Fine-grained Foreground Retrieval Via Teacher-Student Learning Supplementary Material

Anonymous WACV submission

Paper ID 186

7. Additional Quantitative Results

In our comparison with selective convolutional descriptor (SD) aggregation (Tables 1 and 2 in the main paper), we use the SD-FG setting, where the original foreground object is available to the SD approach, which is a much easier setting than that of our fine-grained foreground retrieval setting. Below we report a comparison with the more realistic setting, which attempts to retrieve foreground given only the background (SD-BG). Since SD-BG can only select among foreground candidates with known background, we are only able to use a subset of our full evaluation set for this comparison. For each reference background, there are 13 foregrounds to choose from (instead of 30), and each foreground has known background. While the comparison of our method with SD-FG indicates comparable performance (despite the easier setting of SD-FG), the comparison below to SD-BG shows that our method performs much better. Note that since Tables 1 and 2 in the main paper use the full evaluation set, the numbers reported there are not comparable with the ones reported below.

	Bed	Chair	Couch	Overall
SD: SD-BG	0.394	0.755	0.397	0.516
Ours: full	**0.790**	**0.804**	**0.656**	**0.750**

	Bed	Chair	Couch	Overall
SD: SD-BG	0.815	0.842	0.837	0.831
Ours: full	**0.938**	**0.854**	**0.906**	**0.900**

Table 3: Comparison with SD-BG. Left: in terms of Mean Average Precision, using a score threshold of 2; Right: in terms of Normalized Discounted Cumulative Gain, $k = 5$.

Next, we examine the effect of ablating each of the features used in our approach (shape, azimuth, elevation, style mean, and style std) on the full evaluation set. Overall, our full model perform best in term of nDCG and very close to the best in terms of mAP.

	Bed	Chair	Couch	Overall
Ours: without shape	0.470	**0.601**	0.381	0.484
Ours: without azimuth	0.506	0.567	0.351	0.475
Ours: without elevation	0.539	0.544	0.383	0.489
Ours: without style mean	0.516	0.522	0.425	0.488
Ours: without style std	**0.563**	0.523	**0.436**	**0.507**
Ours: full	0.542	0.574	0.384	0.500

	Bed	Chair	Couch	Overall
Ours: without shape	0.846	**0.781**	0.764	0.797
Ours: without azimuth	0.812	0.768	0.721	0.767
Ours: without elevation	0.848	0.767	0.764	0.793
Ours: without style mean	0.816	0.741	0.769	0.775
Ours: without style std	0.837	0.766	**0.773**	0.792
Ours: full	**0.848**	0.777	0.764	**0.797**

Table 4: Ablation study for individual features. Left: in terms of Mean Average Precision, using a score threshold of 2; Right: in terms of Normalized Discounted Cumulative Gain, $k = 5$.

8. Network Parameters

The mask is of size (192, 256, 1), with 1 in the mask area and 0 otherwise. It is input to the mask branch to get mask embedding. Then mask embedding and image embedding are merged by merge branch to get full embedding. The full embedding is passed to each classifier branches to get each feature embedding. The mask branch is composed of convolution with parameters in Table 5. The merge branch Table 6 is one layer convolution. Each classifier branch Table 7 is a three layer fully connected network, with number of neuron n and m in Table 10.

Type	Configuration
conv2d	c:20, k:3 x 3, s: 2, p: same
batchnorm	
linear	
conv2d	c:40, k:3 x 3, s: 2, p: same
batchnorm	
relu	
conv2d	c:80, k:3 x 3, s: 2, p: same
batchnorm	
relu	
conv2d	c:160, k:3 x 3, s: 2, p: same
batchnorm	
relu	
conv2d	c:320, k:3 x 3, s: 2, p: same
batchnorm	
relu	

Table 5: Parameters for the mask branch. c, k, s and p stand for number of channel, kernel size, stride, padding type.

Type	Configuration
conv2d	c:512, k:3 x 3, s: 2, p: same
global average pooling	
fc	n:200
relu	
batchnorm	

Table 6: Parameters for the merge branch. c, k, s, p, n stand for number of channel, kernel size, stride, padding type and number of neurons.

Type	shape	azimuth	elevation	style mean	style std
fc	200	24	12	400	400
mid activation					
batchnorm					
fc	200	24	12	400	400
mid activation					
batchnorm					
fc	200	2	12	512	512
out activation					

Table 7: Structure of the classifier branches.

9. Saliency masks

Similarly to Figure 2, we show several additional examples of saliency masks, compared to instance segmentation masks.

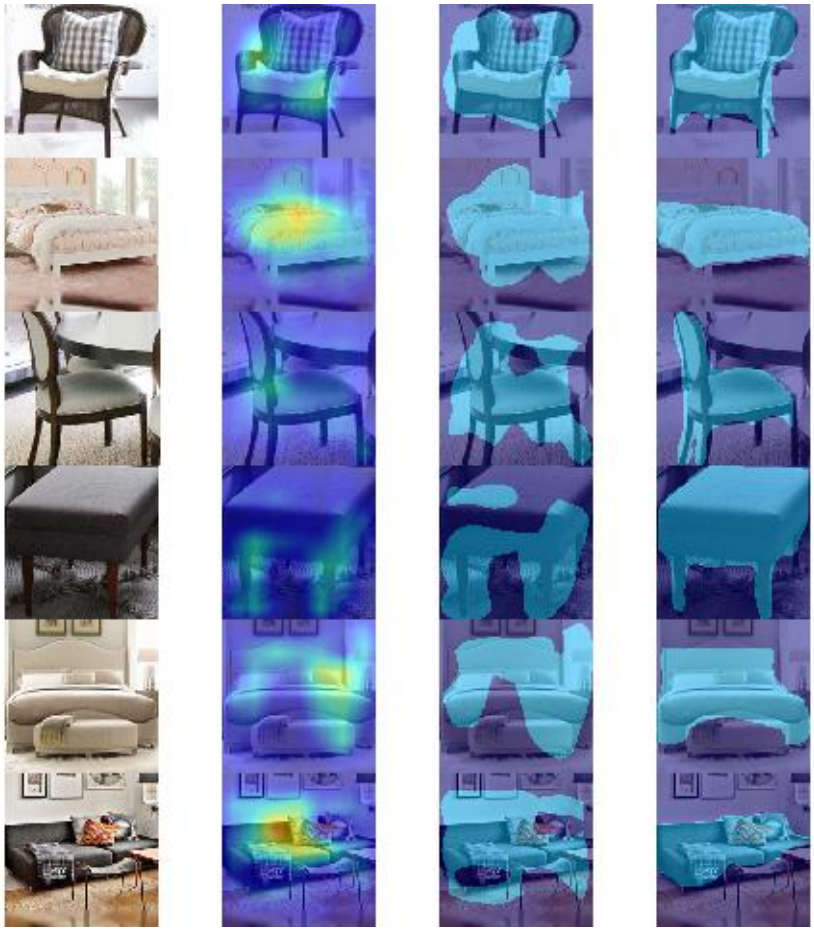

Figure 5: Additional saliency mask examples. From left to right, original image, saliency map, saliency mask, instance segmentation.

10. Dataset Statistics

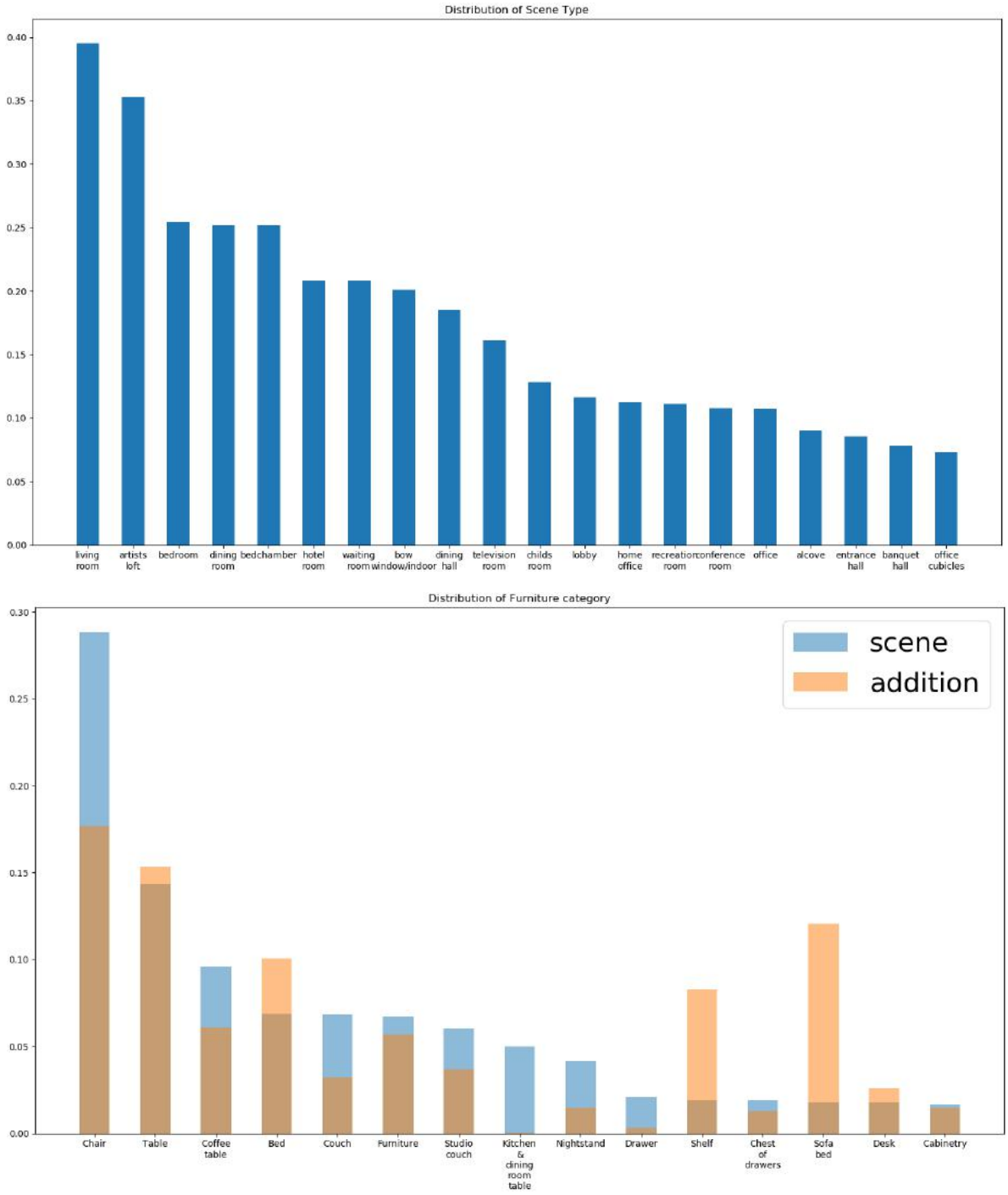

Figure 6: Top: Distribution of top 20 scene types among the 22K scene images. We can see that all of them are indoor scene types. Bottom: Distribution of top 15 furniture categories among the 31K extracted and filtered bounding boxes. These foreground are extracted from scene images with bounding box annotation. And the same 15 furniture categories in 40K additional foreground images. We can see that the dataset covers a wide range of furniture categories.

11. More Qualitative Results

Figures 7 through 15 show additional retrieval results for different categories. Each figure shows the query background scene on the left (with the target location indicated by a black rectangle). Each of the rows on the right shows the top 10 retrieved foreground objects using only one of the extracted features, while the bottom row shows the top 10 retrievals when using a weighted combination of these features.

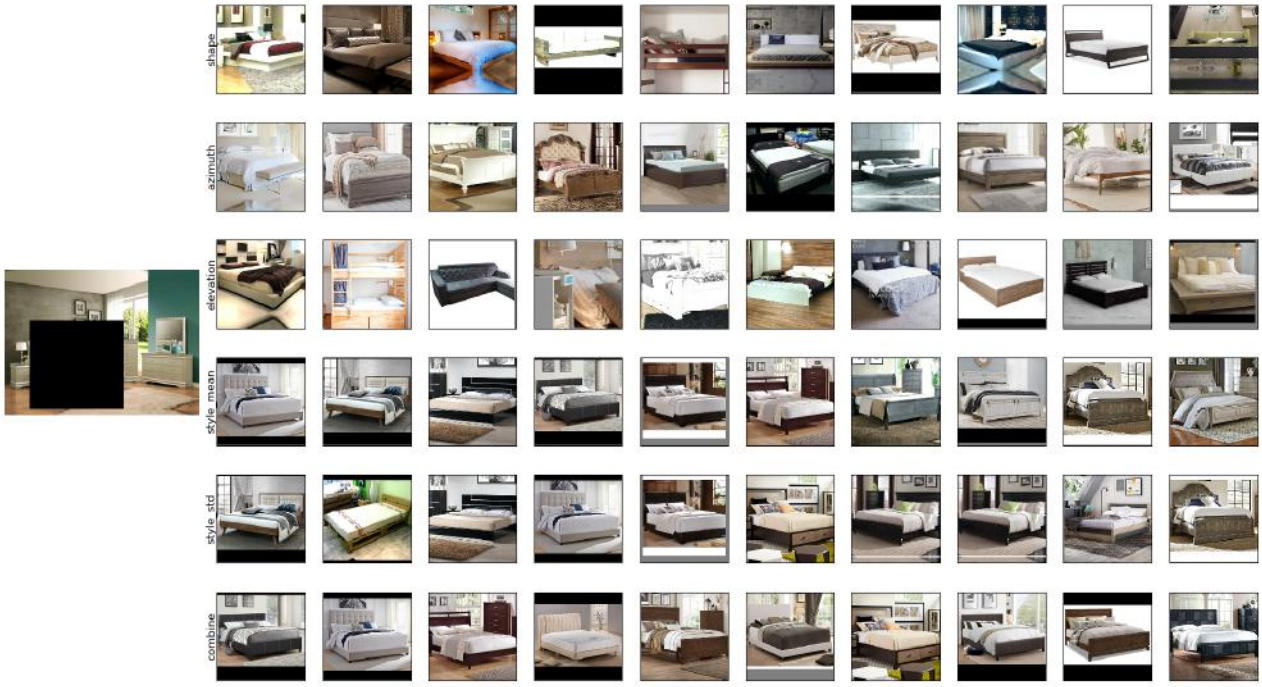

Figure 7: Retrieval for 'Bed' (1)

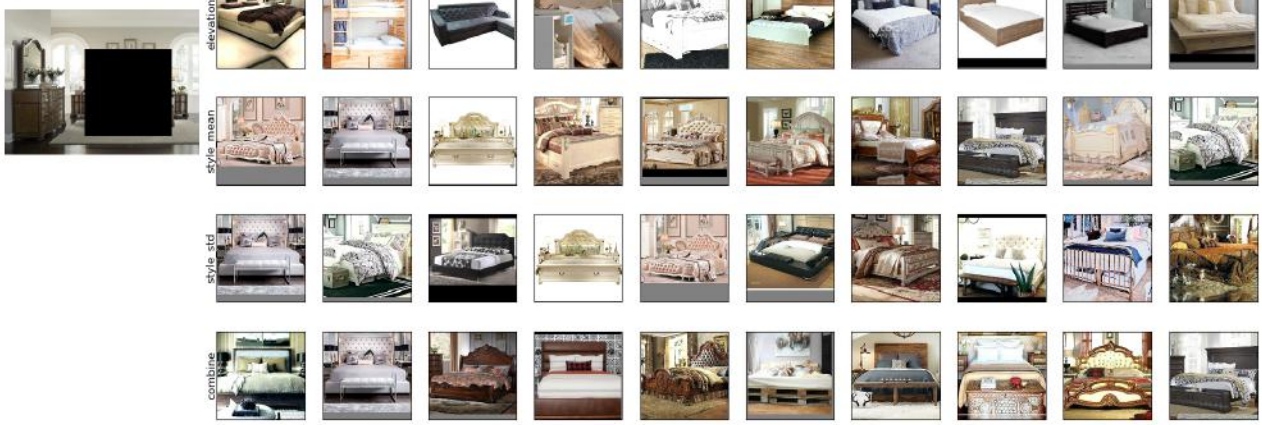

Figure 8: Retrieval for 'Bed' (2)

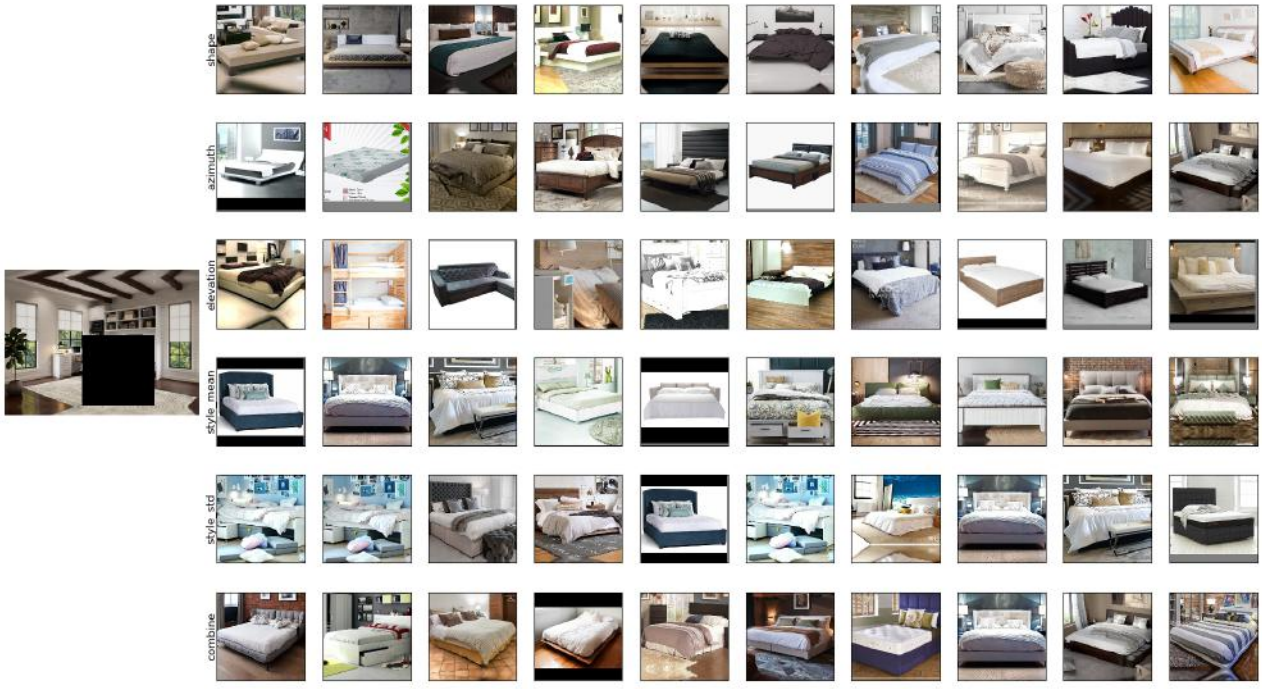

Figure 9: Retrieval for 'Bed' (3)

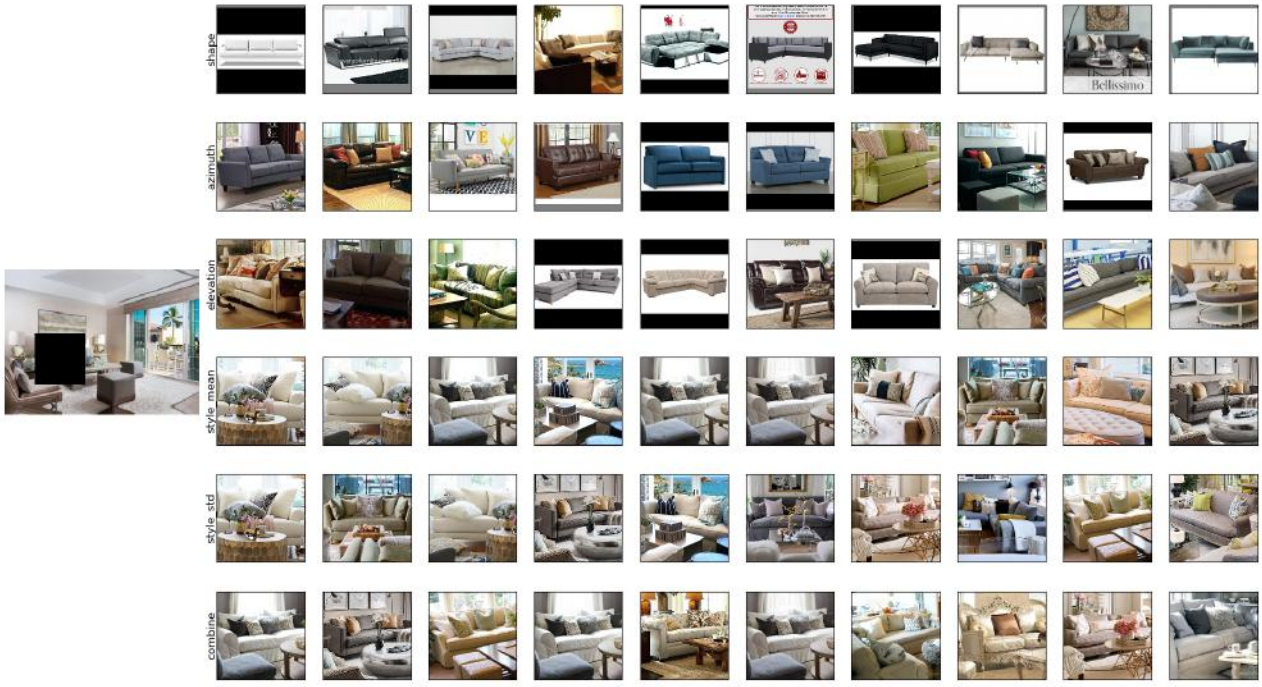

Figure 10: Retrieval for 'Couch' (1)

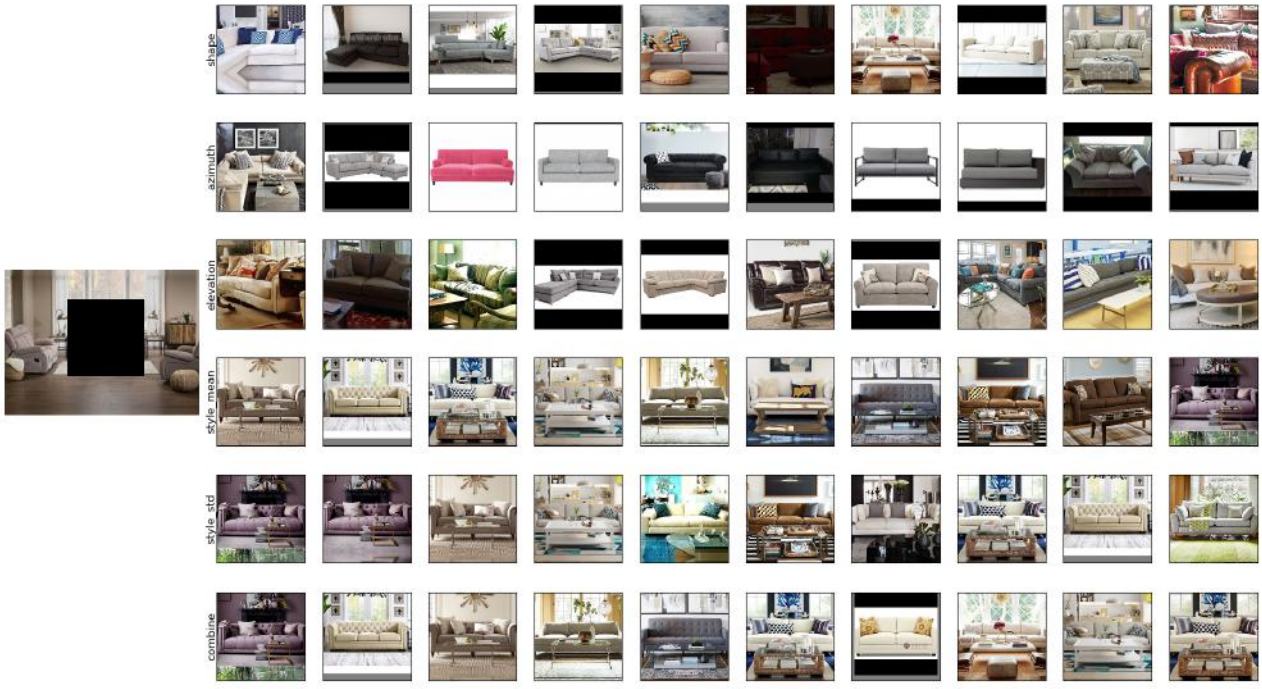

Figure 11: Retrieval for 'Couch' (2)

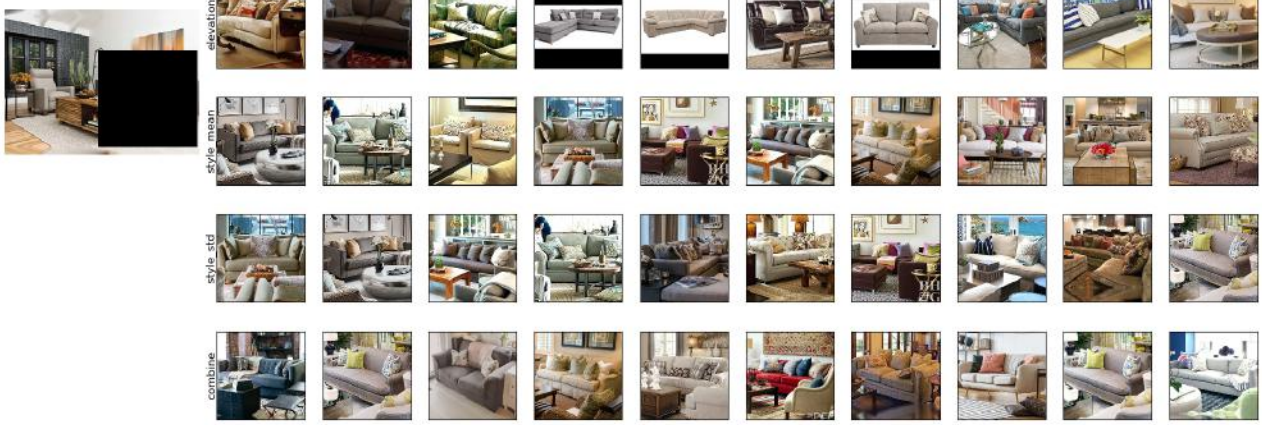

Figure 12: Retrieval for 'Couch' (3)

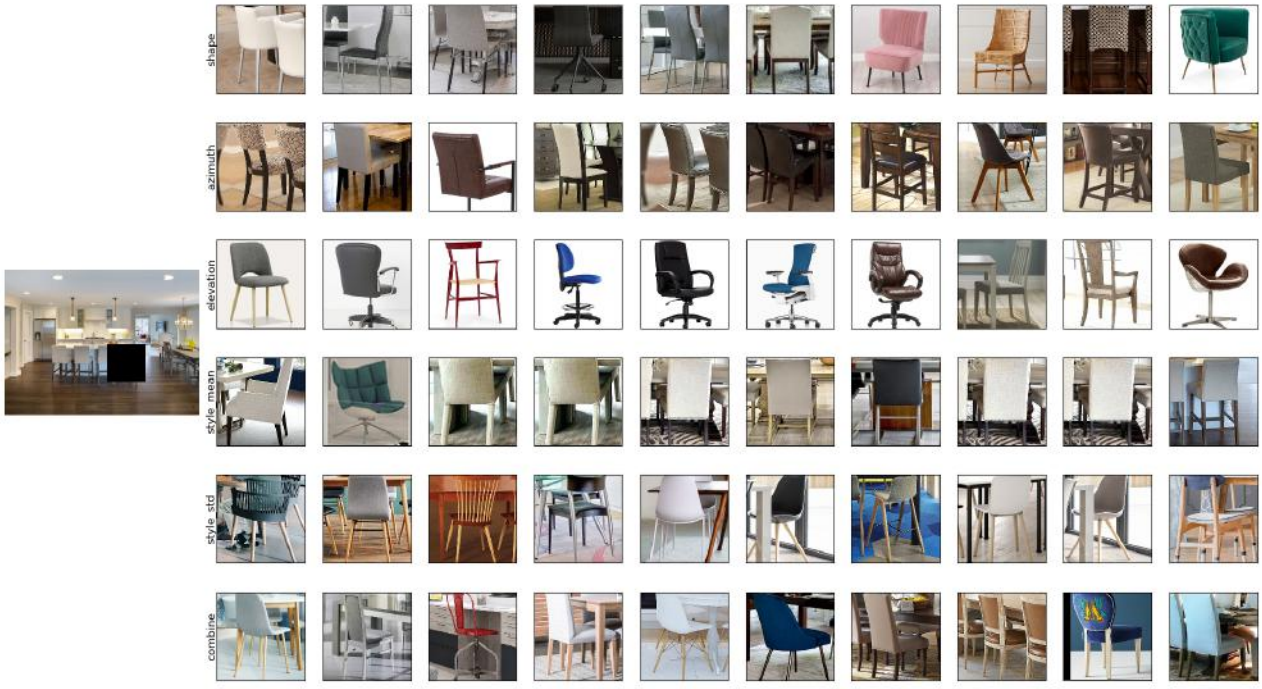

Figure 13: Retrieval for 'Chair' (1)

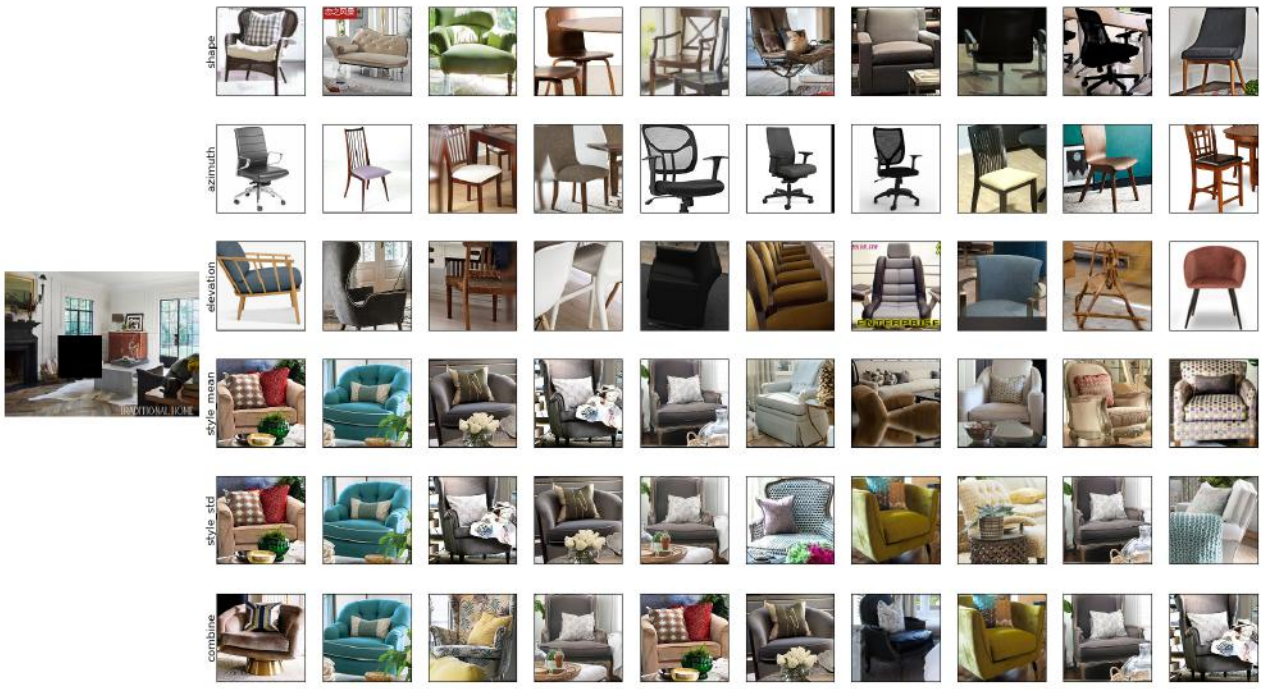

Figure 14: Retrieval for ‘Chair’ (2)

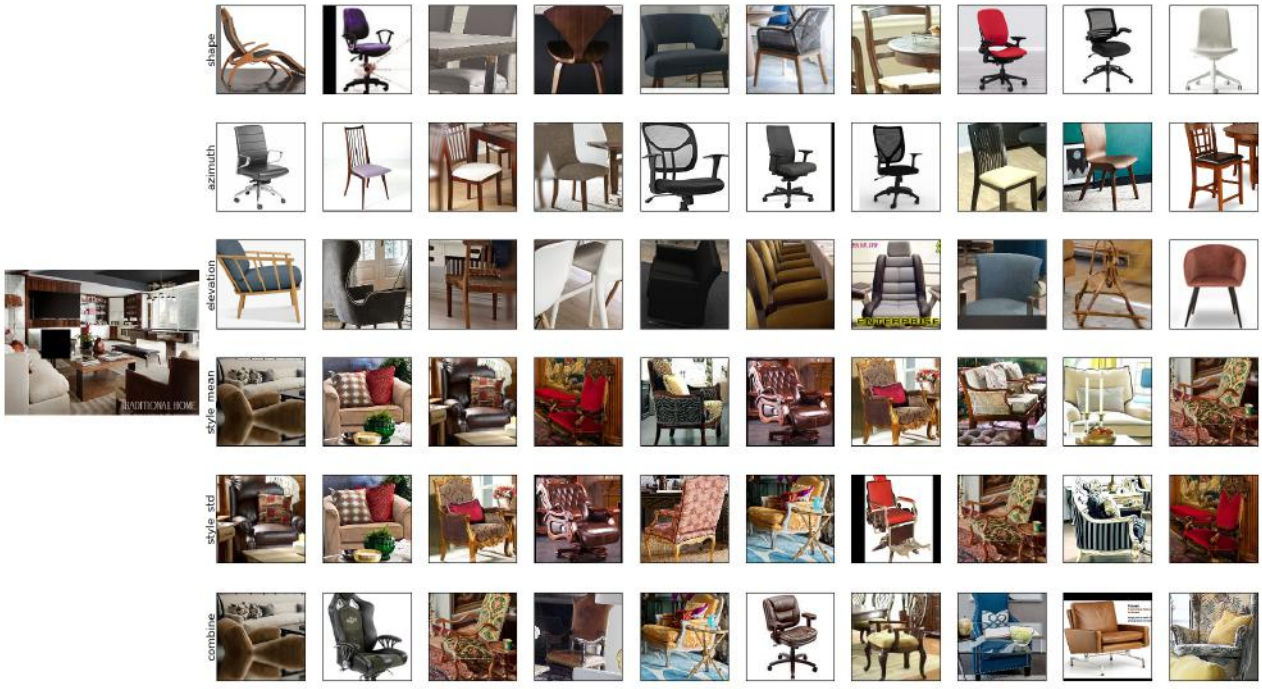

Figure 15: Retrieval for ‘Chair’ (3)

12. Evaluation Set

The remaining figures show the entire set of backgrounds used to construct our evaluation set. Next to each background shown are 30 foreground candidates ranked by the AMT workers (the average score in the range $[0, 3]$ is shown above each foreground). It may be seen that the few top ranked foregrounds indeed fit the query background quite well.

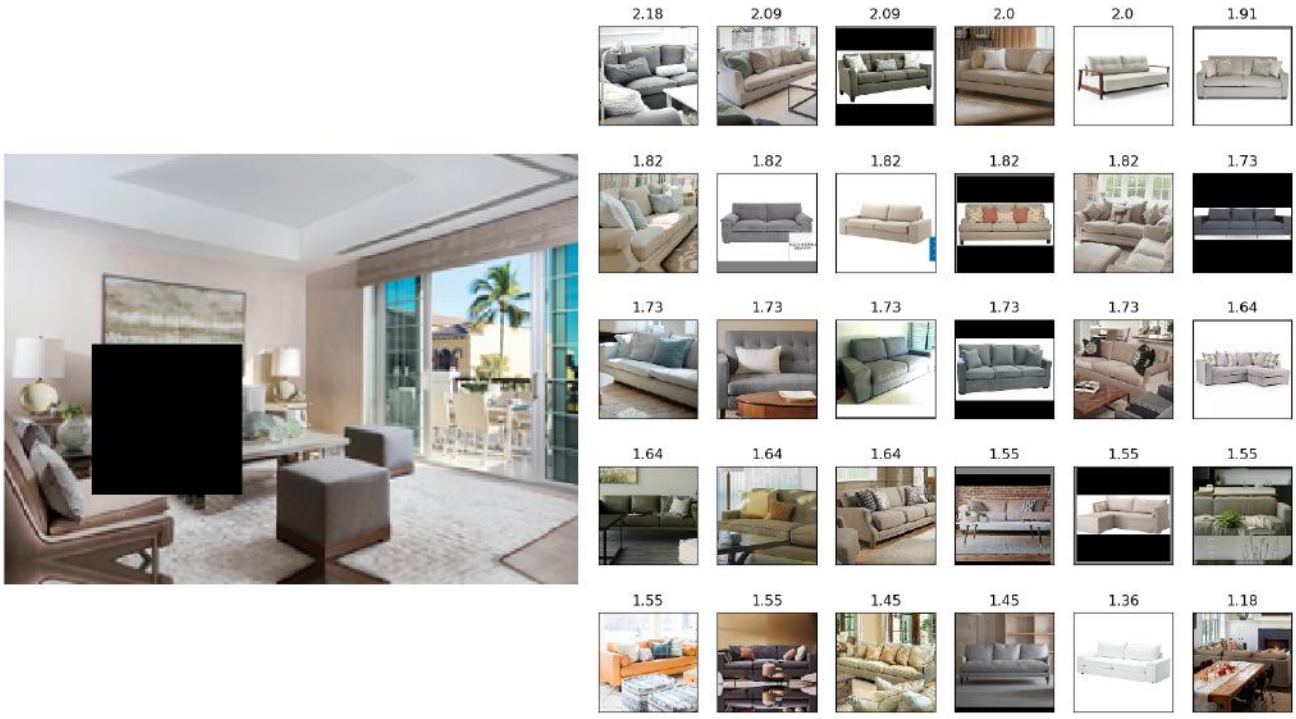

Figure 16: User rankings for background 1

Figure 17: User rankings for background 2

Figure 18: User rankings for background 3

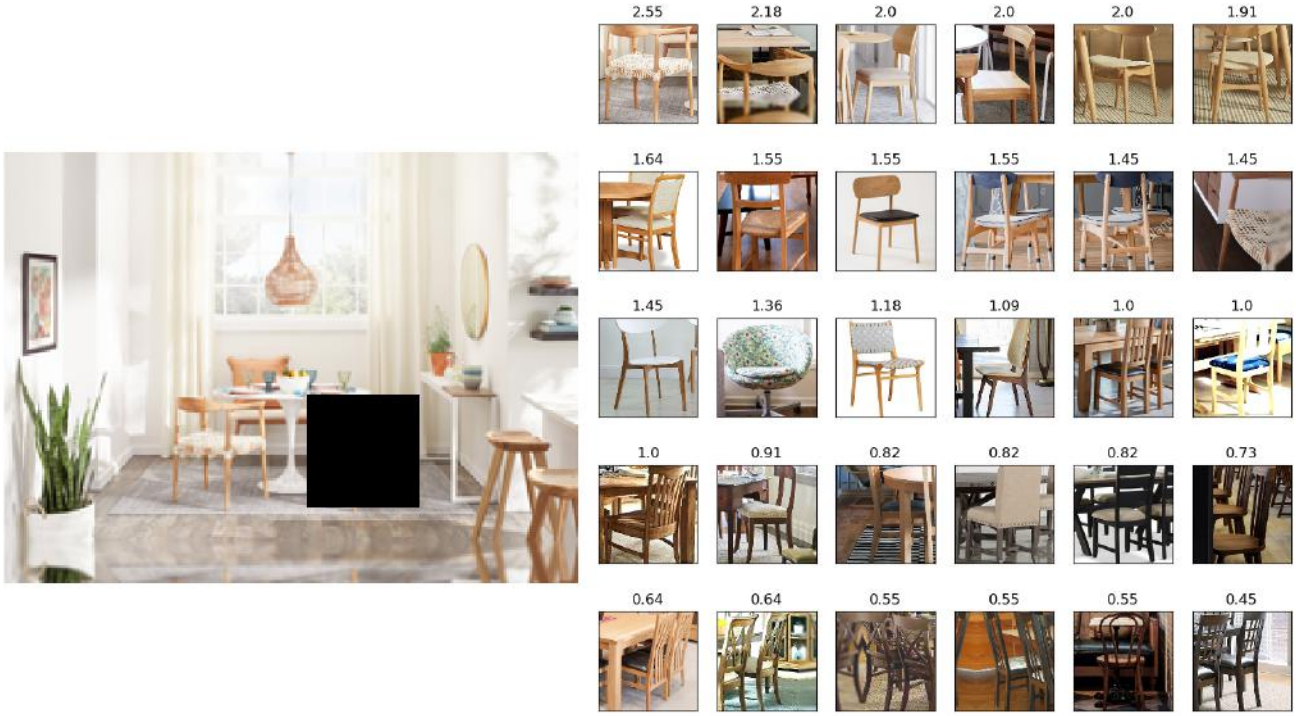

Figure 19: User rankings for background 4

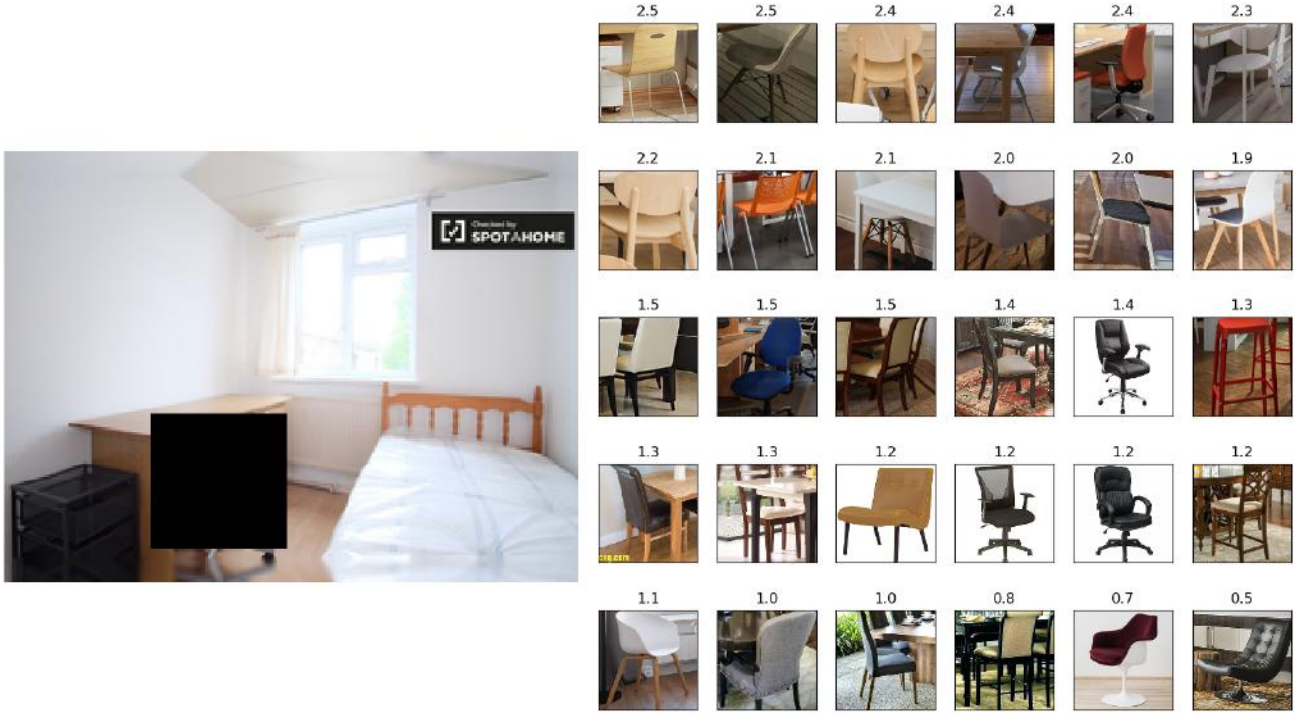

Figure 20: User rankings for background 5

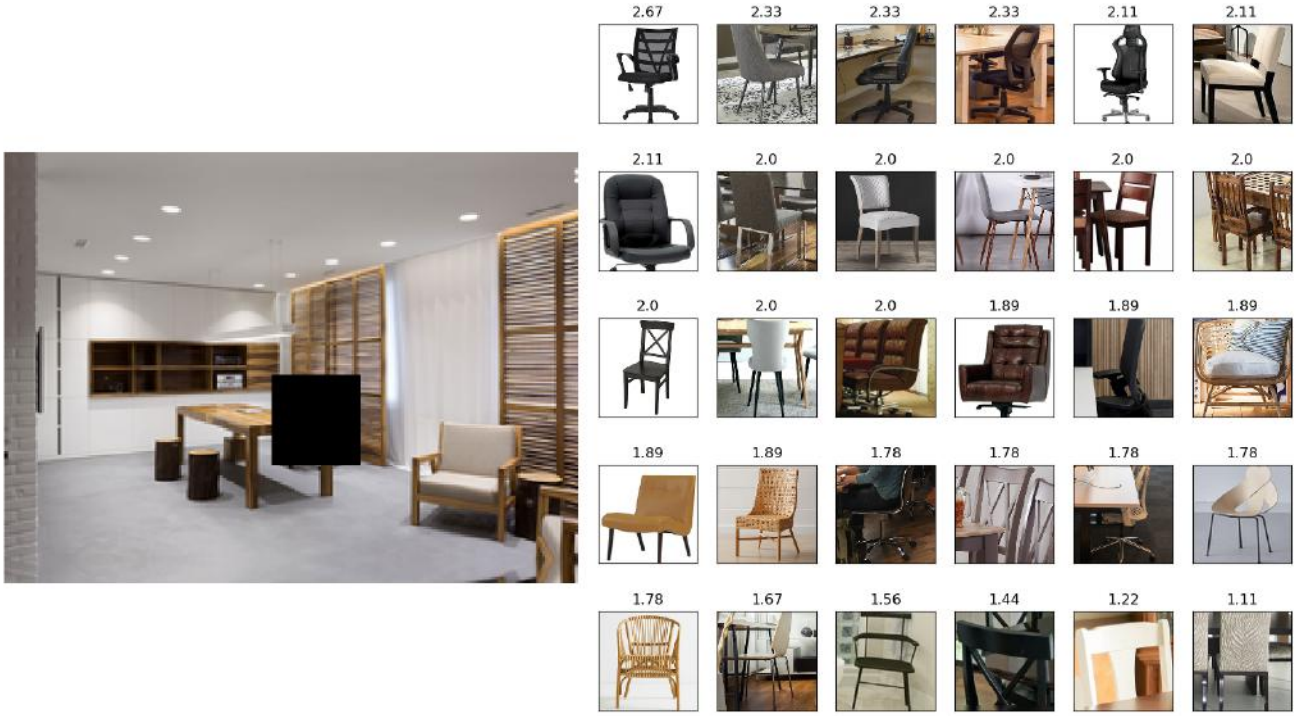

Figure 21: User rankings for background 6

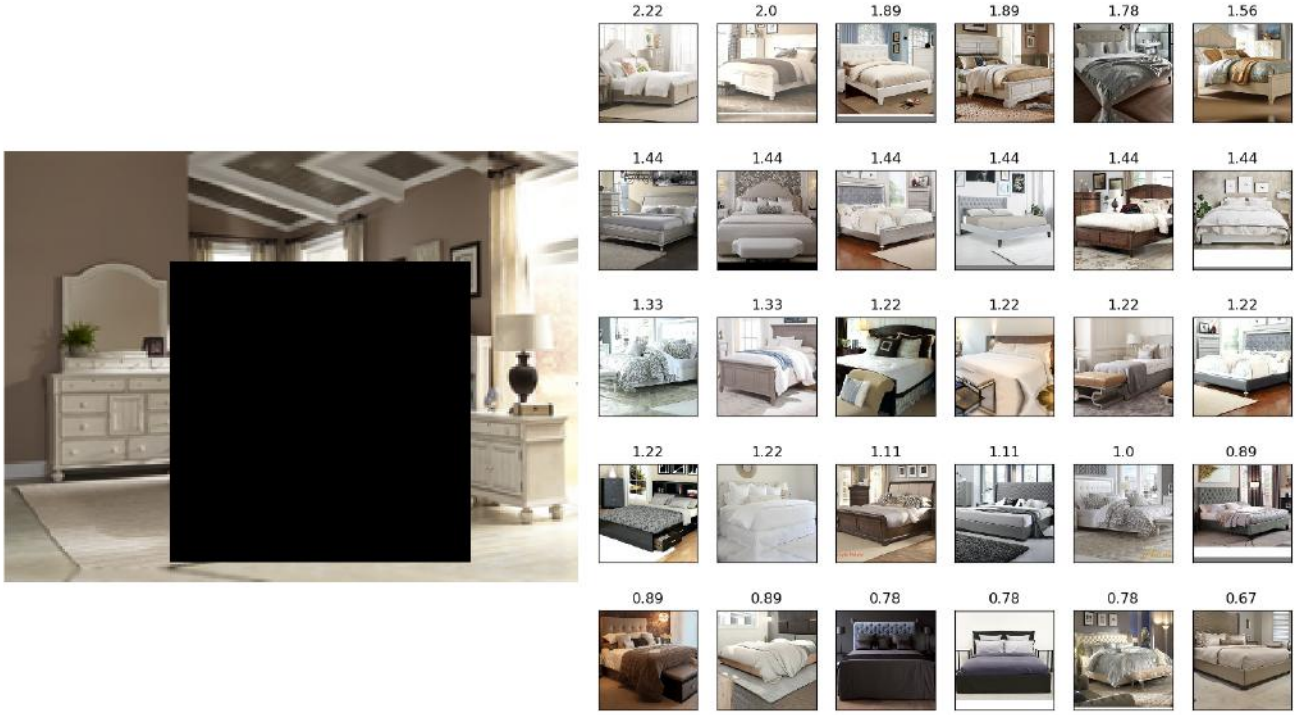

Figure 22: User rankings for background 7

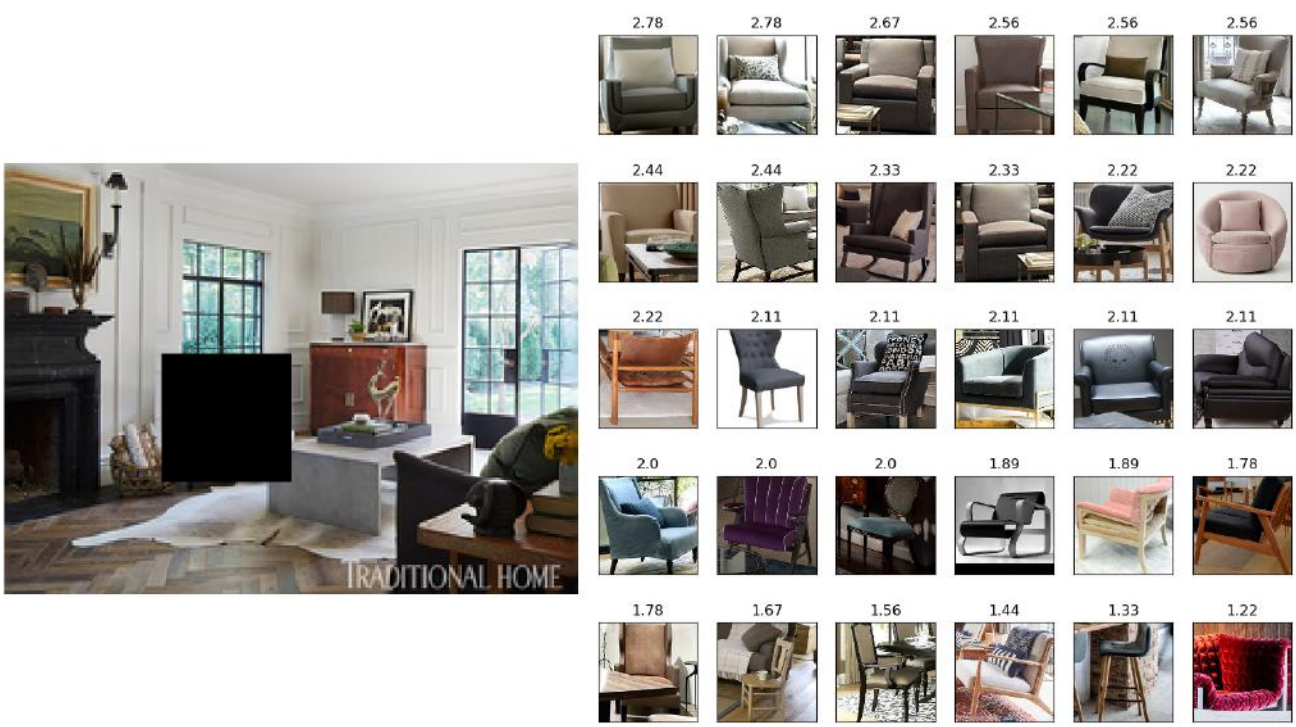

Figure 23: User rankings for background 8

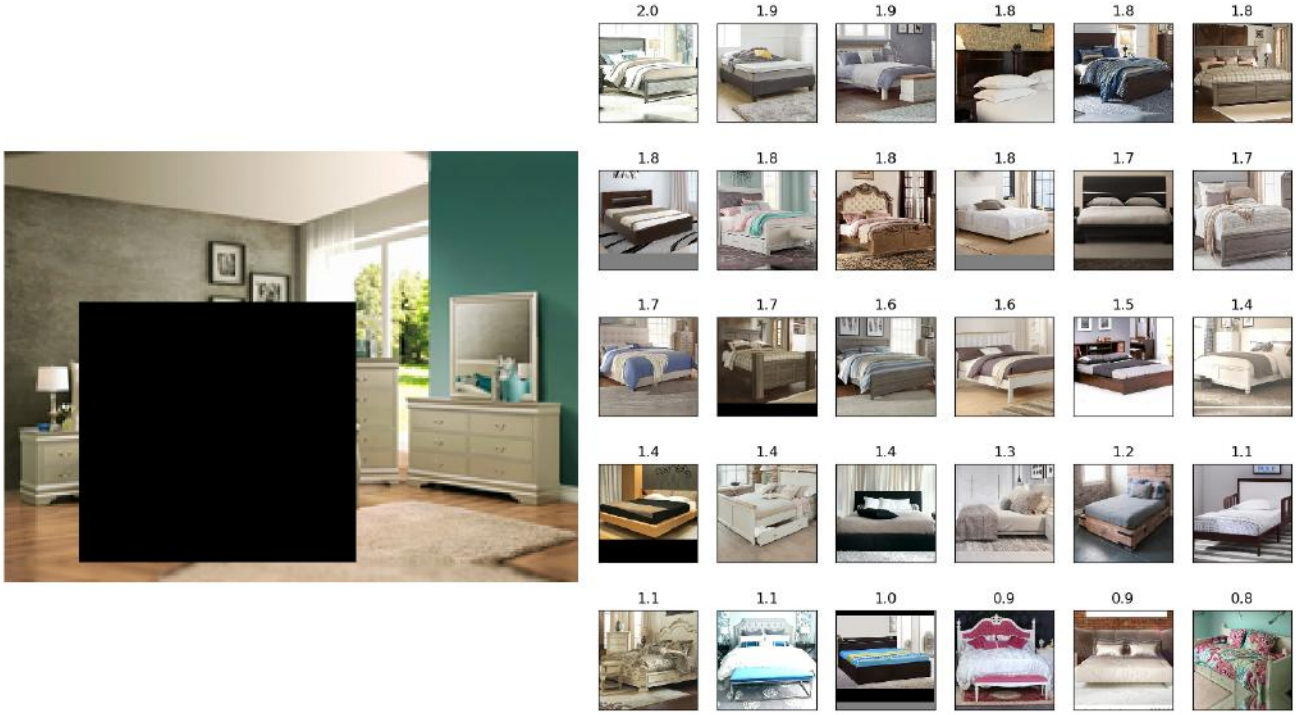

Figure 24: User rankings for background 9

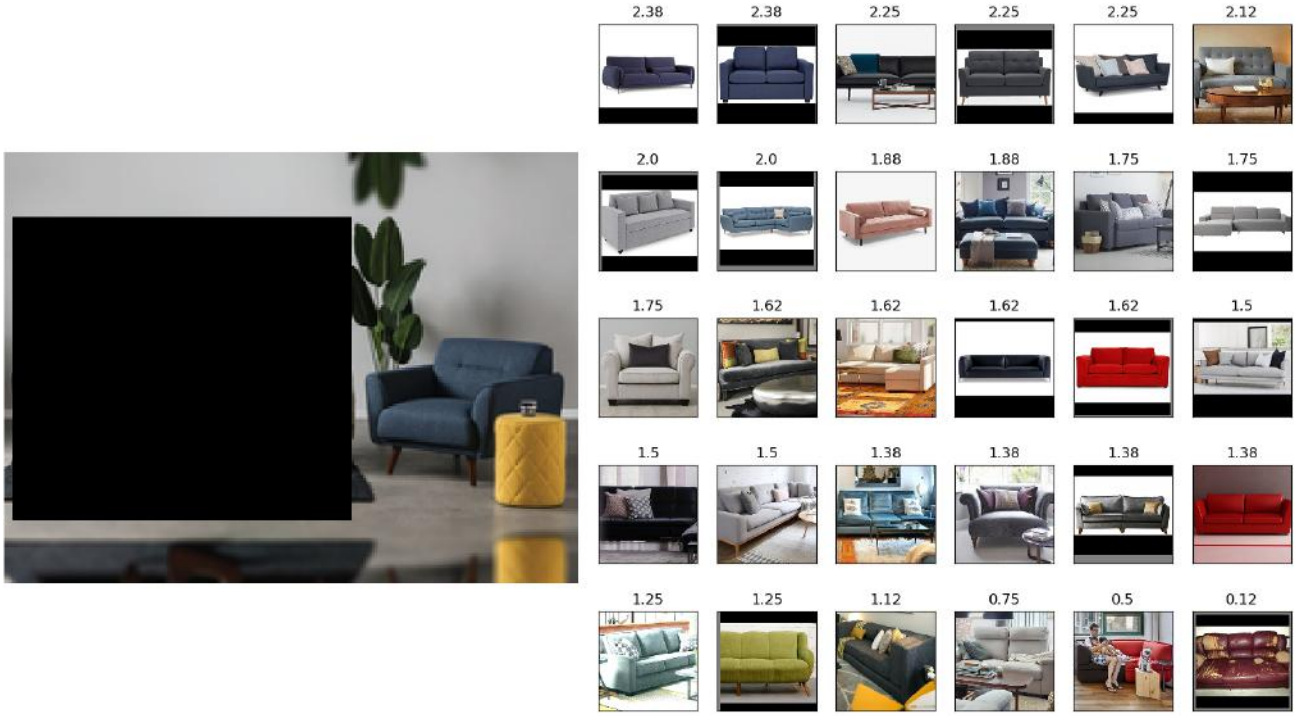

Figure 25: User rankings for background 10

Figure 26: User rankings for background 11

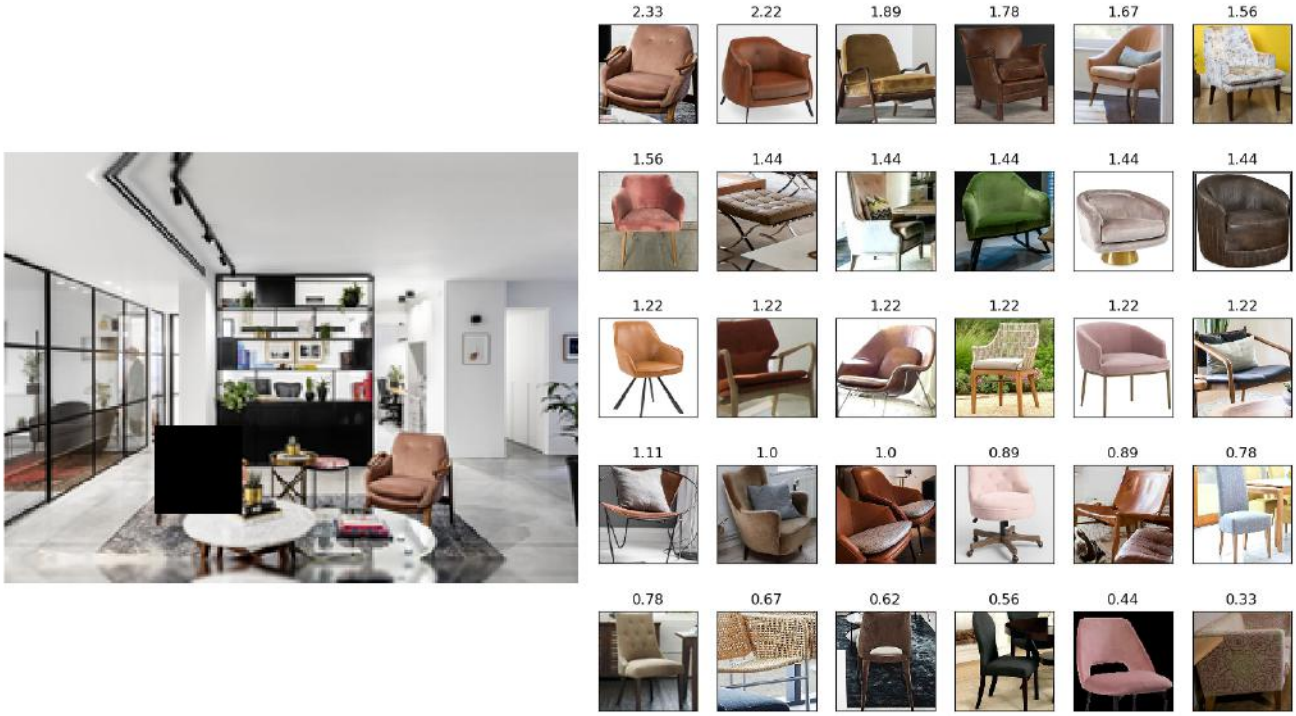

Figure 27: User rankings for background 12

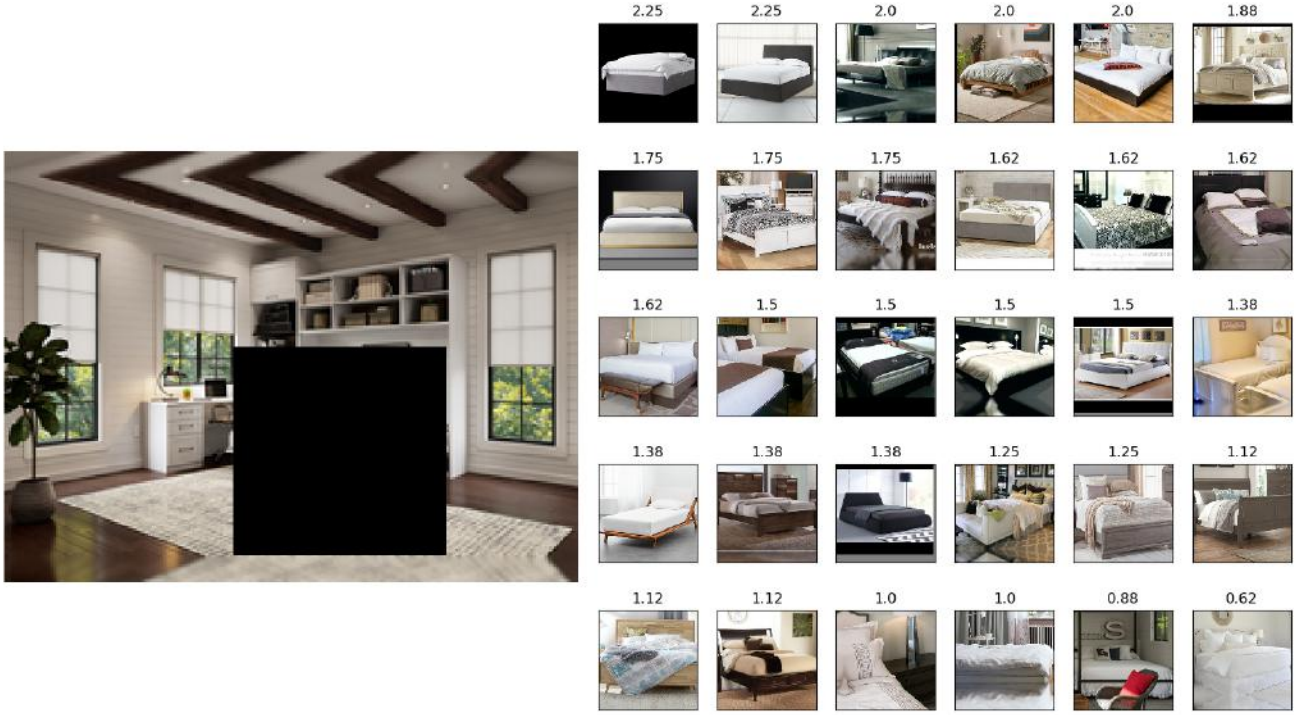

Figure 28: User rankings for background 13

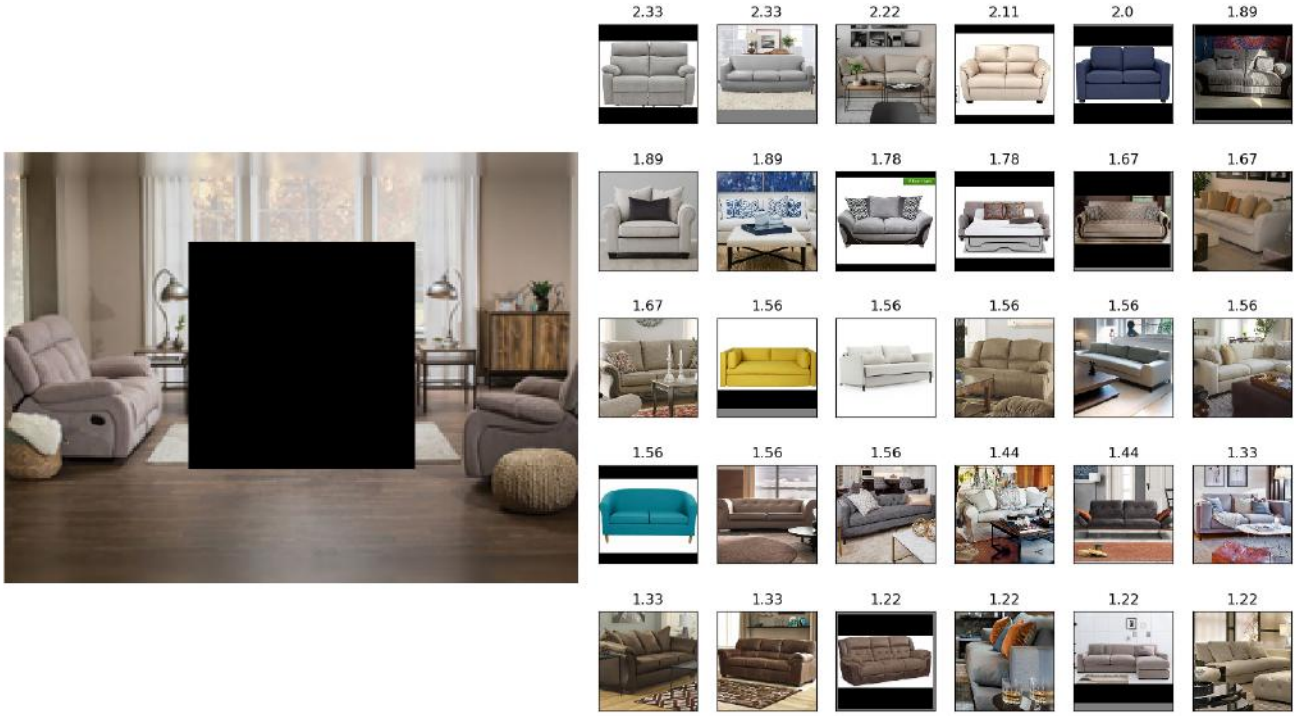

Figure 29: User rankings for background 14

Figure 30: User rankings for background 15