Supplementary Material

Learned Contextual Feature Reweighting for Image Geo-Localization

Hyo Jin Kim
UNC Chapel Hill
hyojin@cs.unc.edu

Enrique Dunn
Stevens Institute of Technology
edunn@stevens.edu

Jan-Michael Frahm
UNC Chapel Hill
jmf@cs.unc.edu

1. Overview

We qualitatively show how feature weights change adaptively based on the context in the proposed image representation as detailed in Section 2 (Fig. 1 and Fig. 2). Additionally, we show examples of retrieval results on the San Francisco dataset [2] (Section 3), and retrieval results on the Oxford Buildings 105k dataset [4] (Section 4) that were not included in the paper due to space constraints.

2. Contextual Feature Reweighting

To show how our Contextual Reweighting Network (CRN) adaptively weights features based on the context, we generated contextual reweighting masks on synthetic images. Figure 1 (left) shows an example where features on the signage on a store front are assigned a comparably high weight by our CRN within the context of its original image. It can be observed that the high weighting is caused by the signage as the weight diminishes when removing some of the letters on the signage as shown in Figure 1 (right). To see how CRN changes weights on a feature from one context to another, we cropped out some image patches containing the letters on the signage on Figure 1 and placed them on other images such that they are surrounded by different visual elements such as pedestrians (Figure 2 (a-b,d)), vehicles (Figure 2 (c,h)), vegetation (Figure 2 (e)), and sky (Figure 2 (f)). We directly overlayed the image patch without resizing. It can be observed that the letters on the signage are no longer assigned high weights as their surroundings changed. For Figure 2 (i-j), we pasted the store signs that are from the same image. The results are generated from our AlexNet [3]-based model.

3. Retrieval Result: Geo-Localization

More examples of retrieval results on the San Francisco benchmark [2] for image geo-localization using our context-aware image representation are depicted in Figure 3 and Figure 4, each using our AlexNet [3] and VGG16 [5] based models. The top 5 retrieved images are shown for each query image. The results of NetVLAD [1] trained on San Francisco in the same pipeline as ours (with the same base architecture) are also shown for comparison.

4. Retrieval Result: Oxford Buildings 105K

In Section 4.4 of the paper, we reported the image retrieval performance of our image representation trained on San Francisco on the Oxford Buildings 105K dataset [4]. The examples of the top 20 retrieved images are shown in Figure 5 and Figure 6. The average precisions (AP) are shown below each query image.

References

[1] R. Arandjelović and A. Zisserman. DisLocation: Scalable descriptor distinctiveness for location recognition. In *ACCV*, 2014. 1, 3, 4

[2] D. Chen, G. Baatz, K. Koser, S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, et al. City-scale landmark identification on mobile devices. In *CVPR*, 2011. 1

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[4] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007. 1

[5] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 1

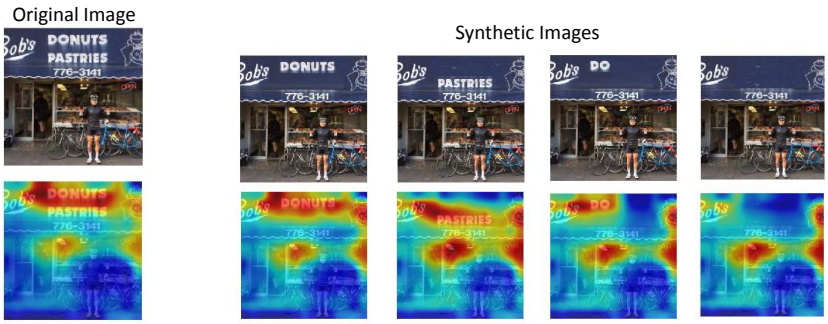

Figure 1. High weights are assigned on features from the signage on a store front. As we removed the letters on the signage, the weights diminish. (top) Input image. (bottom) Generated contextual reweighting mask in a heat map (red: high, blue: low).

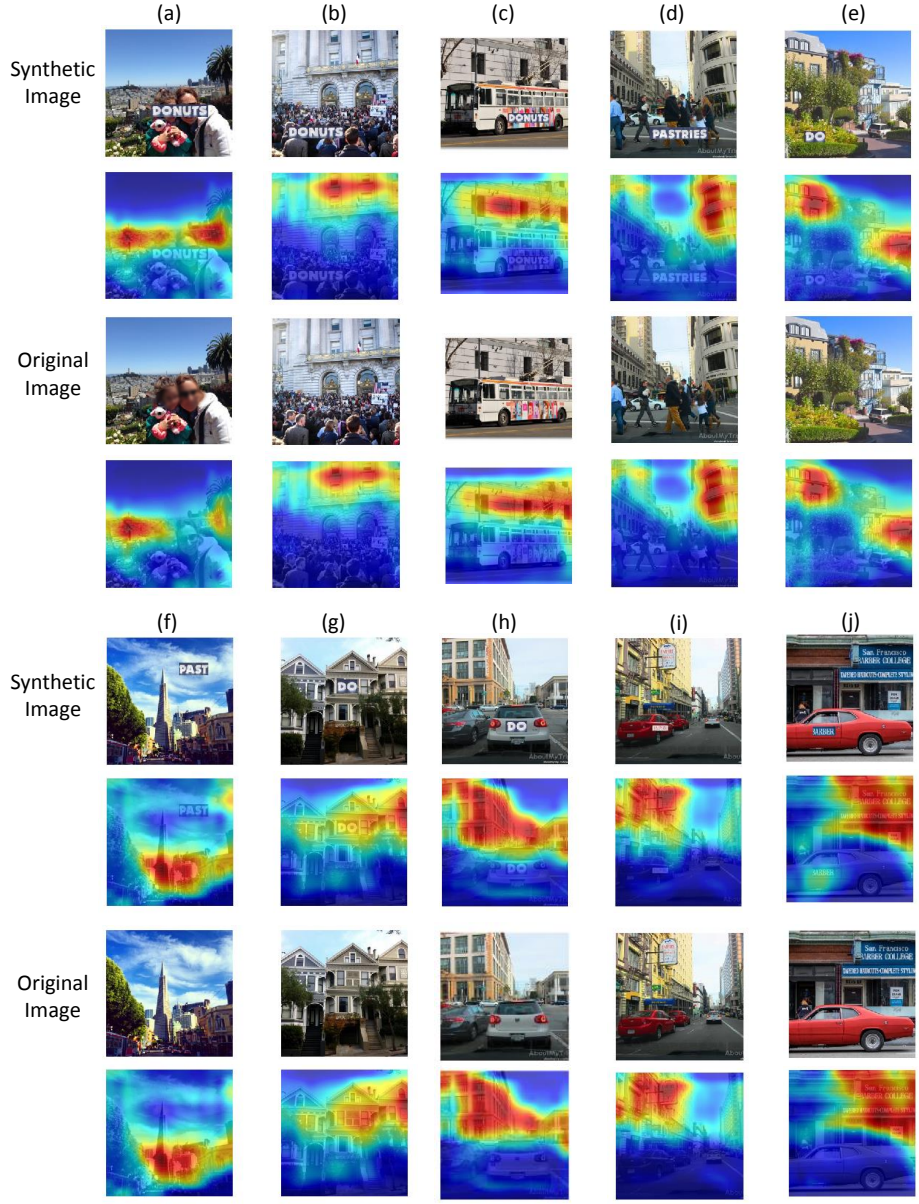

Figure 2. We generated synthetic images by pasting image patches containing the letters of the signage from Figure 1 that was assigned high weights at the store front (a-h). For (i)-(j), we overlayed the store signages from the same image on vehicles. Generated contextual reweighting masks are visualized on the bottom of each image as a heat map (red: high, blue: low). The letters from the signagage are no longer assigned high weights as the surrounding contexts have changed.

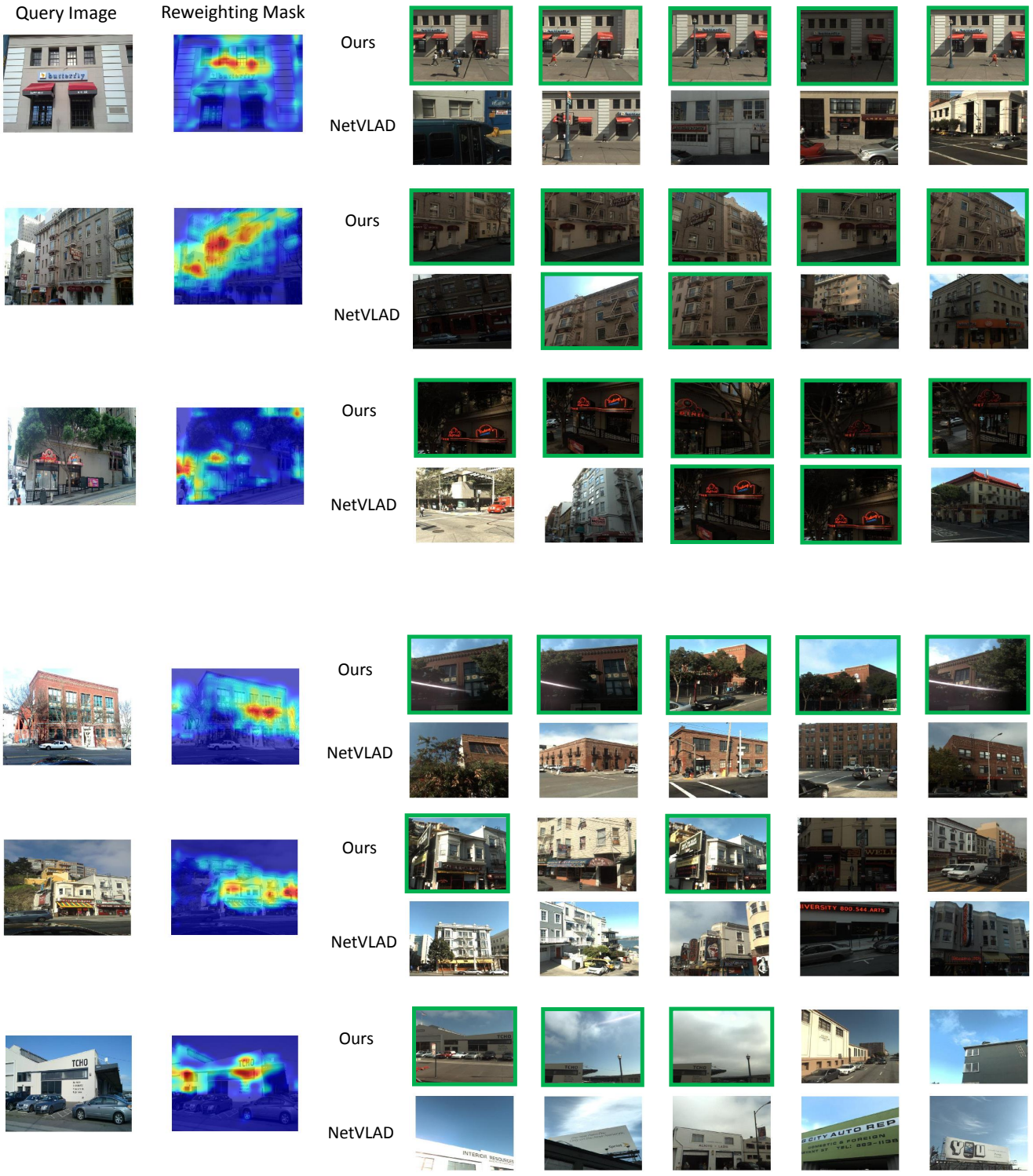

Figure 3. Image geo-localization results. (left) Query images and the corresponding contextual reweighting masks generated by our CRN as heat maps, (right) Top five retrieved images using our method and NetVLAD [1]. The green boxes around the retrieved images denote the correct results. The results are based on our AlexNet-based model.

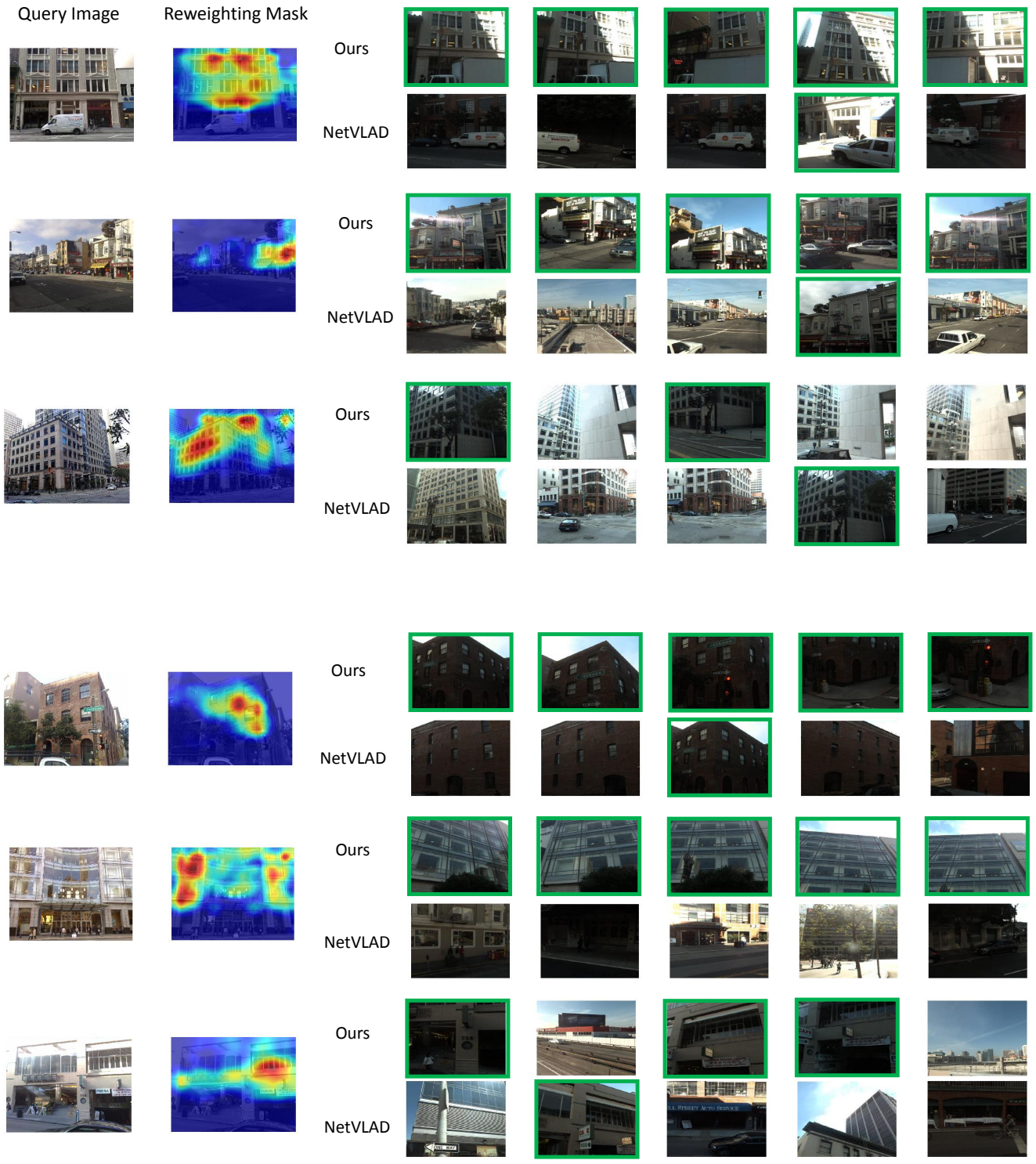

Figure 4. Image geo-localization results. (left) Query images and the corresponding contextual reweighting masks generated by our CRN as heat maps, (right) Top five retrieved images using our method and NetVLAD [1]. The green boxes around the retrieved images denote the correct results. The results are based on our VGG16-based model.

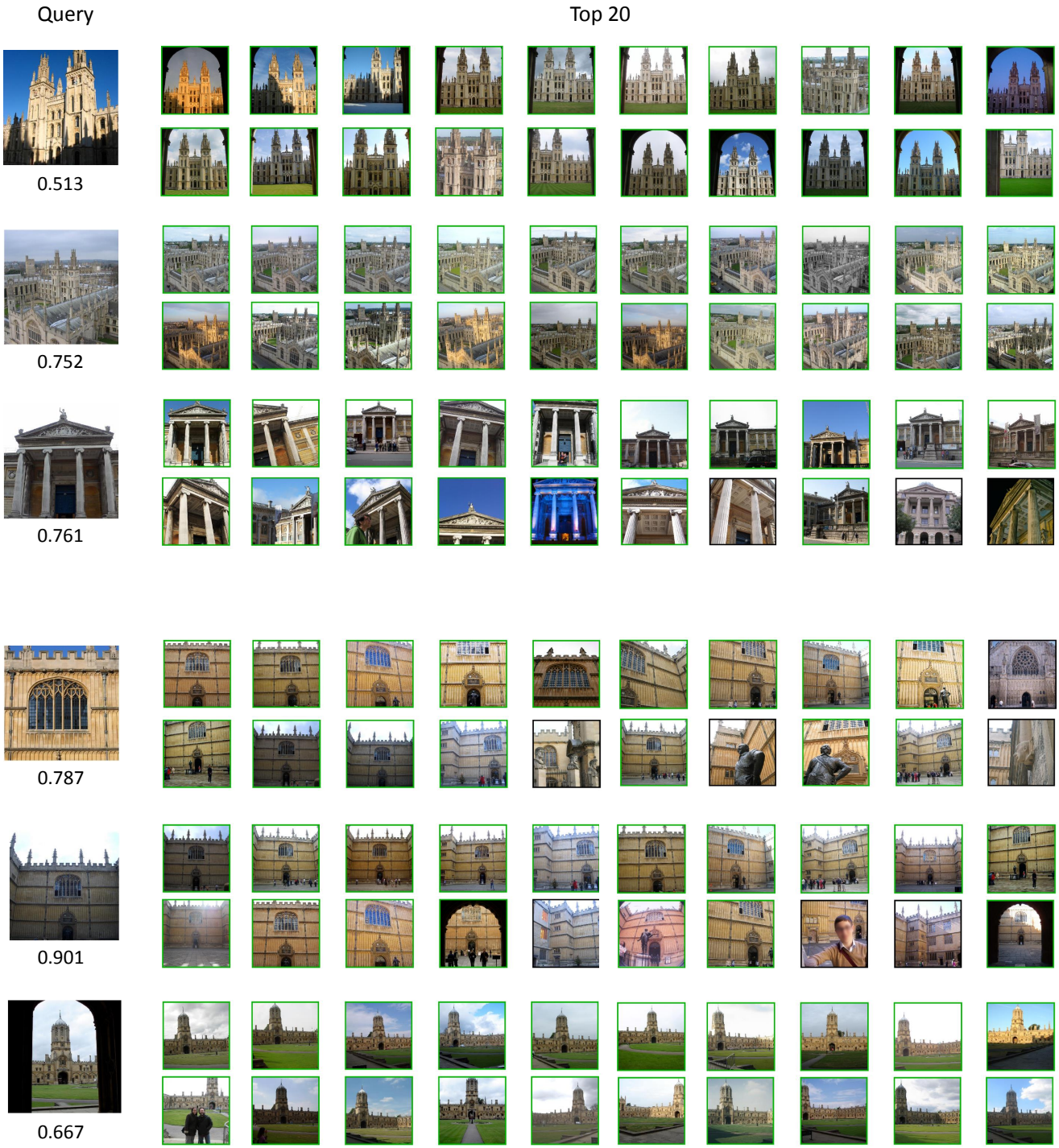

Figure 5. Image retrieval results from the Oxford Buildings 105k dataset. (left) Query images and average precisions (AP) by our method. (right) Top twenty retrieved images using our image representation, where the image with the highest similarity score is shown on the top left. The green boxes around the retrieved images denote the correct retrieval results.

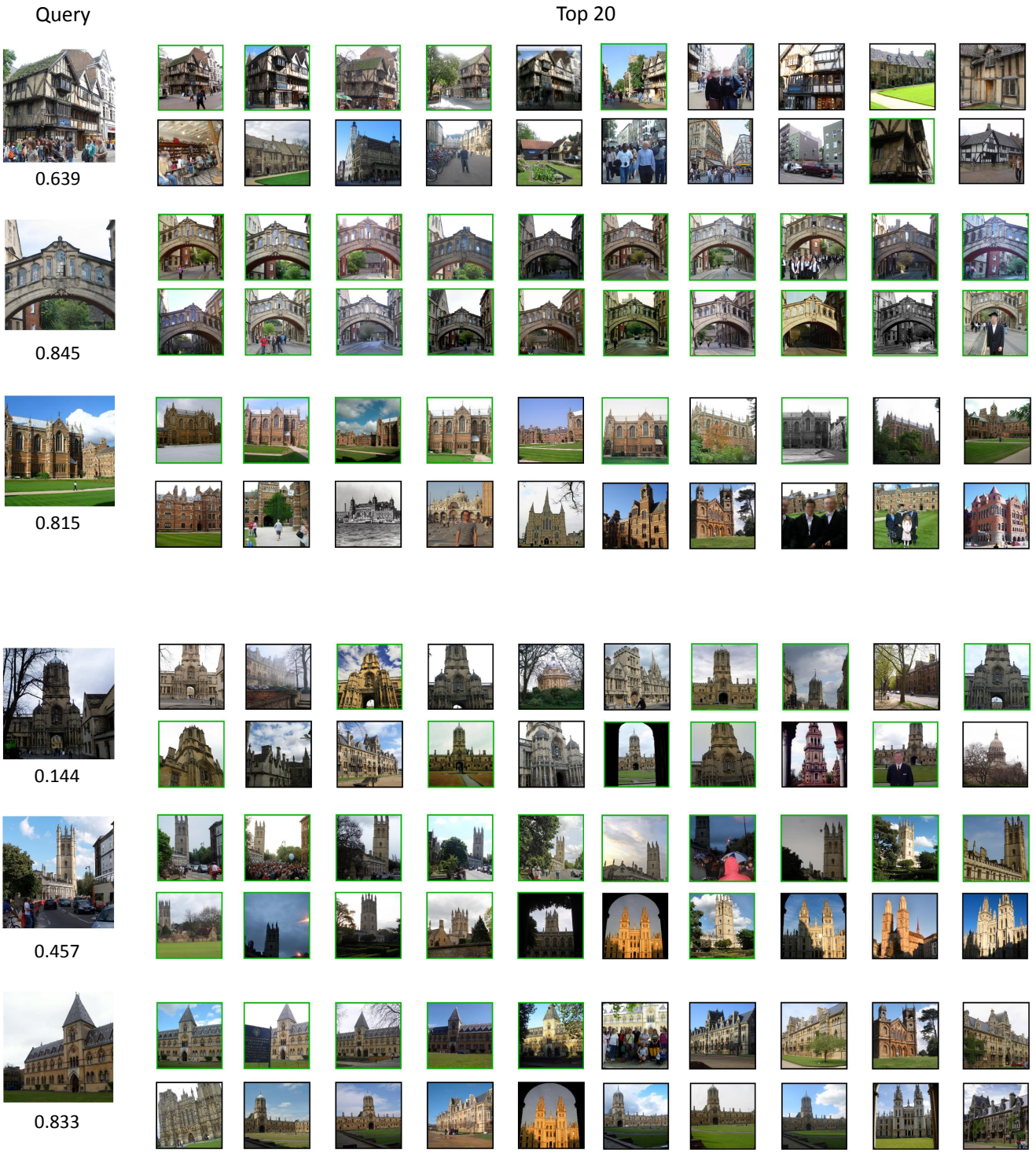

Figure 6. More examples of image retrieval results from the Oxford Buildings 105k dataset. (left) Query images and average precisions (AP) by our method. (right) Top twenty retrieved images using our image representation, where the image with the highest similarity score is shown on the top left. The green boxes around the retrieved images denote the correct retrieval results.