrificed by pursuing the optimal θ_D. This motivates us an alternate training strategy that learns the encoder first, then fine-tune it with the decoder. One may argue that this would potentially undermine the interpretability of the decoder; nevertheless it is the encoder θ_E that this learning process really cares about, and the image synthesis quality is expandable. Specifically, we first minimise $\mathcal{L}_C + \lambda_D \mathcal{L}_\mathcal{D}$ with respect to θ_E and θ_C, then minimise $\mathcal{L}$ with respect to θ_E, θ_C and θ_D. In practice, we find this leads to more stable training behaviour.

Model Deployment Once trained, during testing the decoder, classifier and triplet ranking subnets are stripped off. Given a query sketch x^s, we compute the 256D feature representation and use its euclidean distance

$$Dist_{x^s,x^p} = ||\phi_E(\mathbf{x}^s) - \phi_E(\mathbf{x}^p)||_2 \tag{6}$$

to rank each photo x^p in the gallery set. Note that we can pre-compute the feature for each photo in the gallery set, which means the retrieval process only involves one forward pass of the encoder followed by Euclidean distance computation; it is thus very efficient.

3.3 Discussion on Weight-Sharing Strategies

As illustrated in Figure 2, three different weight-sharing strategies exist. Our multi-branch network is Siamese with weight-sharing everywhere between branches. The same strategy is adopted in [27]. In contrast, the network in [19] is heterogeneous meaning there is not weight sharing in any layer between the photo and sketch branches. The Siamese strategy attempts to align the two domains from the very beginning of feature extraction (convolution layers), whilst the heterogeneous one allows feature extraction filters as well as the embedding layer

to be learned independently and use the discriminative losses at the network output to align them. As far as domain alignment is concerned, leaving it to the end seems to be counter-intuitive; however, the heterogeneous network has one advantage: one could exploit the far bigger auxiliary data in each domain to pretrain each branch as in [19]. There is a third way that lies in-between these two extremes: a hybrid strategy whereby the branches are only tied at the joint embedding layer. In our experiments, all three strategies are evaluated.

4 Experiments

Dataset Experiments are conducted on the Sketchy dataset [19], which is the largest free-hand FG-SBIR dataset to date. It contains 125 categories with 100 photos per category and at least 5 sketches for one photo crowd-sourced from Amazon Mechanical Turk (AMT). We use the same training and testing split as in [19], where the held-out test set consists of 6312 query sketches and 1250 photos spanning all 125 categories. Another noticeable FG-SBIR dataset is the QMUL-Shoe-Bag dataset [27]. However, it is two-magnitudes smaller and contains only two categories. We found that the training of any deep model on this dataset is unstable making it difficult to draw any conclusion. It is thus not selected.

Implementation Details Our model is implemented on Tensorflow with a single NVIDIA Tesla P100 GPU. We set the importance weights for different subnets to: $\lambda_D = 10$, $\lambda_T = 1$, with the triplet loss margin $\Delta = 0.1$. The Adam optimiser [9] is used with the parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 10^{-8}$. The learning rate is set as 10^{-5} at first 20000 iterations and further decreased to 10^{-6} for another 10000 iterations with a batch size of 32. We used the uniformly scaled and centred version of sketches so that the learned representation is not sensitive to the absolute location and scale of a sketch. We randomly cropped an original 256×256 sketch/photo of size 224×224 for data augmentation during training.

Evaluation Metrics We use the same evaluation metrics of recall $@K$ as in [19], where for one query sketch, recall $@K$ is 1 if the corresponding photo is within the top K retrieved results and 0 otherwise. We report $acc@K$ by averaging over all queries in the test set.

Competitors To our knowledge, only two works report results on the Sketchy dataset [14, 19]. However, the model in [14] is designed for category-level SBIR with different experiment settings to FG-SBIR, and the focus is on retrieval speed using hashing techniques rather than accuracy. This leaves the various models proposed in [19] as the main competitors. These include a heterogeneous GoogLeNet triplet model (Heter-GN-Tri), a heterogeneous GoogLeNet pairwise contrastive model (Heter-GN-Pair) and a heterogeneous AlexNet pairwise contrastive model (Heter-AN-Pair). The other competitor is the Siamese triplet ranking model in [27] (Sia-SN-Tri). Its base network is called Sketch-a-Net (SN) which is a modified version of AlexNet. It takes an additional preprocessing step to extract edgemaps from photos [32] in the hope that the domain gap is reduced. For fair comparison, we pretrain the model in stages exactly as described in [27] and use the stage-3 pretrained model to finetune on the Sketchy dataset (Sketch-a net) with the same classification and triplet ranking losses. The performance of humans on FG-SBIR is also reported in [19].

Comparisons against the state-of-the-art Our model is compared to the state-of-the-art alternatives as well as humans in Table 2. The following observations can be made: (i) Our discriminative-generative hybrid model significantly outperforms all compared models (13.04% improvement over the second best Heter-GN-Tri). (ii) It is now fairly close to the human performance (4.12% lower). (iii) Note that all three heterogeneous baselines in [19] took advantage of extensive within-domain pretraining. Our results suggest that it is not nec-

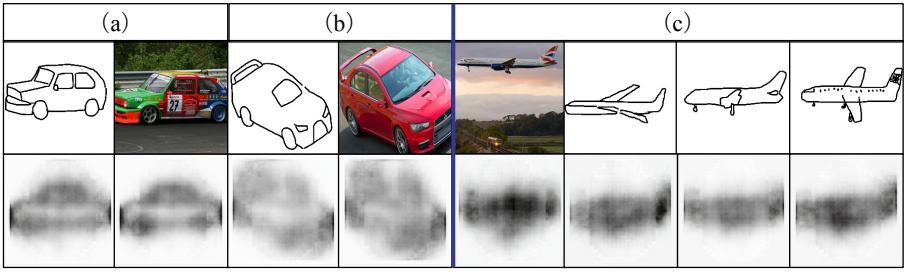

Figure 3: Examples of generated images on unseen test sets. In each sub-figure, top: input sketch/photo; bottom: corresponding generated images using the decoder.

Sia-SN-Tri [27]	Heter-AN-Pair [19]	Heter-GN-Pair [19]	Heter-GN-Tri [19]	Ours	Human [19]
16.17%	21.36%	27.36%	37.10%	**50.14%**	54.27%

Table 2: Comparative results against state-of-the-art FG-SBIR performance (*acc*@1).

essary with our Siamese hybrid network, significantly simplifying the training process. (iv) The poor result of Sia-SN-Tri [27] suggests that replacing natural photos with their edgemap has a negative side-effect given a challenging dataset such as Sketchy. Specifically, as shown in Figure 3, the photos in Sketchy often contain other objects and cluttered background. Removing colour information from the very beginning deprives the model of its ability to learn an implicit foreground-background segmentation mechanism to align photos with sketches that have clean background. Note that the objective of the generative decoder is not to synthesise sharp, visually appealing images. Instead, our goal is to reduce the domain gap and extract domain invariant and discriminative features – images in Figure 3, albeit blurry, are almost identical when a matching pair of photo and sketch are used as input respectively, showing that this goal has been achieved.

Ours-D	Ours	Our_GN-D	Ours_GN
47.18%	**50.14%**	45.52%	48.24%

Table 3: Evaluation of the contributions of the decoder and different basenets. (*acc*@1)

Ablation Study Our model differs from competitors in both the base network and the additional generative decoder. To find out what contributes to the superior performance of our model, we compare a few variants with and without the generative decoder and with different base network in Table 3, where _GN refers to replacing our VGGNet with GoogLeNet, -D means dropping the generative decoder part. The results show that (i) regardless of the choice of basenet, adding an additional generative decoder consistently improves the performance and (ii) compared with GoogLeNet, VGGNet is better for our problem.

Why Generative Learning Helps? To answer this question, let us first examine *what images would the decoder generate*. The decoder is designed to help the encoder preserve domain-invariant information. One thus would expect that given a pair of matching photo and sketch, the generated images would be very similar to each other with any domain discrepancies, such as lack of texture, lighting, occlusions and background information in the sketch domain, removed. Figure 3 shows that this is exactly what a trained model produces on a test set: (i) Despite the drastically different background clutter (Figure 3(a) and (c)) and

Figure 4: Qualitative results. For each query sketch, the top 10 ranked photos out of 1250 candidate photos in the gallery are shown in each row. Green boxes indicate the correct matches and when they are outside the top 10, their actual ranks are given.

Ours-Heter	Ours-Hybrid	Ours
41.52%	49.55%	**50.14%**

Table 4: Comparing different weight sharing strategies (*acc*@1)

occlusions (Figure 3(b)) exhibited in natural photos, the decoder is able to discard these irrelevant information and focus on the main visual structures. (ii) Given a matched sketch-photo pair, the synthesised images are almost identical; they clearly preserve the shared visual cues (i.e., pose, shape) and neglect the unshared ones such as background and other details ignored by the human sketcher (e.g., Figure 3(a), the digit 27 on the side door). (iii) Sketches drawn by different humans for a single photo often varied greatly in the level of abstraction. Figure 3(c) shows that our decoder normalises these variations making the retrieval task easier. We thus conclude that having the generative decoder encourages the learned feature representation in the joint embedding layer to focus on the cross-domain shared semantic visual cues rather than the domain-specific information. Importantly it directly reduces the domain gap and enables the learning of a richer representation useful for model generalisation.

Evaluations on Different Weight Sharing Strategies In this experiment, we compare our model with the three weight-sharing strategies described in Sec. 3.3. Table 4 shows that without any weight-sharing, the heterogeneous version of our model is the weakest in aligning the two domains, whist the partial-sharing strategy results in a slightly inferior performance. Since a Siamese network has much less parameters than the other two, this justifies the use of the Siamese architecture.

Qualitative Results Example retrieval results of the proposed model are shown in Figure 4. The results suggest that the model is very effective in removing other objects in the scene and cluttered background and is able to capture subtle instance-level differences, e.g. the first two rifles are matched correctly among some very similar-looking rifle instances.

Failure cases are those where true matches are ranked outside the top-10. Two of them are shown in the bottom of Figure 4. It is obvious that these failure cases are caused mainly by the poor quality of sketch drawing (e.g., too abstract) with critical details missing in the sketches, giving the model no chance to find the correct matches.

5 Conclusion

In this paper, for the first time, a hybrid discriminative-generative approach is proposed for fine-grained sketch-based image retrieval (FG-SBIR) based on a cross-domain deep encoder-decoder network architecture. The hypothesis was that by introducing the additional generative task, the learned joint embedding space would capture domain-invariant information and explicitly reduce the domain gap between photo and sketch. Extensive experiments validated the hypothesis and demonstrated that the proposed model outperforms existing discriminative models by a large margin.

References

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 2013.

[2] Yang Cao, Changhu Wang, Liqing Zhang, and Lei Zhang. Edgel index for large-scale sketch-based image search. In *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[3] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE transactions on visualization and computer graphics (TVCG)*, 2011.

[4] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems (NIPS)*, 2014.

[6] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 2006.

[7] Rui Hu and John Collomosse. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *Computer Vision and Image Understanding (CVIU)*, 2013.

[8] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[9] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[10] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[11] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] Yi Li, Timothy M. Hospedales, Yi-Zhe Song, and Shaogang Gong. Fine-grained sketch-based image retrieval by matching deformable part models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2014.

[13] Yen-Liang Lin, Cheng-Yu Huang, Hao-Jeng Wang, and Winston Hsu. 3d sub-query expansion for improving sketch-based multi-view image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013.

[14] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[16] Antti Rasmus, Harri Valpola, and Tapani Raiko. Lateral connections in denoising autoencoders support supervised learning. In *Proceedings of the 32rd International Conference on Machine Learning (ICML)*, 2015.

[17] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

[18] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[19] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: Learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (proceedings of SIGGRAPH)*, 2016.

[20] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[22] 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 *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

[23] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[24] Benigno Uria, Marc-Alexandre Côté, Karol Gregor, Iain Murray, and Hugo Larochelle. Neural autoregressive distribution estimation. *Journal of Machine Learning Research (JMLR)*, 2016.

[25] Fang Wang, Le Kang, and Yi Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[26] Donggeun Yoo, Namil Kim, Sunggyun Park, Anthony S Paek, and In So Kweon. Pixel-level domain transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[27] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen-Change Loy. Sketch me that shoe. In *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[28] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. Deconvolutional networks. In *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[29] Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao. Sketchnet: Sketch classification with web images. In *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[30] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[31] Junbo Zhao, Michael Mathieu, Ross Goroshin, and Yann LeCun. Stacked what-where auto-encoders. In *Proceedings of the International Conference on Learning Representations Workshop Track (ICLR Workshop Track)*, 2015.

[32] Larry Zitnick and Piotr Dollar. Edge boxes: Locating object proposals from edges. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.