Supplementary for Adaptive Fine-Grained Sketch-Based Image Retrieval

Ayan Kumar Bhunia[1] Aneeshan Sain[1,2] Parth Hiren Shah*
Animesh Gupta* Pinaki Nath Chowdhury[1,2] Tao Xiang[1,2] Yi-Zhe Song[1,2]

[1]SketchX, CVSSP, University of Surrey, United Kingdom.
[2]iFlyTek-Surrey Joint Research Centre on Artificial Intelligence.
{a.bhunia, a.sain, p.chowdhury, t.xiang, y.song}@surrey.ac.uk

1 Quantitative results on various *fine-grained* SBIR datasets

Table 8. Cross-Dataset adaptation results (Acc@1) based on category-level adaptation setup: Model trained on Sketchy [15] is directly used to evaluate on testing set of four different datasets (used by different FG-SBIR works, e.g. [19, 17, 20, 11, 10]), which is named as baseline (without adaptation). In contrast, our proposed adaptive FG-SBIR can utilise very few labelled sketch-photo pairs ($k \in \{5, 10\}$) through adaptation (one gradient update) step. For upper-limit, we use the complete training-set from the respective datasets for fully supervised-training.

	QMUL-Chair	QMUL-Shoe	QMUL-HandBags	QMUL-ShoeV2
Upper Limit	85.7%	43.2%	52.3%	33.7%
Baseline (Without Adaptation)	18.7%	11.8%	15.8%	10.3%
Ours (k=5)	56.1%	28.3%	33.6%	22.3%
Ours (k=10)	67.6%	34.9%	41.5%	26.4%

2 More details on experimental setup and analysis

(i) For our self-designed baselines, we use the same backbone with spatial attention for feature extraction. We will evaluate the effect of spatial attention on other published works in the future and include in supplementary.

(ii) MAML and ANIL are used to learn both F and M. However, while MAML involves updating both F and M in the inner-loop, ANIL only updates M in the inner-loop. Only classification loss is used as regularizer for both MAML and ANIL.

(iii) Margin is meta-learned only in our final proposed model as it is our own contribution.

(iv) We avoid comparing with Sain *et al.* [13], as it has an unfair advantage over our baseline because it compares the query sketch with every gallery image through paired-embedding. In other words, [13] incurs huge computation overhead by a multiple of at least the number of gallery images, thus we avoid it.

* Interned with SketchX

Although our meta-framework could be implemented on top of it, we avoided such design for faster retrieval and efficiency.

(v) Results reported for Pang *et al.* [9] are based on our own re-implementation, as the code is not publicly available.

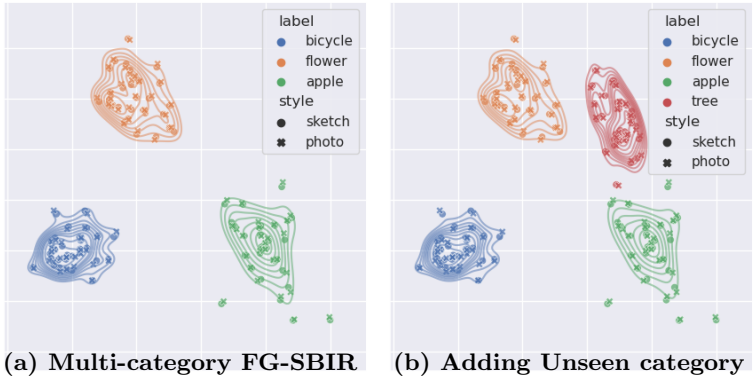
(vi) While the original MAML employs classification loss, we replace it with triplet loss and additionally add one classification head upon $F(\cdot)$ on the outer loop.

(vii) After investigation, we realised that the use of triplet-loss is very critical for FG-SBIR, and is a salient reason why a model such as Doodle2search [2] can yield competitive results as it incorporates triplet-loss while the rest zero-shot models [4, 18, 8] do not.

(viii) Please note that semantic relatedness module is used only for category-level adaptation on Sketchy; not for user style adaptation (on Shoe-V2), as no varying semantic concepts exist across different users.

(ix) In future, we believe a more recent gradient-based meta-learner (like iMAML [12]) could be promising to try out to improve one-shot setting. With reference to implicit differentiation for instance, iMAML [12] can easily tackle multiple gradient steps without the issue of vanishing gradients or memory constraints, while being independent of the inner loop optimiser chosen.

3 Significance of Classification and Semantic Relatedness Loss



(a) Multi-category FG-SBIR **(b) Adding Unseen category**

Fig. 9. (a): Figure illustrating a joint embedding space of sketch-photo pairs from "bicycle", "flower", and "apple" categories for **multi-category FG-SBIR.** (b) The semantic meaning of every category is utilised to ensure a better relative arrangement of class-specific groups in the embedding space. This helps for better adaption to unseen classes. The sketch-photo pairs of an unseen category "tree" (marked in red), lies close its semantically related seen categories "apple" and "flower", but far away from "bicycle". Note that class discriminative objective is still preserved in the embedding space, instead their relative placement helps towards better adaptation.

Classification loss While handling multiple categories (e.g., Sketchy) using a single FG-SBIR model, two design considerations need to be taken in the embedding space: (i) photo-sketch pairs to be separated on category-level (class discriminative), and (ii) every photo-sketch pair within a particular class must be separated from each other as well (instance discriminative). The second objective is handled by triplet loss, while the first is taken care by classification loss. In other words, triplet-loss is the main driving factor, while classification loss acts as a regularizer that helps to group class specific concept for better alignment in the embedding space.

Semantic relatedness loss For multi-category FG-SBIR (e.g., Sketchy), sketch-photo pairs from the same category group together using class discriminative objective. Furthermore, every category encapsulates a semantic concept, which could be used to control the way these class-specific groups are arranged in the embedding space, such that class-specific concept can be transferred from seen to unseen categories (akin to zero-shot SBIR [2, 4, 5]). The entire objective is handled by the semantic relatedness module. This is not to be confused with the instance-specific separation criteria for fine-grained retrieval, which is already handled by triplet-loss. In other words, location of class-specific group in semantic space is so adjusted that adaptation for unseen classes becomes easier.

Note that we do not use semantic-related module for style-adaptation on Shoe-V2 dataset, as no varying semantic concepts exist across different users. Note also that Furthermore, our semantic relatedness module design is generic to any fine-grained SBIR dataset having multiple categories.

4 Difference between our Multi-Category FG-SBIR and SOTA FG-SBIR

Most existing FG-SBIR works [1,9,10,19,16,14] deal with gallery images of a specific category. The problem we deal with, **multi-category FG-SBIR** (Sketchy), involves a single model handling instance-specific retrieval from multiple categories, which is less explored. One obvious choice is to include all the categories during training. However collecting fine-grained sketch-photo pairs for every class is not practical. Therefore, earlier attempts have been made through zero-shot SBIR [2] (at category-level retrieval though, not fine-grained), however, performance is significantly limited compared to its supervised counterpart.

Instead, we are after a single FG-SBIR model that can quickly adapt to a new style/category, with just a few samples during testing. Achieving this offers a best-of-both-worlds solution: (i) the model has a better chance at adaptation having observed new style/category data, as opposed to no data for generalisation or zero-shot, and (ii) the few samples requirement still falls within the practical remit of sketch data, i.e., one can always sketch just a few.

5 Summarize the Novelty

We introduced a novel single FG-SBIR model that can quickly adapt to **both** new style or category, with just a few samples during testing (few-shot style).

Furthermore, ours is not a direct trivial adaptation of MAML, but a design specific to the cross-modal problem in hand:

(a) We simplify the MAML training, by performing inner loop updates only for the final joint-feature embedding layer. This avoids model-overfitting during adaptation, as not all parameters are updated during adaptation.

(b) We *meta-learn* the *margin*-hyperparameter of triplet-loss inside our meta-training process, that would adaptively decide the optimal value for a specific category during inference.

(c) Three additional regularisation losses are introduced in the outer loop, to increase the meta-learned model's efficiency for category/style adaptation (§ 4.3).

6 Difference with classical few-shot learning (classification)

Standard few-shot literature usually deals with classification [6], whereas ours is the *first work employing few-shot adaptation for fine-grained retrieval*. We show potential under two objectives: category and user's style adaptation.

7 Why not other fine-grained problems (ID/re-ID) undergo such generalization issue as FG-SBIR

Domain gap existing across various categories in multi-category FG-SBIR, is much larger than different person-identities in Re-ID, as shape morphology varies highly across new categories (not limited to just human shapes). Note that FG-SBIR model tries to learn *shape correspondences* between sketches and photos. As *shape* itself becomes almost unknown for unseen categories, discovering fine-grained correspondence becomes even harder. Moreover, one single model handling fine-grained retrieval for multiple categories is limited by *model's representation potential* (Fig. 6).

8 Difference with Dou *et al.* [3]

Dou *et al.* [3] adopted MAML for *domain generalisation* purpose where triplet loss acts as an auxiliary loss to encourage class specific feature clustering. On the contrary, ours involves a few-shot *adaptation paradigm* which requires executing inner loop update using triplet-loss during inference. Therefore, the design of inner-loop update using triplet loss is more critical to our framework. Furthermore, unlike [3], margin value of inner-loop triplet loss is meta-learned to facilitate better and stable adaptation.

9 Difference with Hu *et al.* [7]

Hu *et al.* [7] introduced a few-shot sketch classifier, which however does not apply directly to SBIR as SBIR involves *cross-modal retrieval*, other than zxas classifier's weight generation. Nevertheless, the HyperNetworks used in [7] could be exploited in future to generate weights of the final feature embedding layer using sketch-photo pairs, but it is beyond the scope of this work.

10 Ensuring no leakage of class information from word-vectors

We use the pre-computed word vectors provided by the authors of Doodle2search [2], which already ensured no leakage of class information. The fair way of evaluation is to remove those word embeddings from the unseen or novel classes and recompute them and use those embeddings in the semantic relatedness model.

11 Clarification on relatively less improvement on Shoe-V2

We conjecture that the generalisation on unseen Sketchy classes is much more challenging, thus having more room for improvement. Furthermore, the user-specific subtle differences are more difficult to model compared to category-level modelling.

12 A much stronger upper-limit baseline on Sketchy

We did evaluate with a much stronger baseline: Train N different models for every unseen class. This gives an Acc@1 of 36.43% on Sketchy. We decided against including in the main paper as it seems unfair to compare with a baseline having N times the number of parameters and which requires several thousands of iterations to train each model.

13 Relevance of Adaptive FG-SBIR

We use the term "adaptive" from the application point of view, i.e., the model quickly "adapts" to new users/categories.

Note that our setting is beneficial in that: (a) A gain of nearly 6%(10%) over zero-shot using only 5(10)-shot setting. (b) Our adaptation requires *only one gradient* update, for *just* the final feature embedding layer M. Please note that model parameters are not updated for every sketch, rather only when adapting the model to a specific category or user.

14 Does our method perform End-to-end training?

We do train *end-to-end* using *bi-level*, inner (Eq 8) and outer-loop (Eq 9) based optimisation (§ 3.2).

15 Why do we use Sketchy and ShoeV2?

While we used Sketchy dataset for multi-category FG-SBIR under category-level adaptation, Shoe-V2 is used for *user specific* style adaptation. Note, Sketchy is *the only* dataset having multi-category fine-grained sketch-photo pairs, and only Shoe-V2 provides user specific sketches for style adaptation.

16 Is it possible to do cross-dataset adaptation (beyond Sketchy)?

Furthermore, we show the potential of cross-dataset adaptation where model trained from Sketchy is adapted using 5 or 10 samples to the completely different and unseen Shoe-V2 dataset – thus signifying the generalization potential of our method.

References

1. Bhunia, A.K., Yang, Y., Hospedales, T.M., Xiang, T., Song, Y.Z.: Sketch less for more: On-the-fly fine-grained sketch based image retrieval. In: CVPR (2020)
2. Dey, S., Riba, P., Dutta, A., Llados, J., Song, Y.Z.: Doodle to search: Practical zero-shot sketch-based image retrieval. In: CVPR (2019)
3. Dou, Q., de Castro, D.C., Kamnitsas, K., Glocker, B.: Domain generalization via model-agnostic learning of semantic features. In: NeurIPS (2019)
4. Dutta, A., Akata, Z.: Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In: CVPR (2019)
5. Dutta, A., Akata, Z.: Semantically tied paired cycle consistency for any-shot sketch-based image retrieval. IJCV (2020)
6. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML (2017)
7. Hu, C., Li, D., Song, Y.Z., Xiang, T., Hospedales, T.M.: Sketch-a-classifier: Sketch-based photo classifier generation. In: CVPR (2018)
8. Liu, Q., Xie, L., Wang, H., Yuille, A.: Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In: ICCV (2019)
9. Pang, K., Li, K., Yang, Y., Zhang, H., Hospedales, T.M., Xiang, T., Song, Y.Z.: Generalising fine-grained sketch-based image retrieval. In: CVPR (2019)
10. Pang, K., Yang, Y., Hospedales, T.M., Xiang, T., Song, Y.Z.: Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval. In: CVPR (2020)
11. Radenovic, F., Tolias, G., Chum, O.: Deep shape matching. In: ECCV (2018)
12. Rajeswaran, A., Finn, C., Kakade, S., Levine, S.: Meta-learning with implicit gradients. In: NeurIPS (2019)
13. Sain, A., Bhunia, A.K., Yang, Y., Xiang, T., Song, Y.Z.: Cross-modal hierarchical modelling forfine-grained sketch based image retrieval. In: BMVC (2020)
14. Sain, A., Bhunia, A.K., Yang, Y., Xiang, T., Song, Y.Z.: Stylemeup: Towards style-agnostic sketch-based image retrieval. In: CVPR (2021)
15. Sangkloy, P., Burnell, N., Ham, C., Hays, J.: The sketchy database: learning to retrieve badly drawn bunnies. ACM TOG (2016)
16. Song, J., Song, Y.Z., Xiang, T., Hospedales, T.M.: Fine-grained image retrieval: the text/sketch input dilemma. In: BMVC (2017)
17. Song, J., Yu, Q., Song, Y.Z., Xiang, T., Hospedales, T.M.: Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In: ICCV (2017)
18. Yelamarthi, S.K., Reddy, S.K., Mishra, A., Mittal, A.: A zero-shot framework for sketch based image retrieval. In: ECCV (2018)
19. Yu, Q., Liu, F., Song, Y.Z., Xiang, T., Hospedales, T.M., Loy, C.C.: Sketch me that shoe. In: CVPR (2016)
20. Zhang, J., Shen, F., Liu, L., Zhu, F., Yu, M., Shao, L., Tao Shen, H., Van Gool, L.: Generative domain-migration hashing for sketch-to-image retrieval. In: ECCV (2018)

17 Qualitative Retrieval Results for Category-Level Adaptation

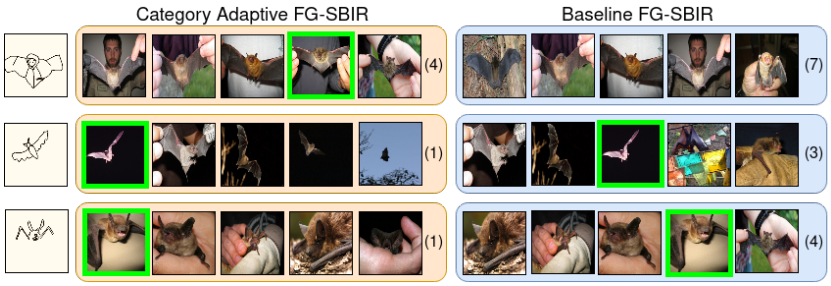


Fig. 10. Retrieval Result Comparison for "bat" category.

Fig. 11. Retrieval Result Comparison for "cabin" category.

Fig. 12. Retrieval Result Comparison for "cow" category.

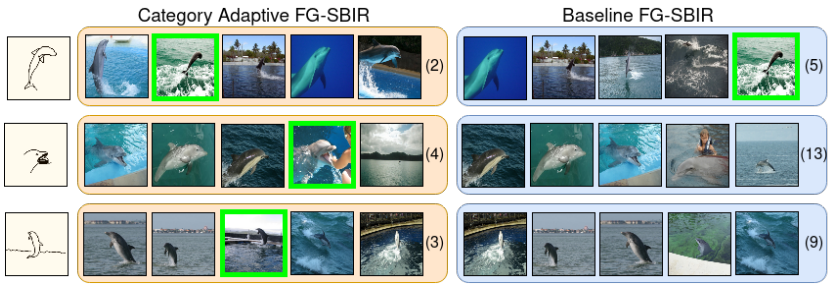


Fig. 13. Retrieval Result Comparison for "dolphin" category.

Fig. 14. Retrieval Result Comparison for "door" category.

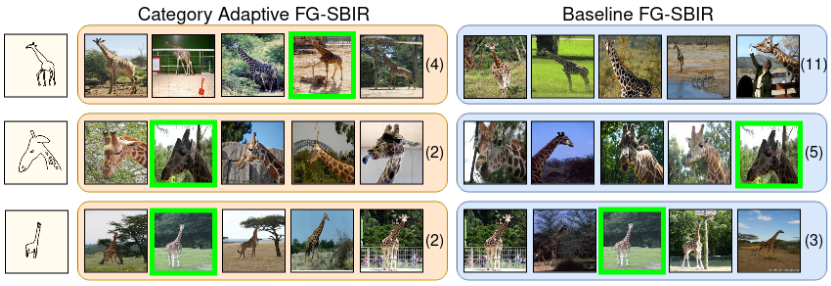


Fig. 15. Retrieval Result Comparison for "giraffe" category.

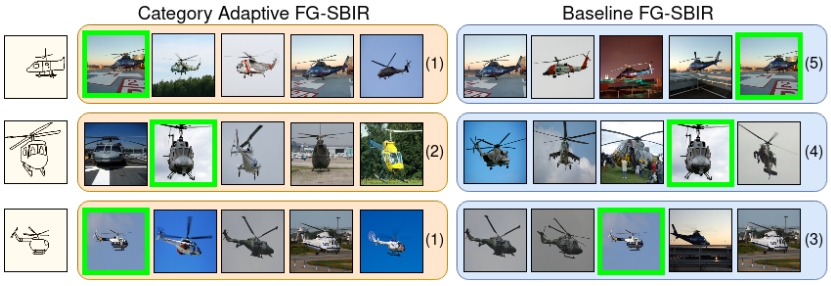


Fig. 16. Retrieval Result Comparison for "helicopter" category.

Fig. 17. Retrieval Result Comparison for "mouse" category.

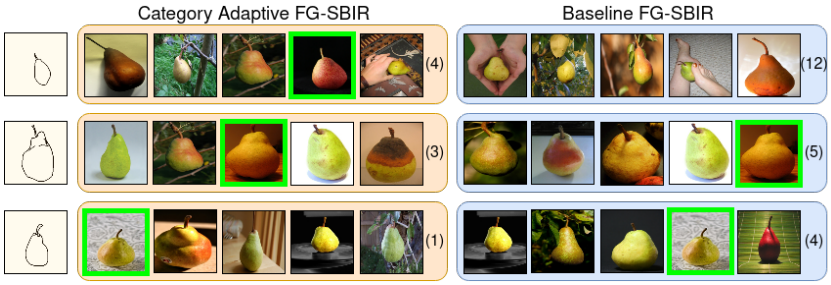


Fig. 18. Retrieval Result Comparison for "pear" category.

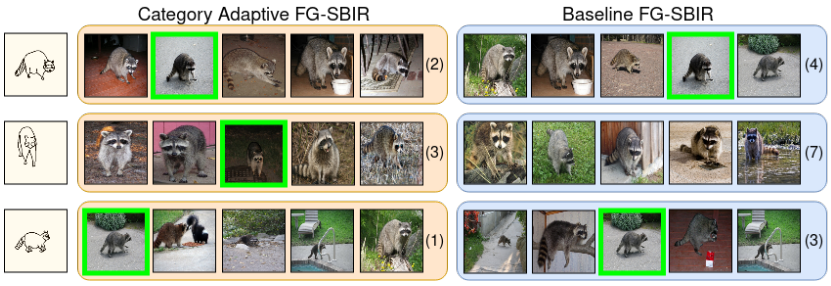


Fig. 19. Retrieval Result Comparison for "racoon" category.

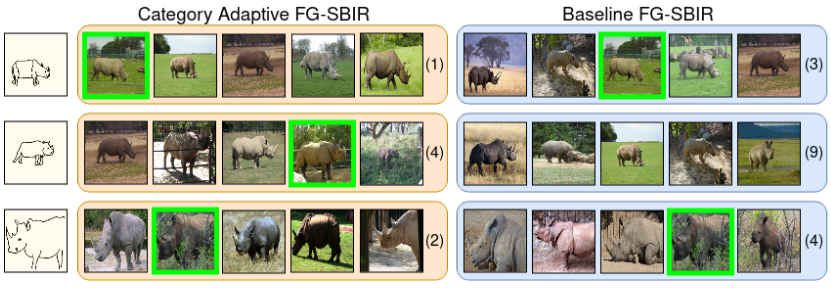


Fig. 20. Retrieval Result Comparison for "rhinocerous" category.

Fig. 21. Retrieval Result Comparison for "saw" category.

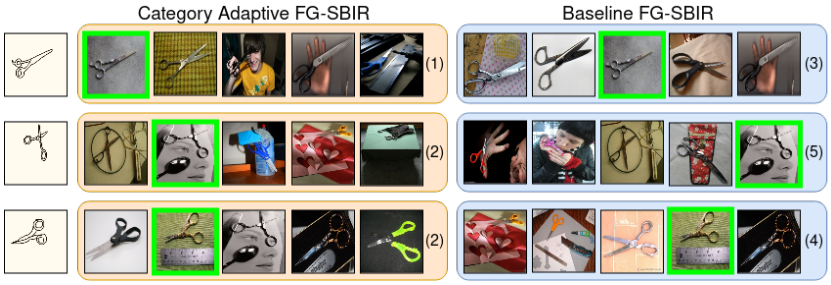


Fig. 22. Retrieval Result Comparison for "scissors" category.

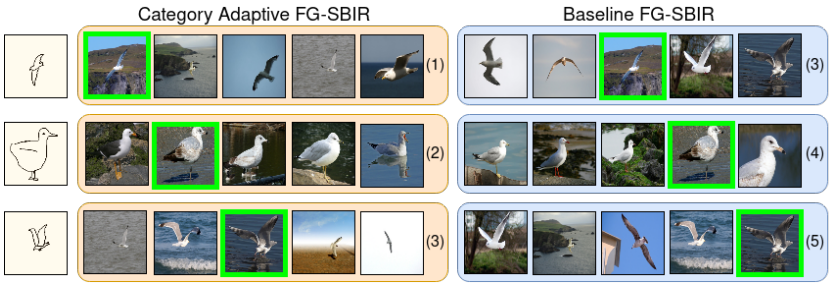


Fig. 23. Retrieval Result Comparison for "seagull" category.

Fig. 24. Retrieval Result Comparison for "skyscraper" category.

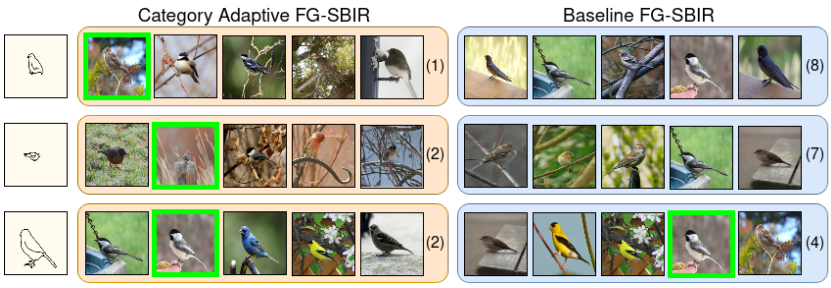


Fig. 25. Retrieval Result Comparison for "songbird" category.

Fig. 26. Retrieval Result Comparison for "sword" category.

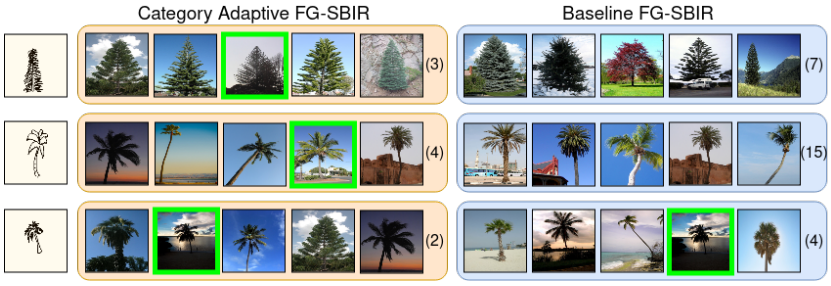


Fig. 27. Retrieval Result Comparison for "tree" category.

Fig. 28. Retrieval Result Comparison for "wheelchair" category.

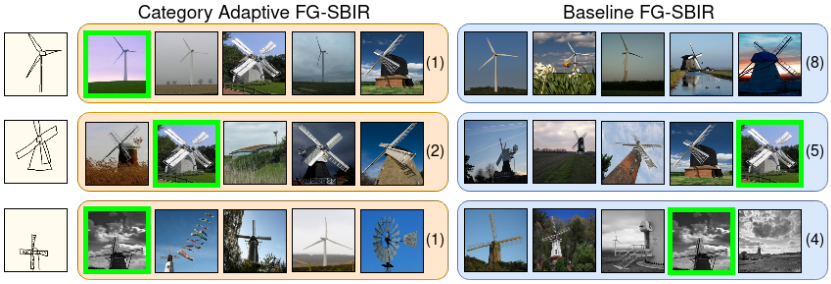


Fig. 29. Retrieval Result Comparison for "windmill" category.

Fig. 30. Retrieval Result Comparison for "window" category.

18 Qualitative Retrieval Results for User-Level Adaptation

Fig. 31. Retrieval Result Comparison among 3 different users for the same shoe instance.

Fig. 32. Retrieval Result Comparison among 3 different users for the same shoe instance.

Fig. 33. Retrieval Result Comparison among 3 different users for the same shoe instance.

Fig. 34. Retrieval Result Comparison among 3 different users for the same shoe instance.

Fig. 35. Retrieval Result Comparison among 3 different users for the same shoe instance.

Fig. 36. Retrieval Result Comparison among 3 different users for the same shoe instance.

Fig. 37. Retrieval Result Comparison among 3 different users for the same shoe instance.

Fig. 38. Retrieval Result Comparison among 3 different users for the same shoe instance.

Fig. 39. Retrieval Result Comparison among 3 different users for the same shoe instance.