A. Supplementary materials

In this supplementary material, we discuss and analyse more experimental evidences which corroborate the effectiveness of the proposed *Semantic Neighbourhood and Mixture Prediction Network* (SnMpNet). We start this discussion with the in-depth details of the two baseline methods developed for UCDR.

A.1. Implementation Details for UCDR Baseline Methods

As discussed in the main paper, we modified two recent algorithms from the field of Domain Generalization and Zero-shot Domain Generalization so that they can serve as the baseline retrieval models for UCDR. Both these baselines are subjected to the same train-time augmentations as SnMpNet.

EISNet-Retrieval. We replace the backbone network of EISNet [28] (originally ResNet50) with SE-ResNet50 (same as our SnMpNet). On top of the last layer of this backbone, we insert a 300-d linear layer, which is connected to the classification branch (the branch with the CE-loss in [28]) of the network. We further modify this CE-loss in [28], and train the model with the semantic-embedding similarity based CE-loss, described in equation (7). This modification in CE-loss accounts for the semantic information used in proposed SnMpNet (which was not present in [28]'s original architecture) and thus it is fair to compare the performance of this modified EISNet-retrieval model with proposed SnMpNet. We leave the other branches of EISNet-architecture [28] and the corresponding losses unchanged.

CuMix-Retrieval. Similar to the previous case, here also we use SE-ResNet50 as the backbone network. In addition, we leverage the image-level and feature-level cross-domain and cross-category mixing of samples, as introduced in [17]; and thus we refer to this model as CuMix-Retrieval. To account for category-discrimination of these mixed-up samples in the learned feature-space, we first apply the mix-up classification loss $\mathcal{L}_{CE}^{mix}$ (equation (8) in main paper) on samples mixed-up at the image-level. Towards that goal, we employ a 300-d linear layer on top of the last layer of the SE-ResNet50 backbone, so that the output of this 300-d layer can be used to compute the semantic similarities with class-embeddings, which are required for $\mathcal{L}_{CE}^{mix}$. To make use of the feature-level mixing, we generate $\mathbf{g}^* = \alpha\mathbf{g}_i^{c,d} + (1-\alpha)[\beta\mathbf{g}_j^{p,d} + (1-\beta)\mathbf{g}_k^{r,n}]$, at the output of the backbone network; and consequently pass it through the above-mentioned 300-d layer to obtain the final feature $\mathbf{f}^*$. We now re-compute $\mathcal{L}_{CE}^{mix}$ as $\mathcal{L}_{CE}^{mix-feat}$ with respect to these $\mathbf{f}^*$-features. Thus, the final loss becomes, $\mathcal{L}_{CuMix-Retrieval} = \mathcal{L}_{CE} + \omega_1\mathcal{L}_{CE}^{mix} + \omega_2\mathcal{L}_{CE}^{mix-feat}$, where $\mathcal{L}_{CE}$ is the standard cross-entropy loss computed for original (not generated through mix-up) samples, and ω_1 and ω_2 are two experimental hyper-parameters. Thus, this CuMix-Retrieval model now has access to the additional knowledge introduced by the class-name embeddings and the same backbone network as SnMpNet; and hence it is fair to compare its performance with SnMpNet for UCDR.

It can be observed that the final features of test samples (both query and search set data) obtained from these models are 300-dimensional. Thus, in absence of any existing methods in literature, these baseline models serve as the logical and fair competitors of proposed SnMpNet.

Now, we move on to the additional experimental analysis presented in the following section.

A.2. Experiments and Analysis contd.

In this section, we explore the robustness of the proposed model by analyzing it for different variations of protocols.

1) U^{c}CDR Evaluation on additional datasets. We discussed the performance of proposed SnMpNet for U^{c}CDR-protocol in Table 3, and compared it with existing ZS-SBIR methods in Table 1, ZS-SBIR being a special form of the same protocol, where sketch and real images are used as query and search-set domains, respectively. We extend this analysis further on two other challenging datasets: 1) Sketchy-extended [24], with split proposed in [26] - randomly chosen 25-classes are used as *unseen* test classes and out of the rest 100-classes, 90 are used for training and 10 used for validation; and 2) TU-Berlin[3], with split followed in [26][5] etc. - out of total 250-classes, randomly chosen 30-classes (with at least 400-images per category) are kept for testing, and 220-classes (200 for training and 20 for validation) are used as *seen*-classes.

The results are summarized in Table 6. We use mAP@all and Prec@100 as the evaluation metrics. For SEM-PCYC [5] and StyleGuide [7], we compare with the reported results in the respective papers. On the other hand, we use the implementation provided by the authors of [15], but unfortunately were unable to reproduce their reported numbers. Thus, we present

[3]M. Eitz, J. Hayes and M. Alexa, How do humans sketch objects?, *SIGGRAPH*, 2012

	Method	Backbone Network	output dim.	TU-Berlin extended		Sketchy extended	
				mAP@all	Prec@100	mAP@all	Prec@100
Existing SOTA	SEM-PCYC [5] (CVPR'19)	VGG-16	64	0.297	0.426	0.349	0.463
	Style-guide [7] (TMM'20)	VGG-16	200	0.2543	0.3551	0.3756	0.4842
	SAKE-512 [15] (ICCV'19)	SE-ResNet50	512	0.475	0.599	0.547	0.692
	SAKE-512 (our evaluation)	SE-ResNet50	512	0.3468	0.5225	0.4690	0.6665
SAKE-Variants	SAKE-512-w/o Label	SE-ResNet50	512	0.3314	0.5052	0.3599	0.5216
	SAKE-300-w/o Label	SE-ResNet50	300	0.3233	0.4959	0.3312	0.4841
SnMpNet		SE-ResNet50	300	**0.3568**	**0.5226**	**0.4412**	**0.5887**

Table 6: Peformance comparison for ZS-SBIR on Sketchy extended and TU-Berlin.

the retrieval accuracies obtained by us as SAKE (our evaluation) in Table 6, as well as the originally reported mAP-values for reference. Out of the SOTA-methods listed in the table, owing to the nature of the algorithms, it is feasible to create domain-independent variants only for SAKE [15]. Thus, similar to Table 1 in the main paper, here also we evaluate the two SAKE-variants for comparison. Following the same pattern as before, we observe that SAKE's performance deteriorates for both of these selected datasets when the domain-indicator is removed. Our SnMpNet outperforms both of these variants, as well as SEM-PCYC [5] and Style-guide [7] for both the datasets. Additionally, it has a superior performance even over the original SAKE-model with the domain indicator, on TU-Berlin, further validating its suitability for the UCDR protocol.

2) U^{d}CDR-Evaluation for QuickDraw. Previously in Table 4 in the main paper, we have evaluated and compared SnMpNet with the two retrieval baselines, EISNet-retrieval and CuMix-retrieval, on the U^dCDR protocol using *Sketch* as the unseen domain. For a more exhaustive analysis and completeness, we now repeat the same experiment with Quickdraw as the unseen domain. We construct the query set here with 10% of available samples, selected randomly, from each of the seen classes in QuickDraw. The search set again contains the *seen*-class RGB images as before. We summarize the retrieval

Method	mAP@200	Prec@200
EISNet-retrieval	0.0637	0.0309
CuMix-retrieval	0.0648	0.0298
SnMpNet	**0.1077**	**0.0509**

Table 7: U^{d}CDR-evaluation on DomainNet for *unseen* QuickDraw query domain. The search set contains only seen class real images. The models are trained on 5 domains - *Real*, *Sketch*, *Infograph*, *Painting* and *Clip-art*.

results in Table 7 and observe that SnMpNet outperforms the two baselines here as well.

3) Effect of Multi-domain Training data. Here we explore the effect of using training data from more than 2 domains to address the cross-domain retrieval. Towards this end, we train SnMpNet using only 2-domains (M=2) (*Sketch*/*QuickDraw* and *Real*) and observe the retrieval performance for UCDR and U^{c}CDR protocols. The results are summarized in Table 8 for two configurations of the search set - a) when only *unseen* classes are present in the search set and b) the search set contains samples from both *seen* and *unseen* classes. We also mention SnMpNet's performance using all 5-training domains (a common practice in DG) in Table 8 for ease of comparison. Importantly, we observe that the performance boost on using the three auxiliary domains is more pronounced for UCDR than U^{c}CDR. Thus, we can infer that the information from auxiliary training domains enhances the model's generalization abilities, especially for a new *unseen* domain.

4) Effect of weighting parameter κ. In the main paper, we propose a novel semantic neighbourhood loss $\mathcal{L}_{Sn}$, which implements a strict-to-relaxed weighting scheme based on how closely another class is related to the class of the current sample. It can be observed from equation (2), that the most important hyper-parameter we use in the training process of SnMpNet is κ, which effectively controls this above-said weighting. In Figure 4, we perform a simple experiment to observe the effect of this hyper-parameter κ in proposed $\mathcal{L}_{Sn}$. To this end, we vary κ over a range $0 \leq \kappa \leq 4$ and observe the retrieval performance (mAP@200) on the validation set data, of a primitive variant of *SnMpNet* - Base N/W + $\mathcal{L}_{Sn}$ (Table 5 in the main paper). We again perform this experiment for ZS-SBIR protocol on the Sketchy-extended dataset [32].

Protocol	Query Domain	Training Domains	*Unseen*-class Search Set		*Seen+Unseen*-class Search Set	
			mAP@200	Prec@200	mAP@200	Prec@200
$UCDR$	*QuickDraw*	*Sketch*, *Real*	0.1540	0.1138	0.1332	0.0972
		Sketch, *Real*, *Infograph*, *Painting*, *Clip-art*	0.1736	0.1284	0.1512	0.1111
	Sketch	*QuickDraw*, *Real*	0.2490	0.1953	0.2188	0.1741
		QuickDraw, *Real*, *Infograph*, *Painting*, *Clip-art*	0.3007	0.2432	0.2624	0.2134
U^cCDR	*Sketch*	*Sketch*, *Real*	0.4163	0.3455	0.3696	0.3066
		Sketch, *Real*, *Infograph*, *Painting*, *Clip-art*	0.4221	0.3496	0.3767	0.3109
	QuickDraw	*QuickDraw*, *Real*	0.2763	0.2181	0.2215	0.1832
		QuickDraw, *Real*, *Infograph*, *Painting*, *Clip-art*	0.2888	0.2314	0.2366	0.1918

Table 8: Effect of training data from multiple-domains on Retrieval Performance of *SnMpNet*.

From equation (2) in the main paper, it follows that $\kappa = 0$ corresponds to having equal weights for all classes, i.e. on all elements of the difference, $||\mathbb{D}(\mathbf{f}_i^{c,d}) - \mathbb{D}_{gt}(\mathbf{f}_i^{c,d})||^2$. For $\kappa \neq 0$, $\mathbf{w}(c)_c = 1$ and $\mathbf{w}(c)_k = e^{-\kappa}$, where k is the index of the most dissimilar class of c, in terms of their semantic distance $D(\mathbf{a}^c, \mathbf{a}^k)$. For any other class pair (c, j) $(j \neq k)$, $e^{-\kappa} < \mathbf{w}(c)_j < 1$, thereby enforcing the strict-to-relaxed criterion for preserving relative distances between classes.

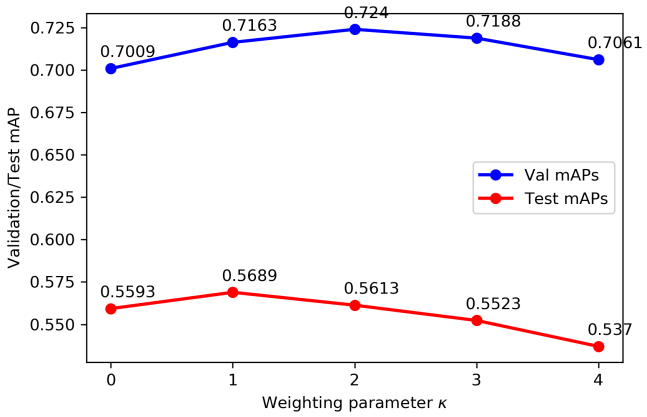


Figure 4: Effect of κ on validation and test mAP@200 for Sketchy extended split [32].

As shown before in Table 5, we select the model corresponding to $\kappa = 2$ for testing based on its highest retrieval accuracy on the validation set. For completeness, the variation of test set retrieval performance with κ is also plotted in the same figure.

A.3. Qualitative Analysis contd.

Here we further analyse SnMpNet to observe the feature-space and retrieved samples in details.

1) Visualization of the feature-space. We further visualize the learned feature-space through SnMpNet using a t-SNE [3] plot. Towards that goal, we train SnMpNet with samples from seen classes, which belong to the domains - Real, Quickdraw, Clip-art, Painting and Infograph. For visualization, we pick the same 10-categories as in Figure 3. We

project features from these selected seen and unseen classes onto the feature-space demonstrated in Figure 5. Moreover, we project features from both Quickdraw (seen domain, protocol U^{c}CDR in figure) and Sketch (unseen domain, protocol UCDR in figure). For comparative analysis, we also simulate the feature-space using the CuMix-retrieval algorithm and present the visualizations in the same figure. We can clearly observe better categorical distinction using SnMpNet over its close-competitor baseline CuMix-Retrieval. Although CuMix-Retrieval is able to place the 4 *unseen*-classes in the neighborhood of the 4 related *seen*-classes, it is not able to separate the object classes into clear and well-separated clusters for 5 out of the 10 classes. Only *van*, *ambulance*, *cow*, *laptop*, and *shark* have easily distinguishable cluster maps. SnMpNet,

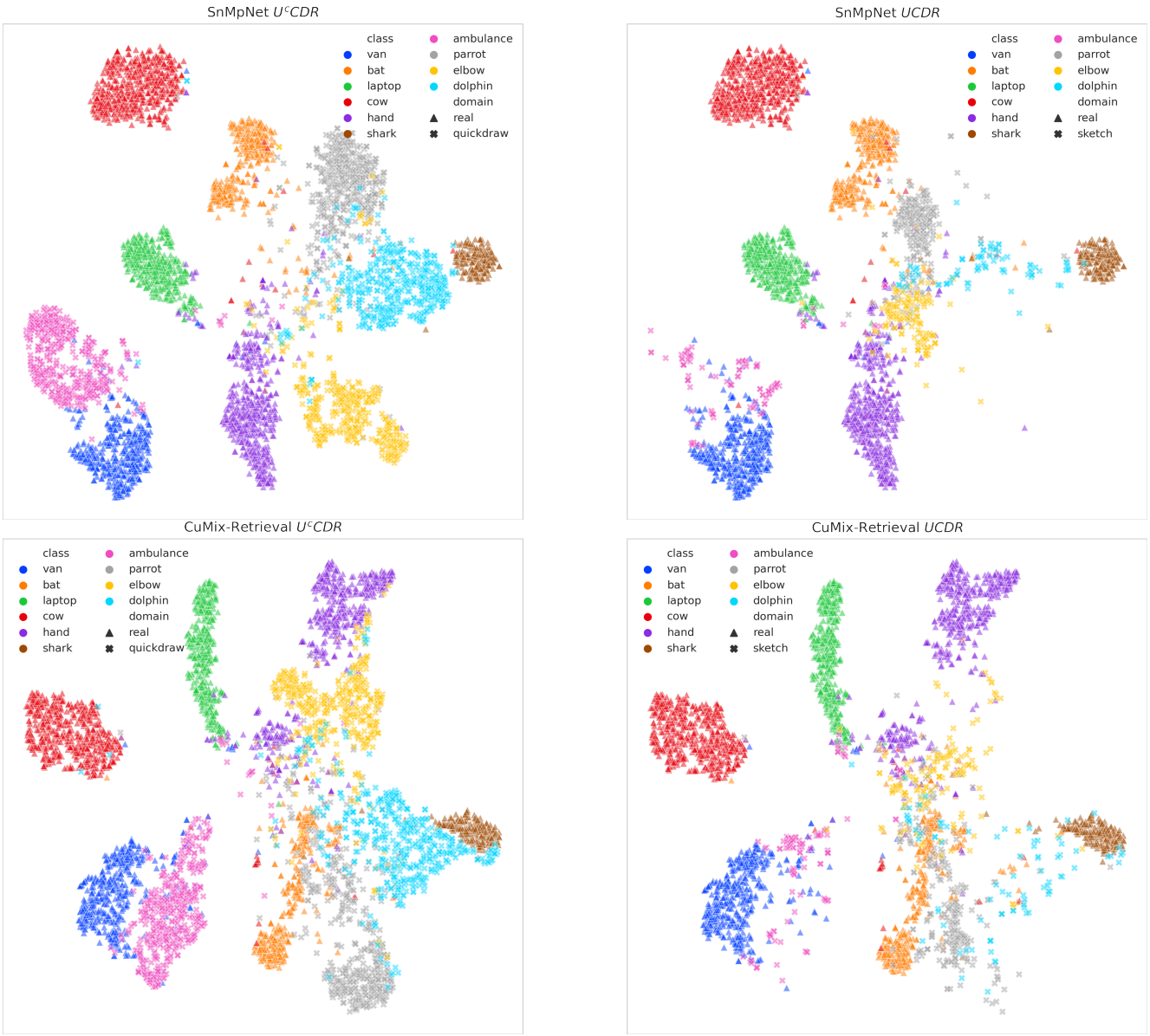


Figure 5: t-SNE plots for UCDR and U^{c}CDR protocols with CuMix-Retrieval and proposed SnMpNet. Here *Sketch* is unseen to the models, while *QuickDraw* and *Real* are seen. (best viewed in color)

on the other hand, succeeds at both the above objectives and creates near-distinct cluster boundaries for all 10 object classes.

2) Sample Retrieval Results. In continuation to Figure 2 in the main paper, we provide some additional retrieval results in Figure 6. In contrast to Figure 2, here we use *Sketch* as the unseen domain for UCDR, and report the top-8

retrieved images against each query for two configurations of the search set, as before. We have also presented sample retrieved images for a few randomly selected sample queries from a seen domain, Quickdraw, for complete analysis of this instance of trained SnMpNet. Thus these results correspond to the protocol U^{c}CDR. We observe the similar *mistake*-pattern as in Figure 2, i.e. the model struggles when the source *queries* (*sketch* or *quickdraw*) are ambiguous, poorly drawn or lack sufficient detail pertaining to the category. This results in retrieval of images from categories, which are either semantically related, or unrelated, but have high shape resemblance with the query. This is observed in several examples: *axe* being retrieved for *sailboat*; *finger*, *skateboard*, *tornado* for *asparagus*; *lightning*, *megaphone* for *axe*; *windmill*, *campfire* for *helicopter*; *cake*, *suitcase*, *ladder* for *skyscraper* etc. In the more challenging search set configuration containing both *seen* and *unseen* classes, incorrect retrievals are from more visually similar or closely related categories. For example, *bird*, *duck*, *swan*, *penguin* are retrieved for *parrot*; *zebra*, *horse* for *giraffe*; *monkey* for *octopus*; *door* for *skyscraper* etc.

Thus, we have analyzed the proposed model SnMpNet for a wide variation of experimental protocols. We have also explained the hyper-parameters associated with our model for better understanding. Now, we finally conclude with a brief discussion on how the proposed $UCDR$ protocol is different from other related works in literature.

A.4. Difference between proposed UCDR and other related works

Here, we discuss the differences between SnMpNet and [27]. A open cross-domain visual search protocol has been proposed in [27], which is significantly different from the traditional cross-domain data retrieval, which addresses the problem of retrieving data from one fixed target domain, and relevant to the query from another fixed source domain. This newly proposed protocol is again significantly different from our UCDR. We summarize the differences here.

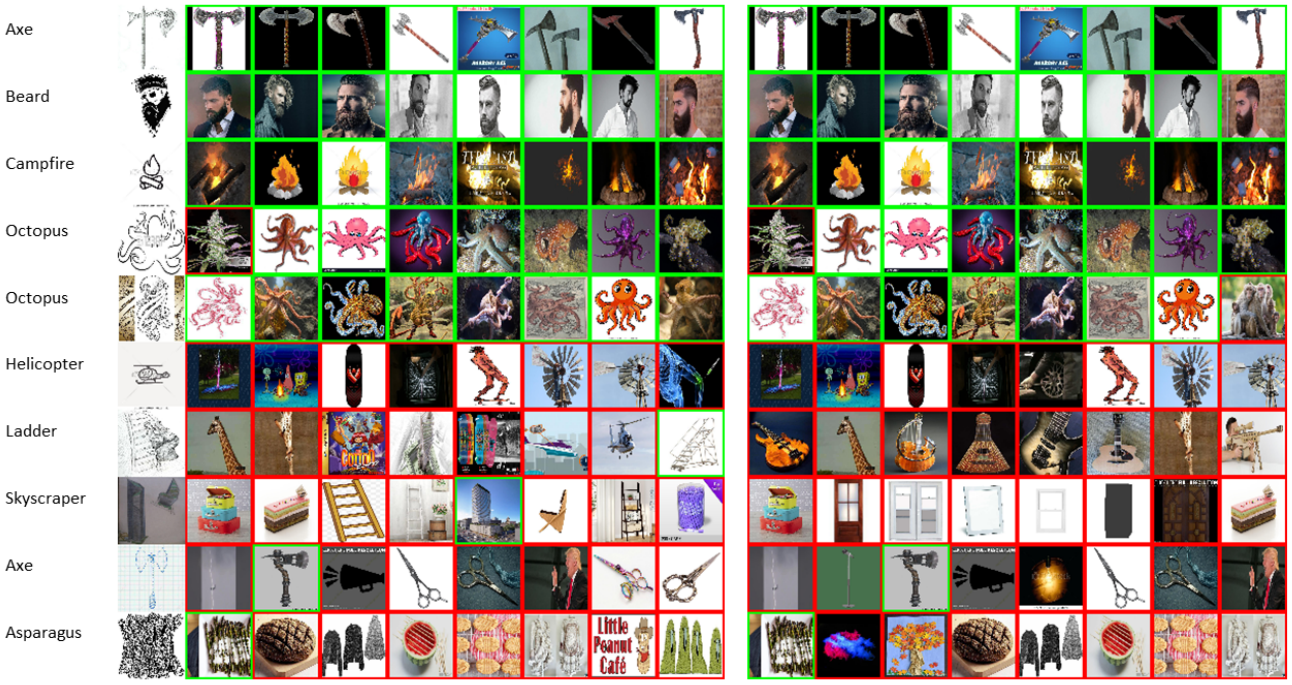
1. Open cross-domain visual search [27] proposes the cross-domain search among any two domains, provided they have been used during training. In contrast, proposed UCDR focuses on cross-domain retrieval scenario, when the query domain is not *seen* during training.

2. Proposed SnMpNet uses multiple domains of data (more than two) for training, as in [27]; however SnMpNet processes all domains through one single network (feature-extractor + classifier), instead of the separate domain-specific prototypical networks that learn a common semantic space in [27]. This results in significant decrease in the number of trainable parameters and model complexity.

3. [27] requires separate learning of a new semantic mapping function, whenever a new source / target domain emerges. Since the proposed model in [27] requires a-priori knowledge about the query and target domains, it cannot be used for UCDR protocol, without additional training. In contrast, SnMpNet can be seamlessly extended to the proposed multi-domain query / target conditions, proposed in [27].

It can be noticed that the focus of our work is more towards the generalization ability of the network for *unseen* classes and *unseen* domains, whereas [27] works towards generalizing retrieval in case of any query-search set pairs from *seen*-domains. Thus our work is significantly different from [27].

(a) $U^{c}CDR$ for QuickDraw



(b) $UCDR$ for Sketch

Figure 6: Top-8 Retrieved Images for UCDR and U^{c}CDR protocols on DomainNet with Sketch being the unseen query domain. Same query is considered for both the search set configurations. *Green* and *Red* borders indicate correct and incorrect retrievals respectively. (best viewed in color)