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

Kaiyue Pang[1], Yi-Zhe Song[1], Tao Xiang[1], Timothy M. Hospedales[1,2]

[1] SketchX Research Lab, Queen Mary University of London, London, UK
[2]School of Informatics, University of Edinburgh, Edingurgh, UK

Introduction

Problem:

- **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.

Challenges:

- Domain gap: sketch and photo are two distinctive image domains.
- Abstraction gap: human sketching exhibits a varied level of abstraction and sophistication.
- Model learning: discriminative loss alone to address cross-domain fine-grained comparisons

Solution:

- Discriminative learning for fine-grained comparisons, generative learning for domain alignments.
- A discriminative-generative hybrid model by introducing a generative task of cross-domain image synthesis.

Generative Effects:

- Domains are explicitly aligned.
- Cross-domain shared visual cues are preserved.
- Human sketching variations are normalized.

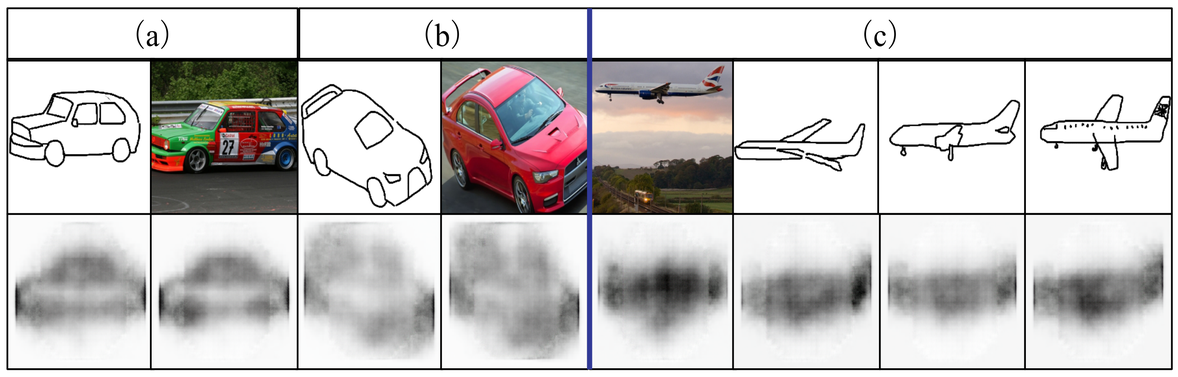

Figure 1: Examples of synthesized images via generative decoder on unseen test sets.

Methodology

Contribution 1: a cross-domain encoder-decoder FG-SBIR model which consists of four sub-networks:

- a three-branch Siamese encoder with a joint feature embedding space for cross-domain matching.
- a Siamese decoder subnet with ℓ_2 pixel-level reconstruction loss to explicitly align sketch and photo domains to a single sketch domain.
- a classification subnet with softmax cross-entropy loss to be category-discriminative.
- a triplet ranking subnet to be instance-discriminative.

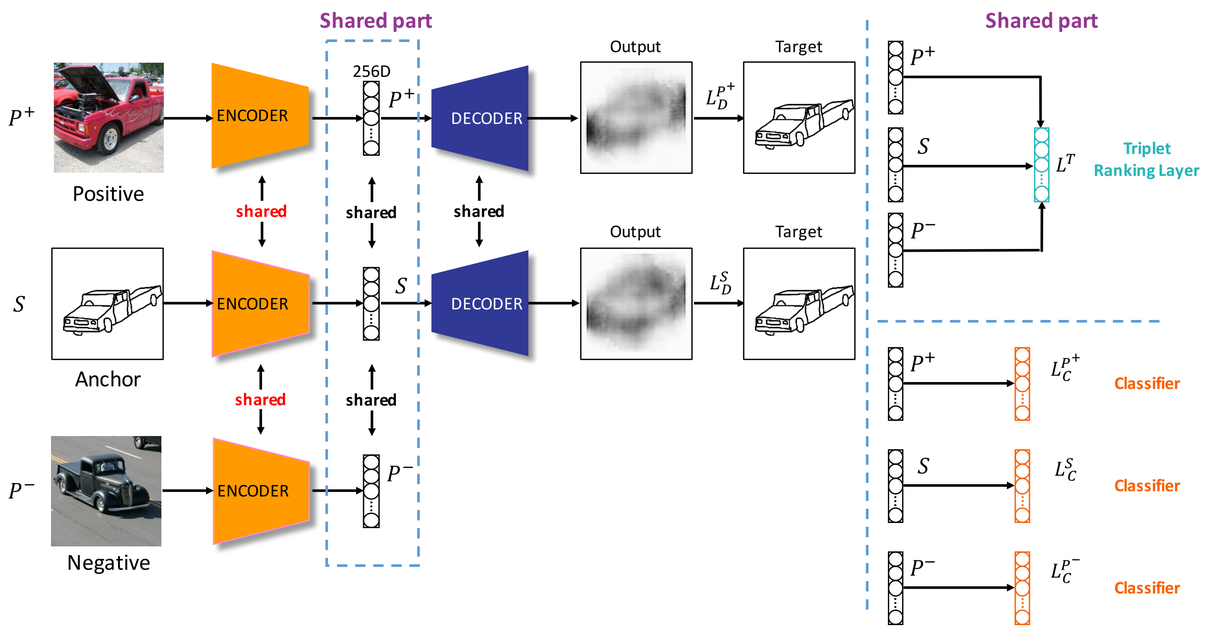

Figure 2: illustration of deep cross-domain encoder-decoder FG-SBIR model.

Contribution 2: a cross-domain generative decoder both architecturally designed and trained to favor the FG-SBIR task:

- Architecture: asymmetric with and shallower than the encoder.
- Training strategy: learning encoder first, fine-tuning with decoder.

Contribution 3: different weight sharing strategies for FG-SBIR task are explored in terms of base-net and final feature embedding layer.

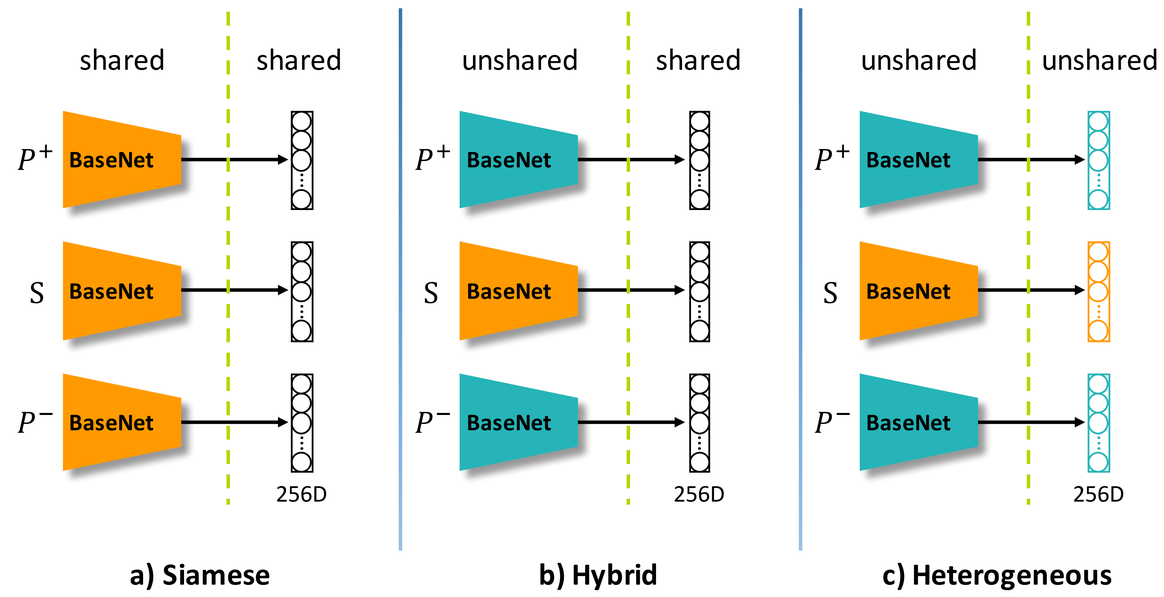

Figure 3: illustration of different weight sharing strategies for FG-SBIR task.

Results

Dataset:

- Largest FG-SBIR dataset to date, Sketchy, containing 125 categories with 100 photos per category and at least 5 human free-hand sketches for one photo.

Comparison against state-of-the-art:

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

Table 1: Comparative results against state-of-the-art FG-SBIR performance (*acc*@1). SN: Sketch-a-Net. AN: AlexNet. GN: GoogleNet.

Contributions of the generative decoder:

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

Table 2: Evaluation of the contributions of the decoder and different basenets. (*acc*@1) -D: without generative decoder. _GN: replace basenet with GoogleNet.

Comparisons of different weight sharing strategies:

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

Table 3: Comparing different weight sharing strategies (*acc*@1)

Qualitative results:

Figure 4: Qualitative results. Green boxes indicate the correct matches and when they are outside the top 10, their actual ranks are given.

Main References

- 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.
- 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.