6. Beyond Cartesian Representations for Local Descriptors: Supplementary Material

6.1. Regarding the Dataset

In order to train scale-invariant models with real data relevant to wide-baseline stereo, it was necessary to collect training data. For this we rely on public collections of photo-tourism images in the Yahoo Flickr Creative Commons 100M (YFCC) dataset. We use COLMAP, a Structure from Motion (SfM) framework, to obtain 3D reconstructions. COLMAP provides us with sparse point clouds and depth maps for each image. We clean up the depth maps following the procedure outlined in the paper and use them, along with the ground truth camera poses, to project keypoints between corresponding images.

We sample pairs of images with a visibility check in order to guarantee that a minimum number of keypoints can be extracted and matched across both views. Specifically, we retrieve the SfM keypoints in common over both views, extract their bounding box, and reject the image pair if it is smaller than a given threshold (we use 0.5) for either image.

We use 11 sequences for training and validation and 9 for testing. We list their details in Table 6, and give some examples in Fig. 6. This data will be made publicly available along with code and pre-trained models.

Sequence name	Num. images
brandenburg_gate	1363
buckingham_palace	1676
colosseum_exterior	2063
grand_place_brussels	1083
notre_dame_front_facade	3765
palace_of_westminster	983
pantheon_exterior	1401
sacre_coeur	1179
st_peters_square	2504
taj_mahal	1312
temple_nara_japan	904
Total	18233

Sequence name	Num. images
british_museum	660
florence_cathedral_side	108
lincoln_memorial_statue	850
milan_cathedral	124
mount_rushmore	138
reichstag	75
sagrada_familia	401
st_pauls_cathedral	615
united_states_capitol	258
Total	4107

Table 6: **Dataset details.** Left: training sequences. Right: Test sequences.

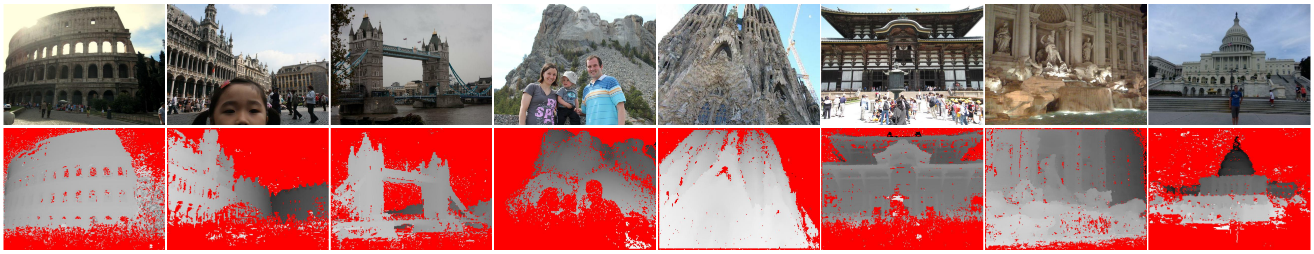

Figure 6: **Dataset samples.** We show the original images and their corresponding depth maps, estimated by COLMAP and post-processed by us as explained in Section 4.1.1. The depth maps are color-coded by depth, in grayscale, with red indicating occlusions and regions for which depth estimates are not available. Notice how despite some noise and occluded areas, the depth estimates are good enough to extract training data.