Estimating Extreme 3D Image Rotations using Cascaded Attention

Supplementary Material

1. Qualitative experimental comparisons

The qualitative results of the rotation estimation are shown in Fig. 1 for the StreetLearn and SUN360 datasets, for the large, small and nonoverlapping cases. We show the full panoramas, the footprints of the cropped images that were used as inputs for the proposed scheme and the footprint of the estimated image crop based on the estimated rotation. In all cases. we archive high estimation accuracy.

2. Attention Maps Visualization

As described in Section 1, our scheme, in conjunction with Cai et al. [1], takes advantage of rotation-informative image signals by assigning them high attention scores. These cues can be both implicit (e.g. shadows, straight lines, analytic shapes, lighting angles) and explicit (e.g. corners used for feature correspondences). The visualization of these attention scores is shown in Fig. 2, where we have overlayed the attention scores on the input images following the approach of Kolesnikov et al. [5]. We applied our model to each pair of images, obtained the activation maps, and superimposed them on the input images. The attention maps emphasize vertical and horizontal lines for the overlapping and non-overlapping image pairs. Our experimental datasets, consist of interior and urban outdoor scenes showing horizontal and vertical lines. This relates to the seminal single-image camera calibration approaches of Coughlan and Yuille [2] and Criminisi et al. [3] and their extensions, which detect parallel lines in an image to estimate vanishing points [4] and relative rotations [2]. Therefore, we postulate this shows that our approach implicitly solves the rotation estimation the same as [2] and [3], by learning both the low-level informative task-specific image cues and the higher-level algorithmic computational flow.

3. Architectural Ablations

These are the network configurations corresponding to the architectural ablations shown in Table 6 of our paper.

References

[1] Ruojin Cai, Bharath Hariharan, Noah Snavely, and Hadar Averbuch-Elor. Extreme rotation estimation using dense correlation volumes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[2] J.M. Coughlan and A.L. Yuille. Manhattan world: compass direction from a single image by bayesian inference. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 941–947 vol.2, 1999. 1

[3] Antonio Criminisi, Ian Reid, and Andrew Zisserman. Single view metrology. In *IEEE International Conference on Computer Vision (ICCV)*, pages 434–441. IEEE Computer Society, 1999. 1

[4] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 1

[5] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1

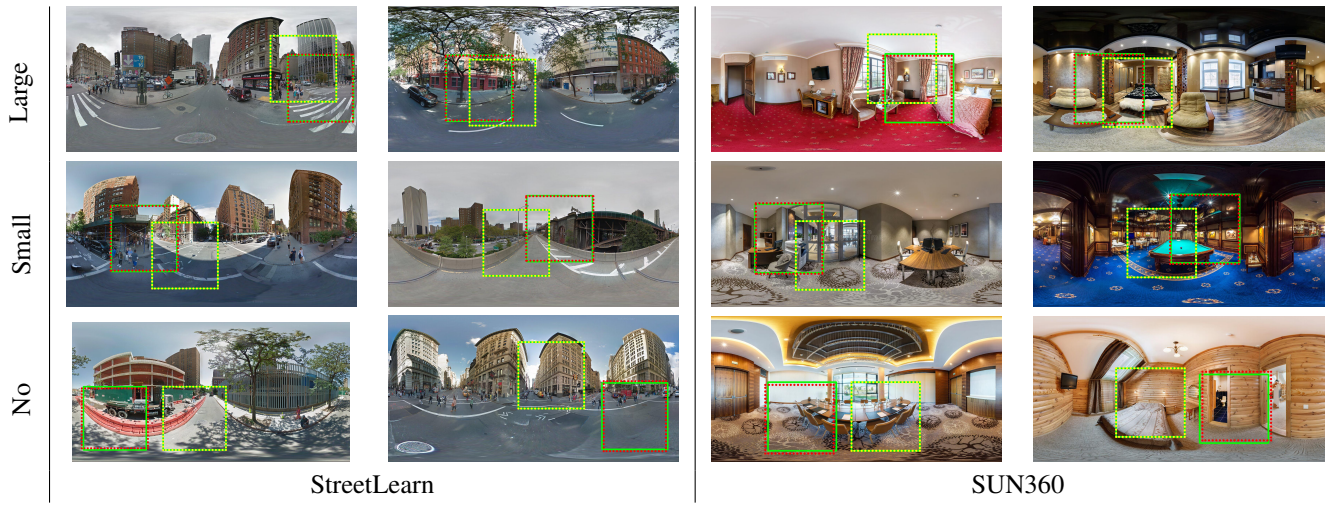

Figure 1. Rotation estimation results. The panoramic and cropped groundtruth images are marked by green and yellow dots. The predicted footprint of one of the cropped images is marked by the red-dotted line. The first row shows the results of the matching of images with large overlaps. The second and last rows show the matching of small overlap and non-overlapping images.

Figure 2. Visualizations of the attention maps. The attention maps generated by our model are overlaid on the image pairs for the overlapping (top) and non-overlapping (bottom) pairs. The input crops are represented by solid lines, and the crops are aligned with the estimated rotation as indicated by the dotted lines.

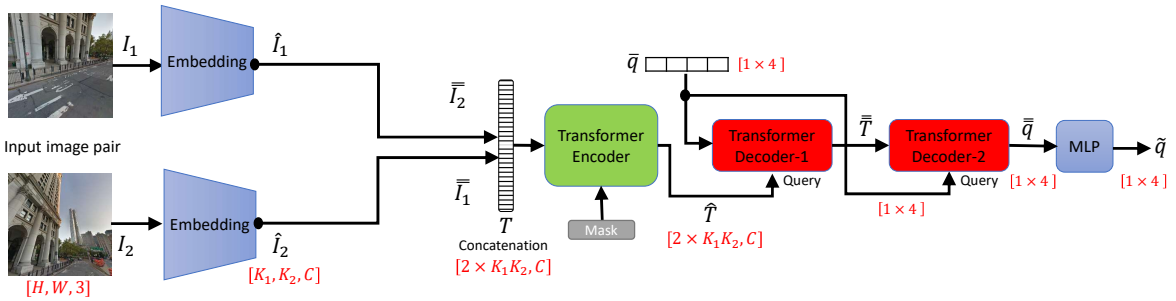

Figure 3. Architectural Ablation #1: without Transformer-Decoder-0.

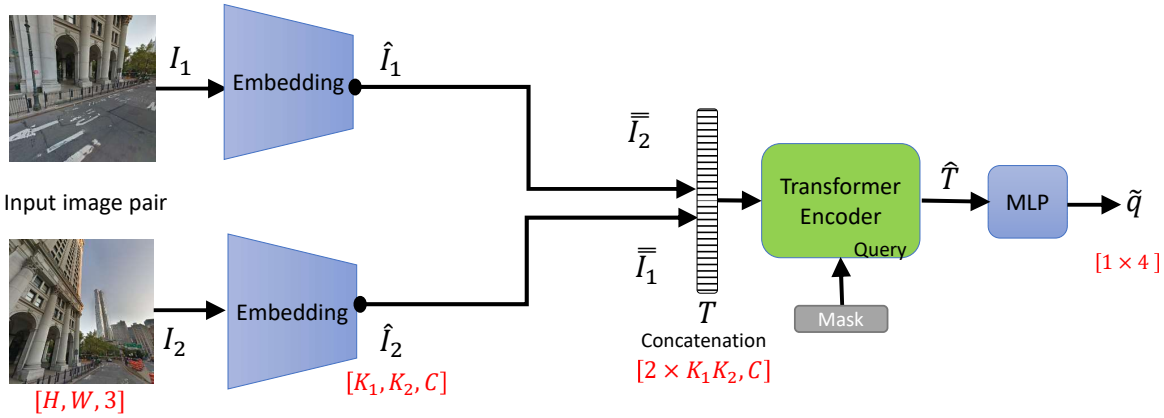

Figure 4. Architectural Ablation #2: without Transformer-Decoder-0, Transformer-Decoder-1 and Transformer-Decoder-2.

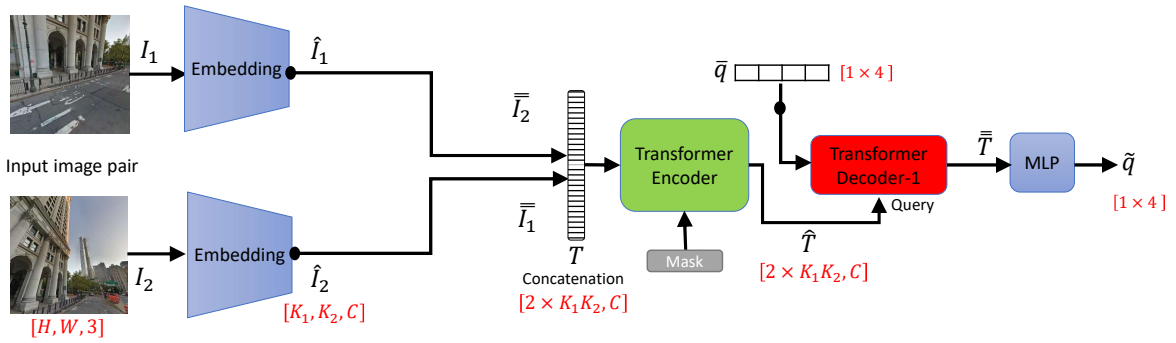

Figure 5. Architectural Ablation #3: without Transformer-Decoder-0 and Transformer-Decoder-2.

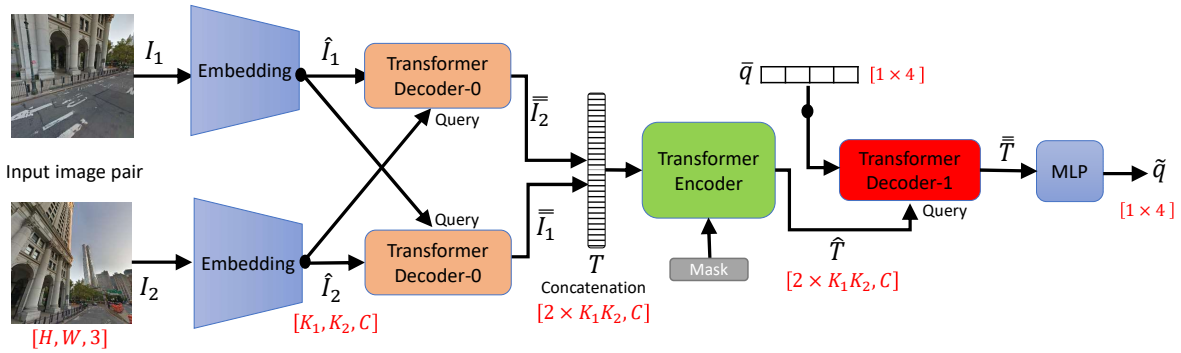

Figure 6. Architectural Ablation #4: without Transformer-Decoder-2.

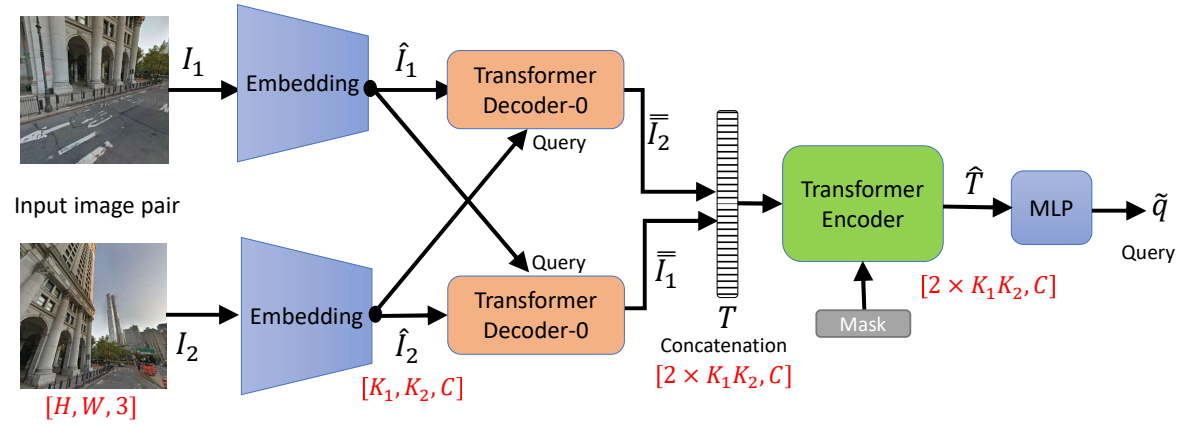

Figure 7. Architectural Ablation #5: without Transformer-Decoder-1 and Transformer-Decoder-2.