000
001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044

Supplementary Material: Retrieving Visually Overlapping Images Through Interpretable Non-Metric Embeddings

Anonymous ECCV submission

Paper ID 4294

1 Interpretable Queries

Further to Section 4.2 in our main work, we demonstrate the interpretability of our method on additional test queries from the MegaDepth dataset .We show three different types of figures:

- Interpretability plots: For a query image from the test set what do the predicted enclosure and concentration of retrievals from the **training** set images tell us about their relationship to one-another?
- Generalization plots: Does our embedding generalize to the images from the **test** set and are pairwise relationship interpretations valid?
- Relative scale plots: Given pairs of images from the test set and their box representations, can we estimate their relative scale difference?

1.1 Interpretability Plots

These plots (Figures 4 5 6 7 8 9 10 11 12) illustrate the interpretability that we gain when using box embeddings trained with normalized visual overlap as the world-space measure. The key takeaway is that we can qualitatively observe the relationship between the query and each of the retrieved images. The relationships can be grouped into four categories: Given a query image, a retrieval can be a:

- zoom-out
- crop-out or oblique-out
- close-up
- clone-like

We show three different examples for the scenes $Venice$ and $Florence$ and two examples for $BigBen$ and $NotreDame$ additionally to Figure 5 in the main work. For each figure a single query and up to 36 retrieved images from the training set are shown. They are placed into one of six buckets according to the **predicted** normalized box overlap $\mathsf{NBO}(\mathbf{b_x} \mapsto \mathbf{b_y})$ and $\mathsf{NBO}(\mathbf{b_y} \mapsto \mathbf{b_x})$. The vertical axis describes the enclosure, so $\mathsf{NBO}(\mathbf{b_x} \mapsto \mathbf{b_y})$, or "*how much surface from the query image is visible in the retrieved image*". The horizontal

axis describes the concentration $\mathsf{NBO}(\mathbf{b_y} \mapsto \mathbf{b_x})$, in other words "*how much surface of the retrieved image is visible in the query image*".

The numbers below each image are enclosure and concentration estimated with box representations as well as ground-truth values estimated with semi-dense depth maps.

1.2 Generalization Plots

The interpretability plots retrieve images from the training set of our box embeddings. Next, we demonstrate qualitatively that the learned representations generalize to the images in the test set. Here the test set are images in MegaDepth that do not have dense depth information. Hence, we can only report qualitative results for this larger test set.

So, *we now retrieve images from the test set* using a random query image from the test set. Here, we plot retrieved images on a 2D grid, using enclosure and concentration of the retrieved image as 2D coordinate, where the x-coordinate denotes the concentration, and the y-coordinate denotes the enclosure.

These plots (Figures 13 14 15 16) also provide interpretability, as one can observe different clusters of images as zoom-outs, crop-outs, close-ups and clones of the query in similar "quadrants" as Interpretability plots.

1.3 Relative Scale Plots

Relative scale plots are further qualitative examples similar to Figure 1 and Figure 7 in the main paper.

We illustrate that we can estimate geometric relationships between two images from the test set using our box embeddings. For two images from the test set we can estimate the relative scale of the first image in the second image. Hence, we plot two test images and a rescaled version of the first image such that any geometric verification between the two images is now easier to do due to matching scale.

This relative scale estimate is relatively accurate if the images have zoom-in/zoom-out relationship. If the images are in crop-out/oblique-out to one-another, then the rescaling is not necessarily going to make matching easier. So, image pairs that seem to be failure cases in terms of estimated relative scale often have low enclosure value ($<80\%$), which means that these image pairs can be detected and treated accordingly. To demonstrate that this filtering approach is effective, we show failure cases with an enclosure of at least 80%.

2 Further Localization Evaluations

In this section we provide further evaluations of different embeddings for localization task.

2.1   7-Scenes

Figure 1 shows sorted rotation errors for each scene, similarly to Figure 6 in the main paper, for different embeddings. Again, rotation error is computed per query, when matched against 10-th and 30-th nearest neighbor from the training set (gallery).
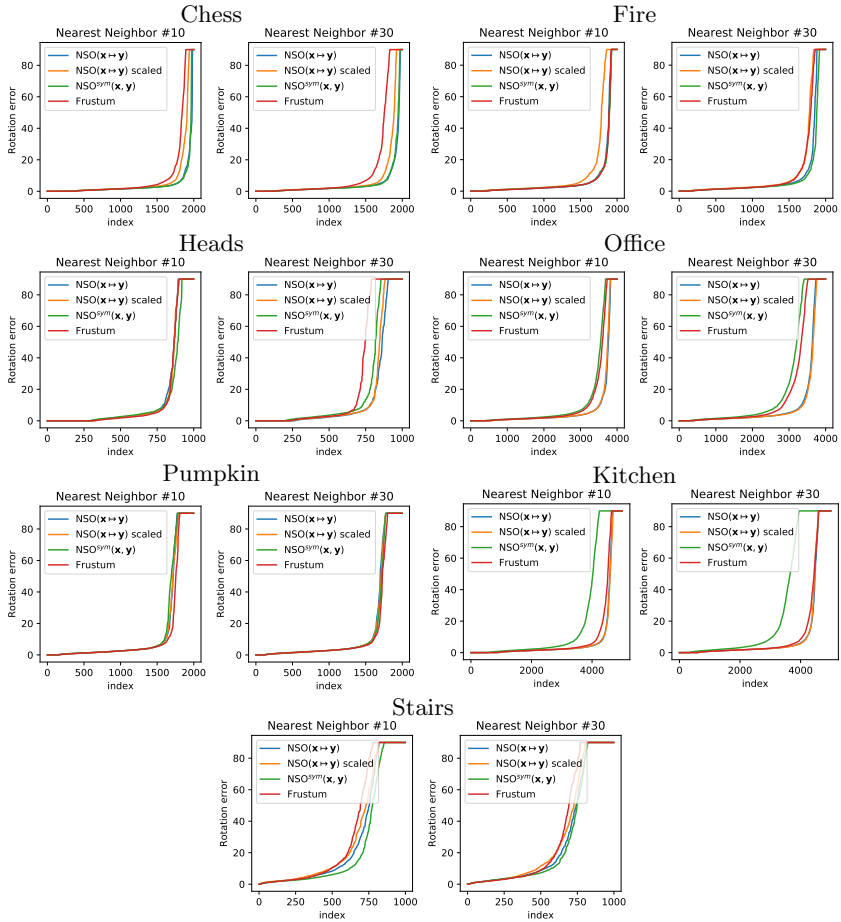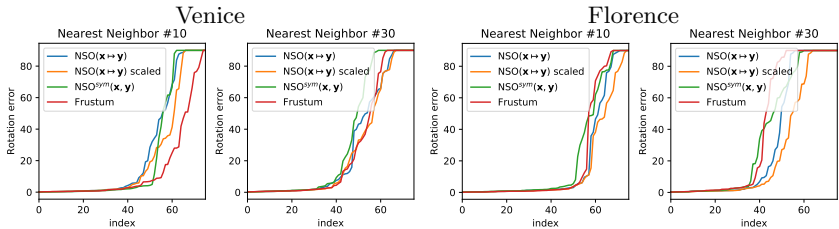

Fig. 1. 7-Scenes pose rotation error. Each plot shows (sorted) rotation error (capped at $90°$) when each test image is matched against 10-th and 30-th closest retrieved image for pose estimation. As we can see, box embeddings with surface overlap measure tend to outperform alternatives, especially when rescaling images according to estimated relative scale.

When solving for the pose, retrieving 10-th nearest neighbor for matching seems to be sufficient to estimate good pose for most of the scenes. However,

even in this setting we can see that Frustum overlap is under-performing compared to embeddings trained with surface overlap. Only in Pumpkin scene with 10-th retrieved image, the frustum overlap is marginally better. Between surface overlap-based embeddings the performance is quite comparable. There are no systematic improvements nor deteriorations with relative scale correction as captured scenes are all rooms with limited scale variation.

2.2   Megadepth

Figure 2 shows evaluation results for Venice and Florence scenes to complement Figure 6 of the main paper.

Fig. 2. Additional results to Fig. 6 on MegaDepth. Each plot shows (sorted) rotation error (capped at $90°$) when each test image is matched against 10-th and 30-th closest retrieved image for pose estimation. As we can see, box embeddings with surface overlap measure tend to outperform alternatives, especially when rescaling images according to estimated relative scale. Results evaluated for 100 images from the test set.

In Table 1 we report median error translation and rotation error of estimated pose for 100 images of the test set (that have corresponding depth maps) similar to Table 2 in the main paper. The Megadepth scenes are not metric, so the scale factor of translation errors is not known. Furthermore, all these errors are relatively low, corresponding to accurate pose, so it is difficult to draw conclusions from these results.

Hence, we also evaluated different embeddings with a larger test set which consists of images in Megadepth that do not have depth maps. This results in 3165, 1931, 2255 and 1159 test set images for Big Ben, Notre-Dame, Venice and Florence, respectively. Table 2 show median errors for the larger test set. As can be seen, correcting for scale using Box embeddings is superior to alternatives on 3 scenes.

Similarly, (sorted) rotation error evaluated for the larger test set could seen in Figure 3. Here the error is computed against 1st nearest neighbor retrieval. These plots indicate a similar conclusion. The surface overlap based embeddings are outperformed by Frustum overlap embedding for Florence scene. Florence scene has images that capture a large area with complex narrow streets, however the training set consists of only 1471 images. We suspect that our CNNs need more training data to learn generalizable surface overlaps.

Training $\mathcal{L}$	Box $\mathrm{NSO}(\mathbf{x} \mapsto \mathbf{y})$	Box Scaled $\mathrm{NSO}(\mathbf{x} \mapsto \mathbf{y})$	Vector $\mathrm{NSO}^{sym}(\mathbf{x}, \mathbf{y})$	Vector Frustum
Retrieval func.	$\mathrm{NBO}(\mathbf{b_x} \mapsto \mathbf{b_y})+$ $\mathrm{NBO}(\mathbf{b_y} \mapsto \mathbf{b_x})$	$\mathrm{NBO}(\mathbf{b_x} \mapsto \mathbf{b_y})+$ $\mathrm{NBO}(\mathbf{b_y} \mapsto \mathbf{b_x})$	Eucl. dist.	Eucl. dist.
Notre-Dame	.038, **0.79°**	.038, 0.87°	**.036**, 0.83°	.047, 1.05°
Big Ben	**.067**, 0.87°	.089, 0.98°	.070, 0.87°	.096, **0.83°**
Venice	.096, 1.01°	.098, 1.23°	**.088, 0.85°**	.085, 0.91°
Florence	.081, 1.08°	.079, 1.10°	.054, 0.87°	**.048, 0.68°**

Table 1. Comparison of rotation and translation errors on the MegaDepth dataset, where boxes learn surface overlap asymmetrically while vectors are trained symmetrically. The first entry of each cell denotes the translation error up to scale, the second entry is the rotation error in degrees. Results evaluated for 100 images from the test set.

Training $\mathcal{L}$	Box $\mathrm{NSO}(\mathbf{x} \mapsto \mathbf{y})$	Box Scaled $\mathrm{NSO}(\mathbf{x} \mapsto \mathbf{y})$	Vector $\mathrm{NSO}^{sym}(\mathbf{x}, \mathbf{y})$	Vector Frustum
Retrieval func.	$\mathrm{NBO}(\mathbf{b_x} \mapsto \mathbf{b_y})+$ $\mathrm{NBO}(\mathbf{b_y} \mapsto \mathbf{b_x})$	$\mathrm{NBO}(\mathbf{b_x} \mapsto \mathbf{b_y})+$ $\mathrm{NBO}(\mathbf{b_y} \mapsto \mathbf{b_x})$	Eucl. dist.	Eucl. dist.
Notre-Dame	0.84, 15.1°	**0.81**, **13.8°**	1.43, 38.2°	0.98, 25.2°
Big Ben	2.91, 58.0°	**2.85**, **53.9°**	2.81, 54.0°	3.30, 69.8°
Venice	3.20, 68.9°	**3.24**, **58.6°**	3.13, 70.8°	2.70, 65.6°
Florence	1.44, 35.0°	1.33, 31.6°	1.67, 42.2°	**0.75, 10.6°**

Table 2. Comparison of rotation and translation errors on the MegaDepth dataset for test set without depth images, where boxes learn surface overlap asymmetrically while vectors are trained symmetrically. The first entry of each cell denotes the translation error up to scale, the second entry is the rotation error in degrees. Total number of images in these test sets (in Figure order): 3165, 1931, 2255 and 1159.
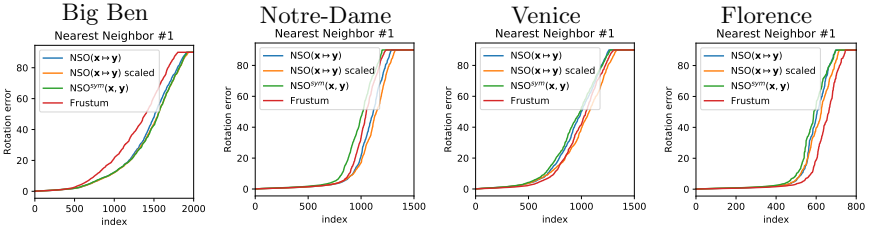
Fig. 3. Results on images without depth (MegaDepth) Each plot shows (sorted) rotation error (capped at $90°$) when each test image is matched against the closest retrieved image for pose estimation. As we can see, box embeddings with surface overlap measure tend to outperform alternatives, especially when rescaling images according to estimated relative scale. Total number of images in these test sets (in Figure order): 3165, 1931, 2255 and 1159.

	High concentration		Low concentration		
Low enclosure	77%/6% (88%/21%)	66%/5% (96%/11%)	42%/8% (0%/0%)	22%/7% (6%/3%)	14%/18% (0%/0%)
	68%/29% (80%/76%)	57%/18% (89%/45%)	42%/18% (20%/13%)	31%/26% (0%/0%)	7%/30% (1%/3%)
	70%/33% (75%/56%)	50%/45% (59%/69%)	34%/45% (23%/59%)	24%/45% (5%/17%)	7%/47% (2%/28%)
High enclosure	69%/66% (64%/55%)	52%/63% (54%/83%)	39%/54% (19%/64%)	20%/51% (11%/44%)	10%/50% (5%/52%)
		57%/80% (66%/66%)	34%/83% (36%/82%)	28%/70% (62%/59%)	6%/72% (6%/64%)
			37%/85% (25%/58%)	17%/90% (25%/88%)	10%/90% (12%/88%)

Query

Results of predicted and ground-truth enclosure and concentration relative to the query image on the left. The numbers below each image indicate the predicted and ground-truth concentration/enclosure. It can be observed that the images in the upper left quadrant are close-ups of the query. The images in the lower left quadrant are clones of the query. The retrieved clones preserve the normals of the surfaces in the query. Images in the lower right quadrant are zoom-outs of the scene in the query. The images in the upper right corner are mostly oblique-outs and show the scene in the query from different angles. *This caption applies to all Interpretability plots unless otherwise stated.*
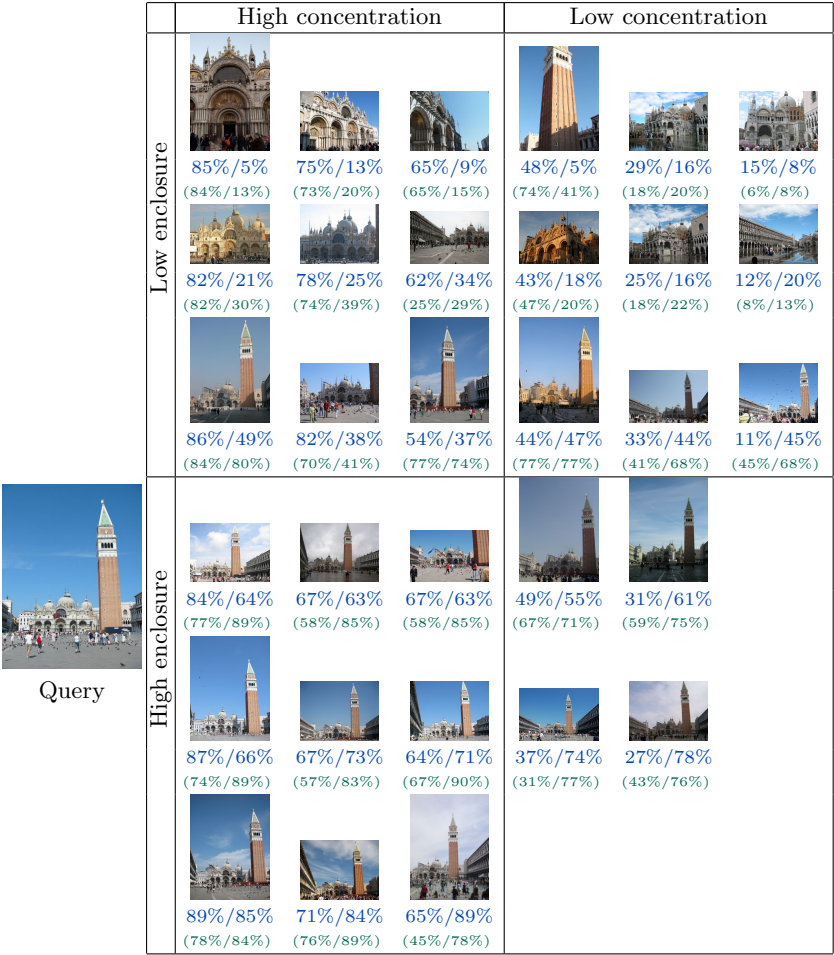
Fig. 4. Interpretability plot: Venice No. 1.

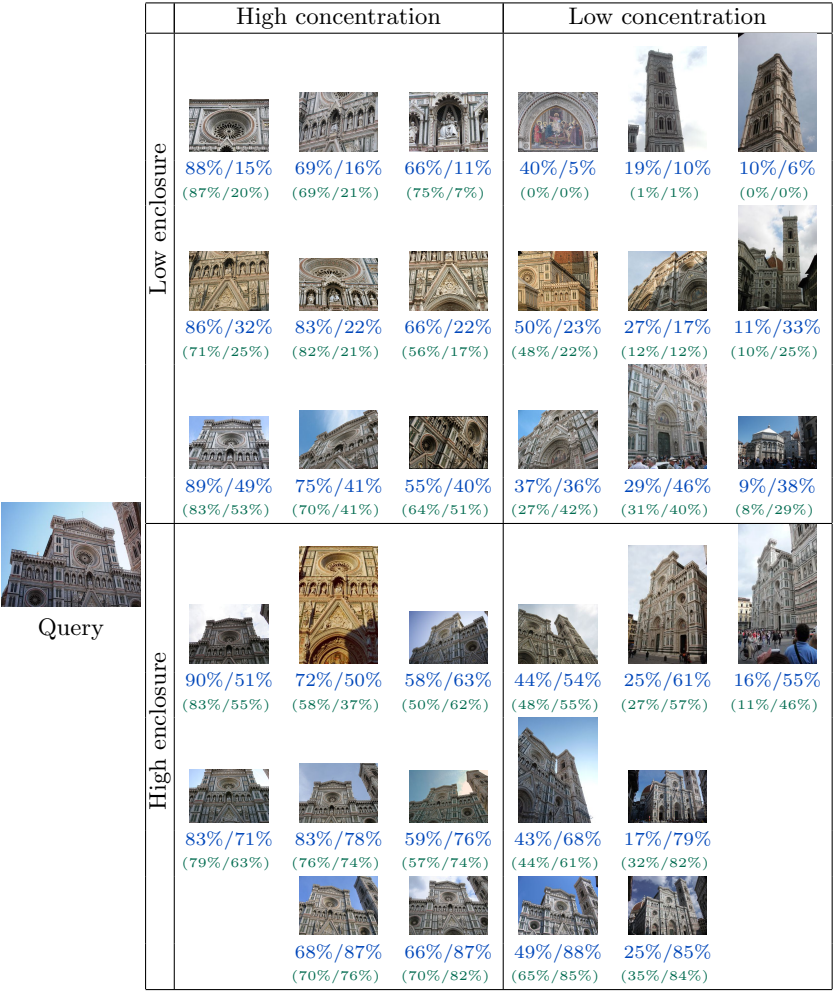**Fig. 5.** Interpretability plot: Venice No. 2.
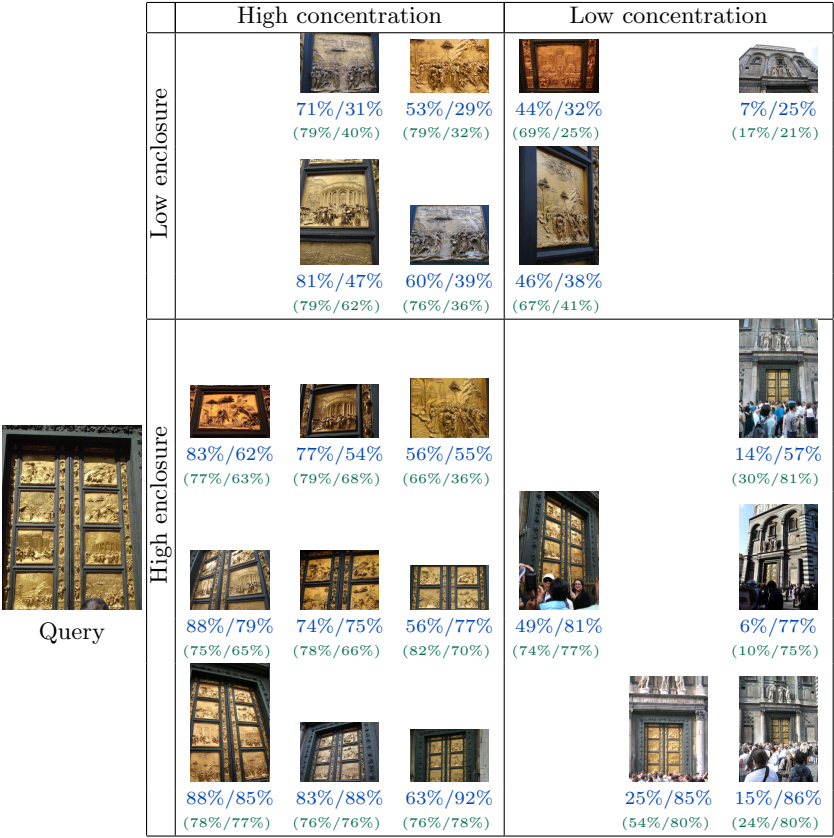
Fig. 6. Interpretability plot: Florence No. 1.
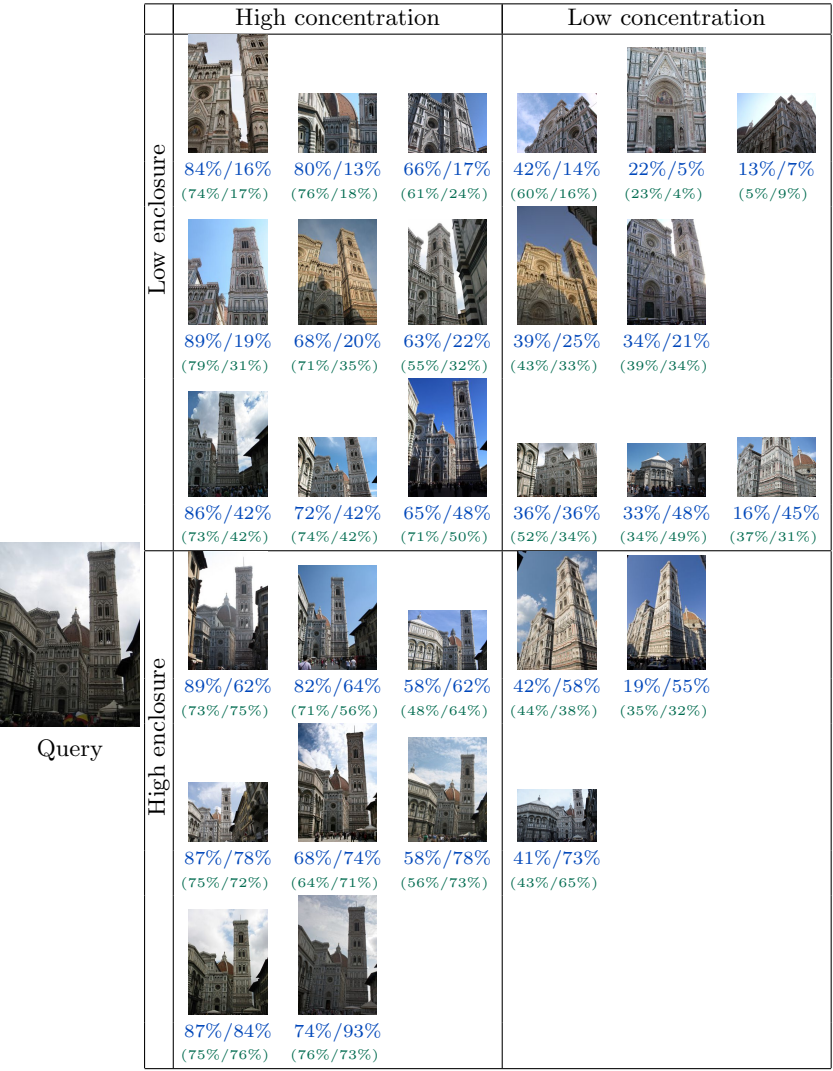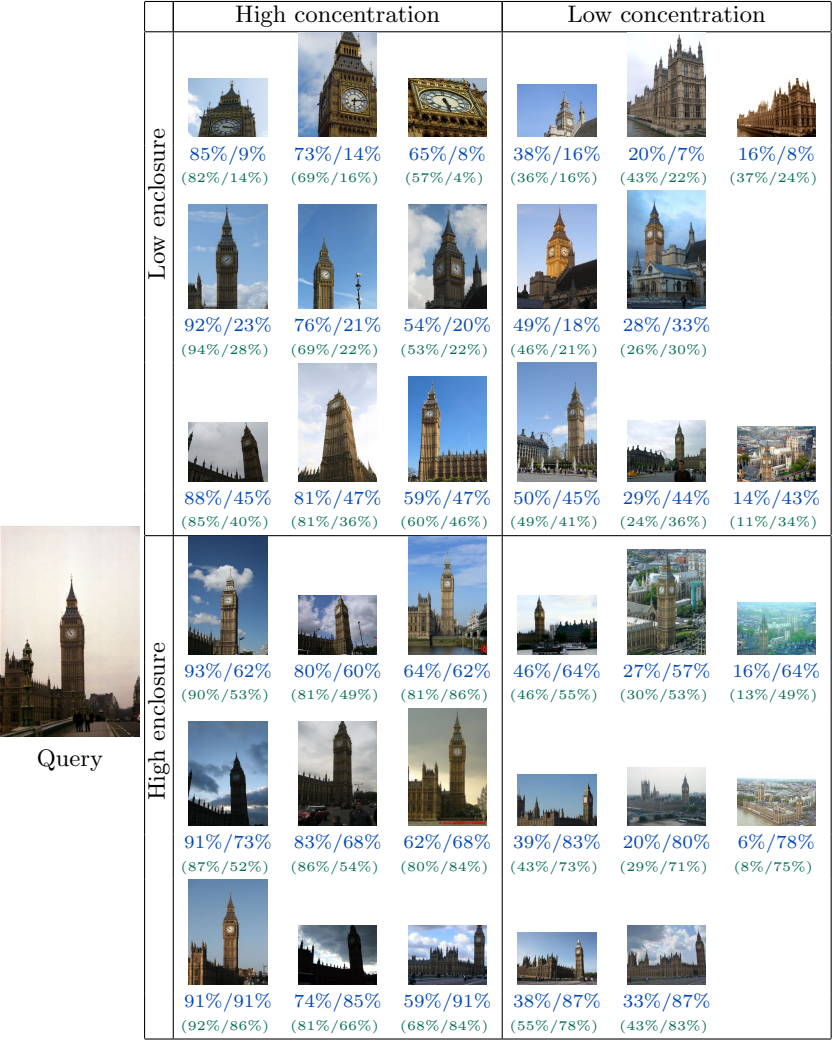
Fig. 7. Interpretability plot: Florence No. 2.

		High concentration			Low concentration		
Low enclosure		84%/16% (74%/17%)	80%/13% (76%/18%)	66%/17% (61%/24%)	42%/14% (60%/16%)	22%/5% (23%/4%)	13%/7% (5%/9%)
		89%/19% (79%/31%)	68%/20% (71%/35%)	63%/22% (55%/32%)	39%/25% (43%/33%)	34%/21% (39%/34%)	
		86%/42% (73%/42%)	72%/42% (74%/42%)	65%/48% (71%/50%)	36%/36% (52%/34%)	33%/48% (34%/49%)	16%/45% (37%/31%)
High enclosure		89%/62% (73%/75%)	82%/64% (71%/56%)	58%/62% (48%/64%)	42%/58% (44%/38%)	19%/55% (35%/32%)	
		87%/78% (75%/72%)	68%/74% (64%/71%)	58%/78% (56%/73%)	41%/73% (43%/65%)		
		87%/84% (75%/76%)	74%/93% (76%/73%)				

Query

Fig. 8. Interpretability plot: Florence No. 3.

Fig. 9. Interpretability plot: Big Ben No. 1.
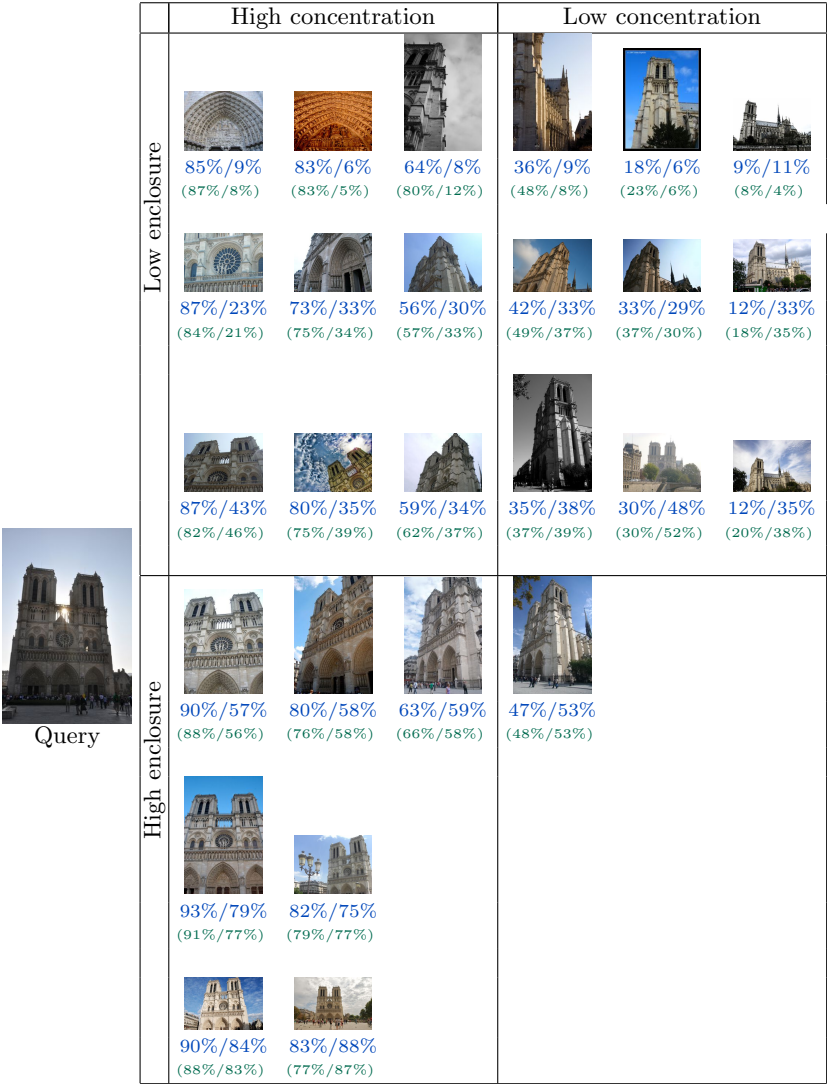
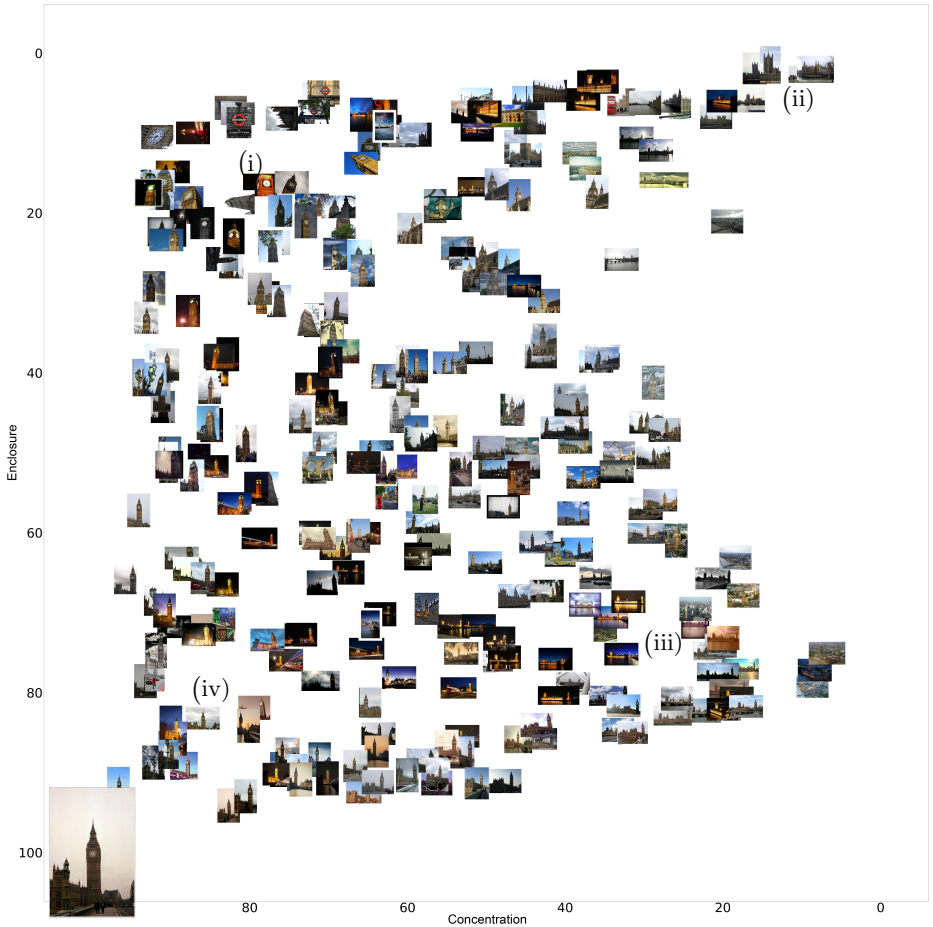**Fig. 10.** Interpretability plot: Big Ben No. 2.

	High concentration			Low concentration		
Low enclosure	85%/9% (87%/8%)	83%/6% (83%/5%)	64%/8% (80%/12%)	36%/9% (48%/8%)	18%/6% (23%/6%)	9%/11% (8%/4%)
	87%/23% (84%/21%)	73%/33% (75%/34%)	56%/30% (57%/33%)	42%/33% (49%/37%)	33%/29% (37%/30%)	12%/33% (18%/35%)
	87%/43% (82%/46%)	80%/35% (75%/39%)	59%/34% (62%/37%)	35%/38% (37%/39%)	30%/48% (30%/52%)	12%/35% (20%/38%)
High enclosure	90%/57% (88%/56%)	80%/58% (76%/58%)	63%/59% (66%/58%)	47%/53% (48%/53%)		
	93%/79% (91%/77%)	82%/75% (79%/77%)				
	90%/84% (88%/83%)	83%/88% (77%/87%)				

Query

Fig. 11. Interpretability plot: Notre Dame No. 1.

	High concentration	Low concentration

Low enclosure

83%/5% (88%/9%) 74%/6% (80%/6%) 57%/7% (56%/6%) 34%/15% (33%/14%) 21%/8% (29%/17%) 7%/16% (30%/9%)

70%/28% (83%/27%) 51%/30% (64%/34%) 36%/18% (45%/9%) 32%/26% (34%/17%) 12%/24% (25%/15%)

74%/47% (68%/42%) 59%/45% (75%/35%) 35%/36% (77%/72%) 31%/38% (34%/22%) 15%/38% (37%/25%)

High enclosure

76%/62% (59%/43%) 64%/49% (84%/50%) 40%/54% (51%/55%) 30%/62% (45%/54%)

76%/81% (89%/79%) 55%/82% (84%/61%) 48%/82% (82%/89%) 28%/67% (28%/72%)

52%/96% (82%/81%) 41%/93% (81%/89%) 29%/90% (76%/88%)

Query

Fig. 12. Interpretability plot: Notre Dame No. 2.

In this caption we expand on details and explanations of Fig.5 (Right) from the main paper. We show a query image from the test set (lower left corner) and the concentration and enclosure between randomly sampled test images from the MegaDepth SfM data set for which no depth maps are provided. The query image shows Big Ben from the view of the Westminster Bridge. (i) It can be observed that close-ups on the tower clock are clustered around the coordinates (80,15) which is consistent with our terminology of retrievals with high concentration and low enclosure. (ii) The images in the upper right corner show the waterfront side of Westminster Palace. These are crop-outs of the query image. In fact, the tower in the lower left corner of the query is one of the two towers that mark the corners of the water-front side of the palace. The retrievals in the upper right quadrant of the cluster therefore extends the view of the query. (iii) The images in the lower right area of the cluster clearly show zoom outs, with the pointy bell tower visible in all images. (iv) Lastly, one can observe that the images in the *clone − like* category are in fact similar views on Big Ben. Note that some of the retrievals are rotated images and sometimes cause outlier predictions. *This caption applies to all generalization plots unless otherwise stated.*

Fig. 13. Generalization plot: Big Ben.
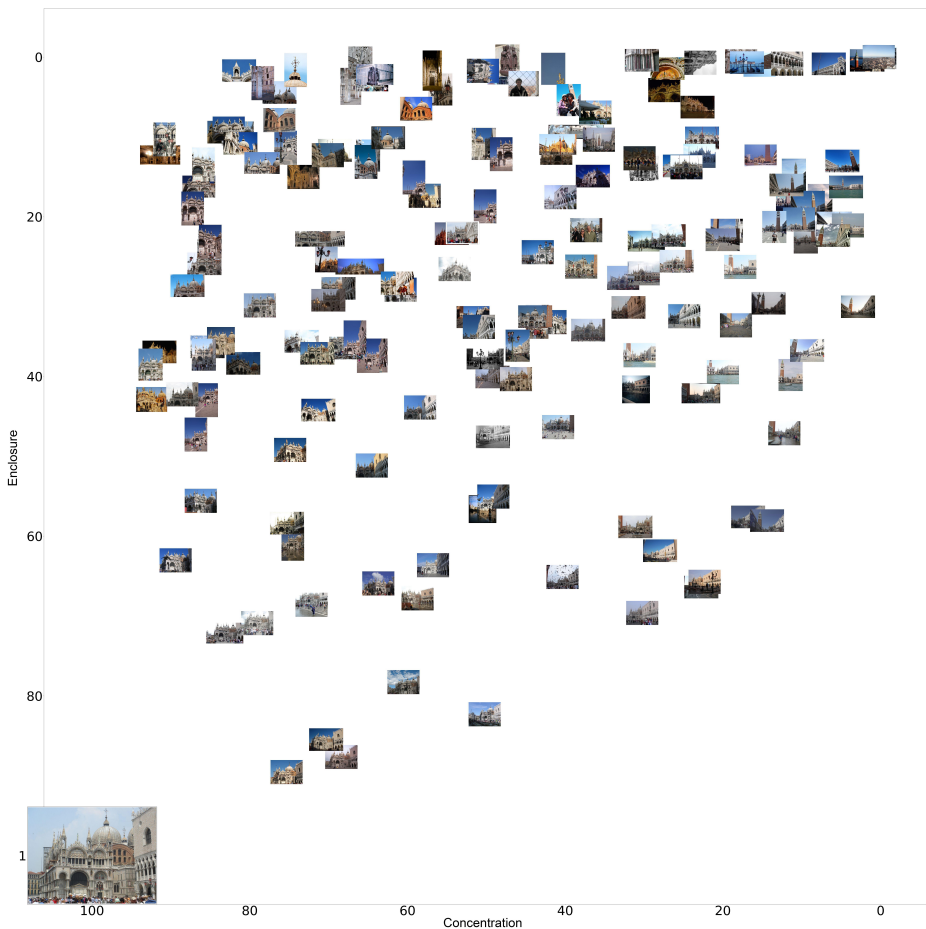
We show a query image from the test set (lower left corner) and the concentration and enclosure between randomly sampled test images from the MegaDepth SfM dataset for which no depth maps are provided. The query image shows the side view on Saint Mark's Basilica. One can observe the front of the Basilica from a very oblique angle in the left-most fifth of the image.

The images in the left upper corner show images with high concentration and low enclosure. According to our classification these are close-ups. Especially around the coordinates (80, 10) one can clearly observe zoomed in views on the side of the Basilica. The right upper quarter of the cluster consist of the crop-outs and oblique-outs. Note the images around the coordinate (10,25). These are mostly front views on the Basilica, and correspond to the left-most part of the query image–from a very different angle. Lastly, observe that images in the left lower corner are similar to the query, and images around (20,60) are zoom-outs.

Fig. 14. Generalization plot: Venice.
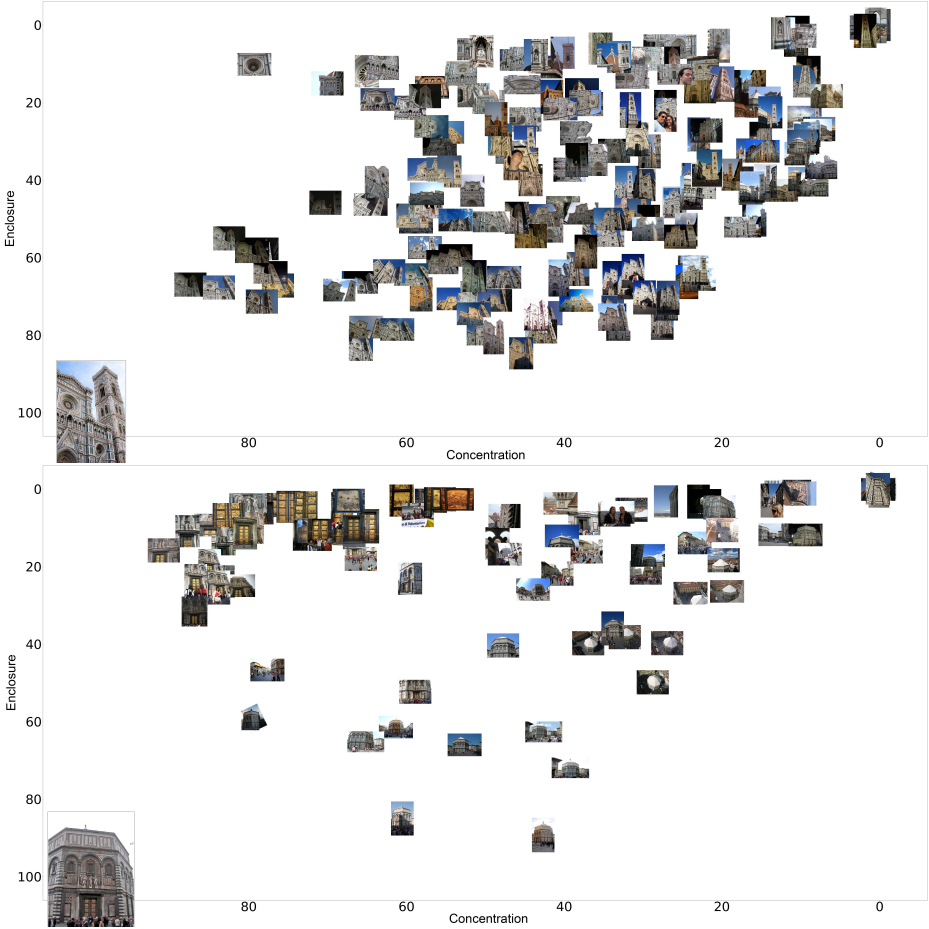
Because the scenes is very complex, streets are narrow, and there are not many images of the same view we show two scenes from Florence.

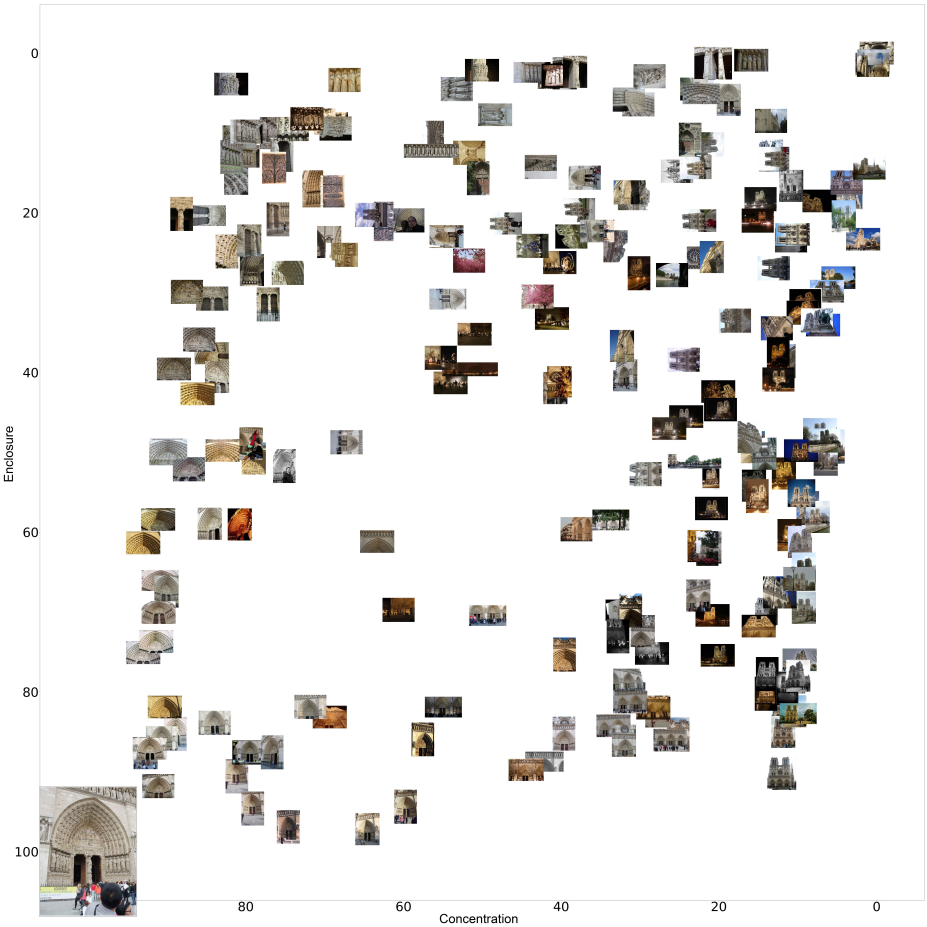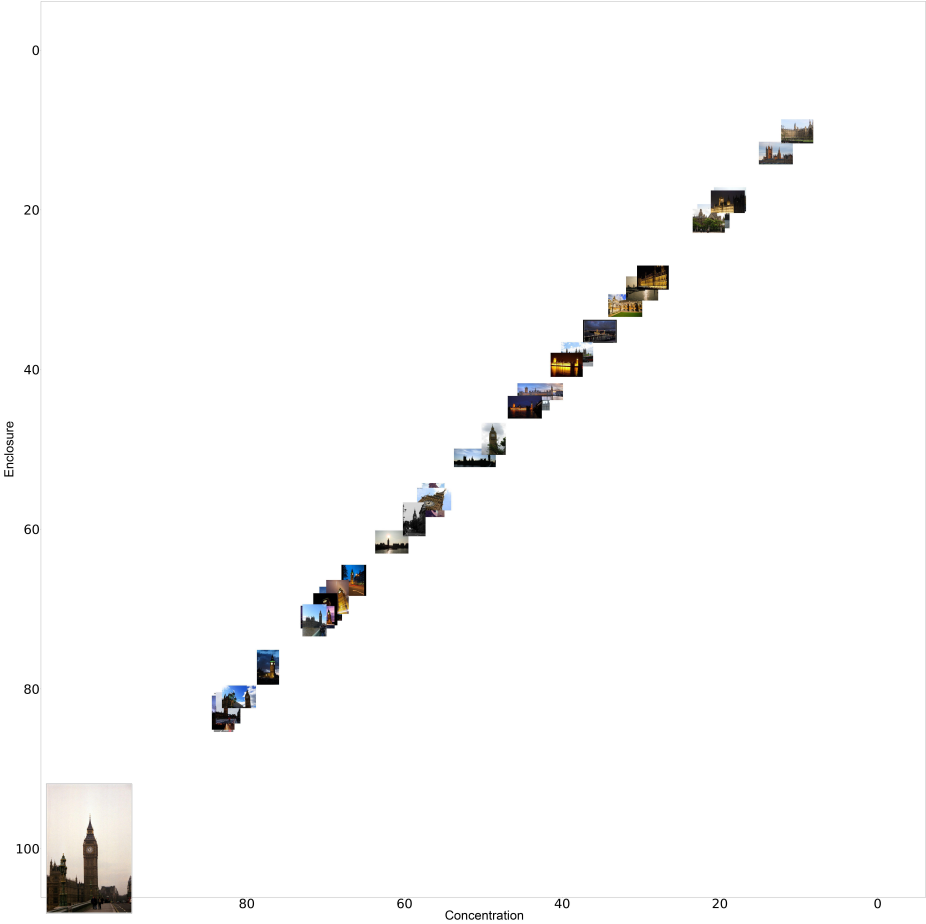**Fig. 15.** Generalization plot: Florence.

Fig. 16. Generalization plot: Notre Dame.

We show a query from the test set and report the predicted symmetric normalized surface overlap on a subset of test images. Because the embedding space measure is symmetric concentration and enclosure are equal for a given image. It can be observed that nearby images show similar views on the scene. However, the distance between the retrievals is not interpretable.

Fig. 17. Generalization plot: Big Ben NSO^{sym}.

87 % / 29%

97% / 33 %

98 % / 5%

65% / 30 %

83 % / 5%

73% / 14 %

79 % / 5%

76% / 27 %

82 % / 5%

93% / 39 %
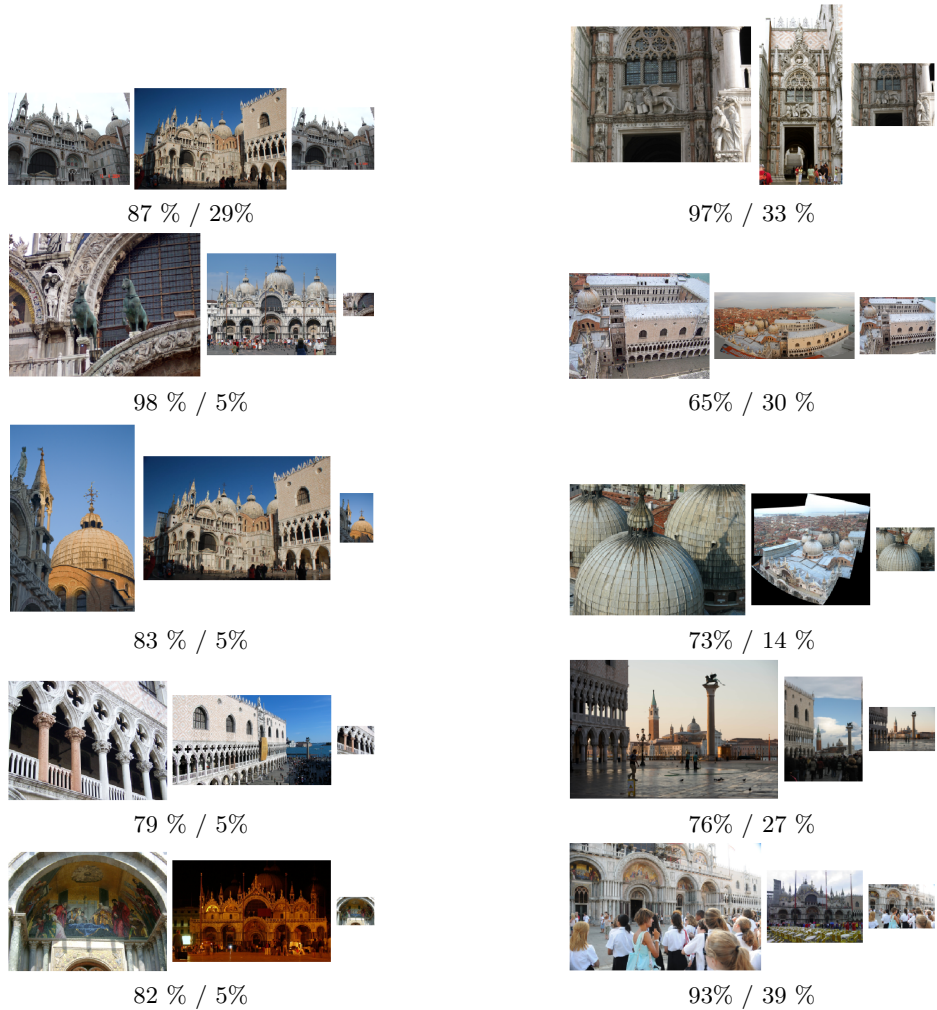
Illustrated are several examples of how our method can estimate geometric relationships between images. For each pair the enclosure and concentration are calculated form which the relative estimated scaled can be derived. Based on that scale, the first image is resized and shown in the third position. The resized images match the scale of the scene in the first image to the scale in the second image. Note, that the resized images are sometimes very small, and the reader is encouraged to zoom into the images. The two numbers below each image pair show the estimated enclosure and concentration. Note that although some scale estimates are inaccurate, overwhelmingly the rescaling does not increase the scale difference between the two images, but only reduces it. *This caption applies to all Relative scale plots.*

Fig. 18. Relative scale plot: Venice No. 1.

95 % / 27%

69% / 18 %

69 % / 30%

94% / 30 %

63 % / 11%

93% / 9 %

95 % / 13%

95% / 36 %

88 % / 9%

85% / 29 %

Fig. 19. Relative scale plot: Venice No 2.

84% / 10%   80% / 7%

Fig. 20. Relative scale plot: Unsuccessful cases for Venice scene (test image pairs here were found by querying the database for images that had enclosure > 0.6 and 0.05 < concentration < 0.4).

90 % / 37%

93% / 38 %

91 % / 33%

88% / 38 %

95 % / 26%

95% / 24 %

87 % / 5%

95% / 31 %

92 % / 39%

96% / 38 %

Fig. 21. Relative scale plot: Big Ben No. 1.

95 % / 24%

92 % / 24%

89 % / 21%

89 % / 21%

95 % / 24%

95% / 24 %

87% / 5 %

95% / 32 %

96% / 38 %

97% / 28 %
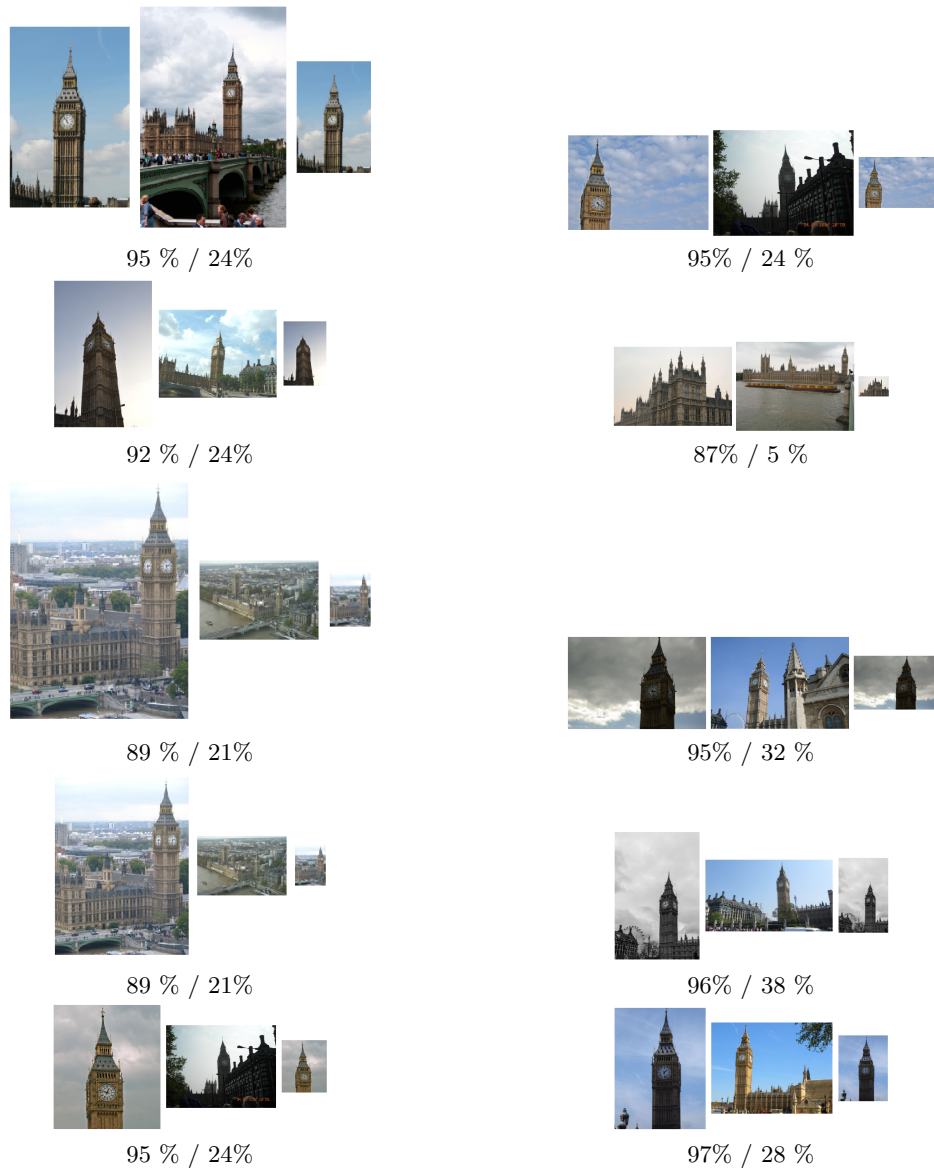
Fig. 22. Relative scale plot: Big Ben No. 2.

90 % / 12%

88% / 39 %

89 % / 34%

93% / 39 %

91 % / 37%

94% / 38 %

89 % / 36%

91 % / 40%

Fig. 23. Relative scale plot: Less successful cases Big Ben (18 out of 95 pairs shown in the document).

88 % / 38%

91% / 39 %

82 % / 26%

81% / 5 %

92 % / 6%

87% / 32 %

71 % / 6%

70% / 10 %

81 % / 37%

91% / 39 %
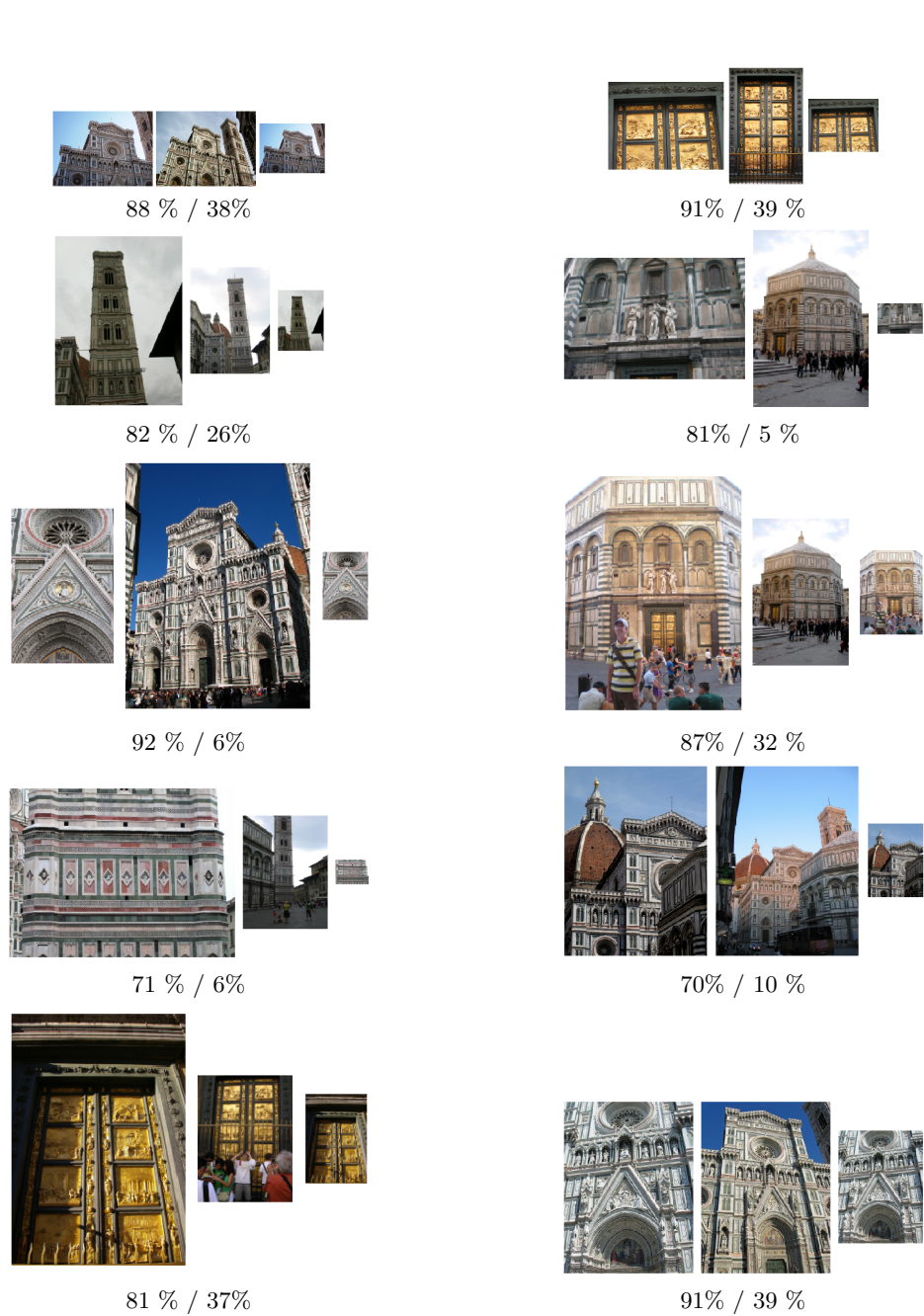
Fig. 24. Relative scale plot: Florence No. 1.

65 % / 32%

81% / 25 %

69 % / 22%

64% / 33 %

88 % / 36%
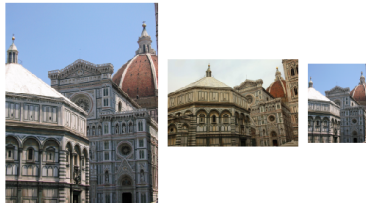
66% / 38 %

82 % / 34%

74% / 16 %

74 % / 37%
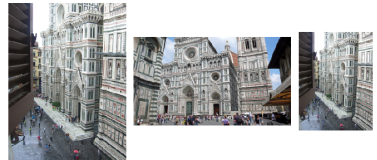
83% / 37 %

Fig. 25. Relative scale plot: Florence No. 2.

82 % / 34%

89 % / 29%
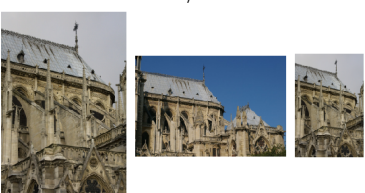
87% / 32 %

81% / 37 %

88 % / 36%

80% /39 %

82 % / 26%
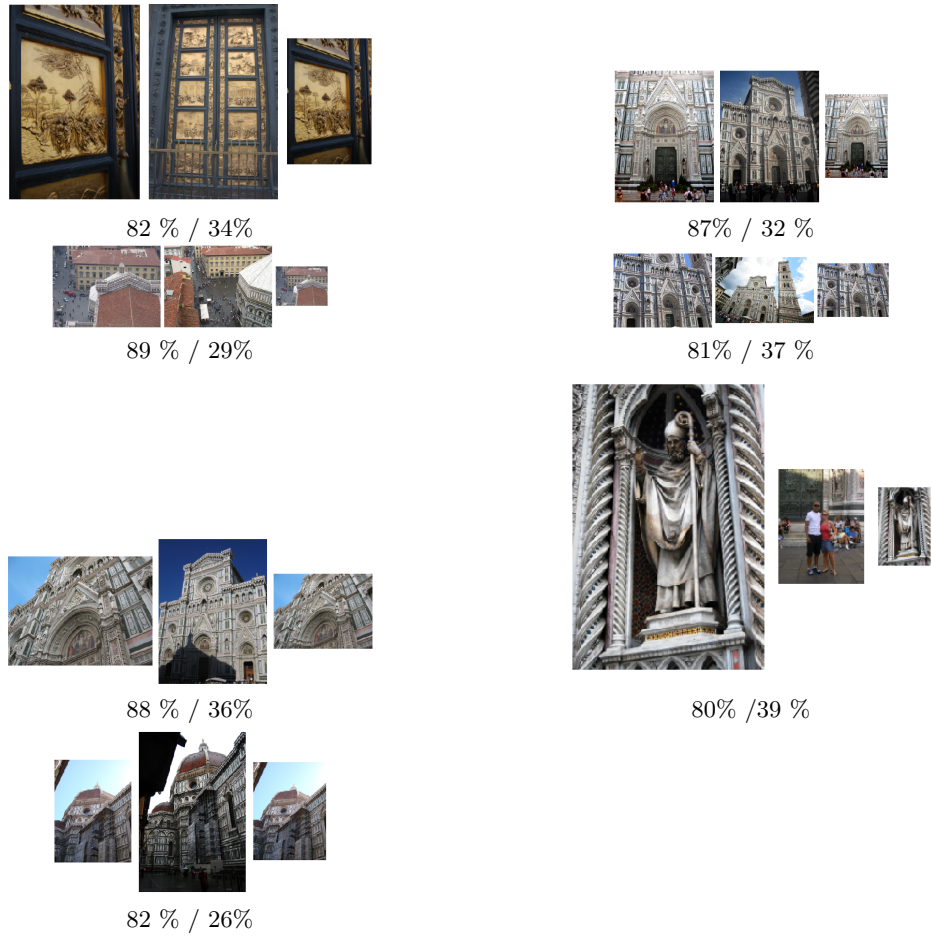
Fig. 26. Relative scale plot: Less successful cases Florence (7 out of 57 pairs shown in the document).

91 % /30%
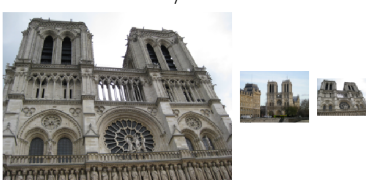
80% / 5 %

91 % / 5%

93% / 37 %

77 % / 36%

81% / 15 %

85 % / 39%

94% / 34 %

94 % / 5%

77% / 39 %

Fig. 27. Relative scale plot: Notre Dame No. 1.

92 % / 10%

68% / 31 %

96 % / 28%

89% / 8 %

87 % / 14%

78% / 17 %

78 % / 13%

71% / 37 %

83 % / 38%

87% / 9 %

Fig. 28. Relative scale plot: Notre Dame No. 2.

64 % / 37%

83% / 35 %

81 % / 39%

85% / 39 %

60 % / 35%

84% / 39 %

79 % / 38%

73% / 39 %
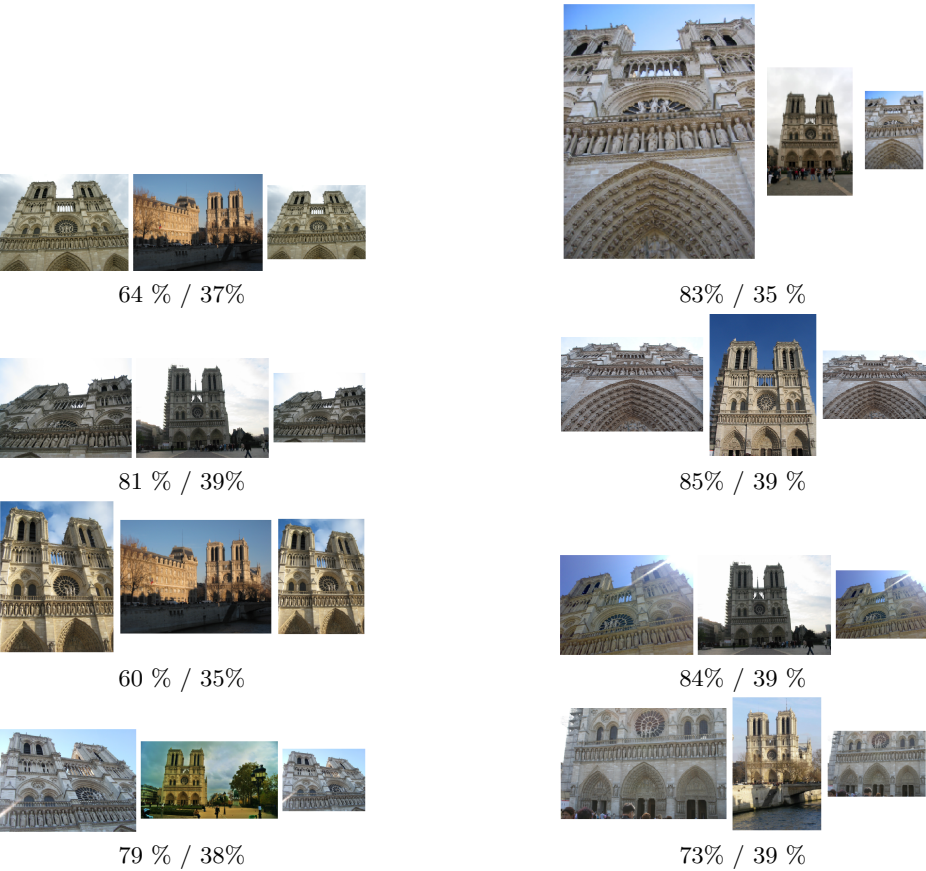
Fig. 29. Relative scale plot: Less successful cases Notre Dame (11 out of 61 pairs shown in the document).