Supplemental Material

1. Implementation

Here we will provide some more detail about our implementation. We use Caffe framework [14]. We use learning rate 0.01 and reduce it several time during the training, to 0.00001 (when the loss seems to stop improving). Minibatch size is 32, momentum is 0.9 and weight decay factor is 0.0005.

We use VGG trained on ImageNet [29] as initialization and train a network with the 1060 ways classification for 500k iterations. Then we use this network as initialization for training every other networks (usually just another 100k-200k iterations), we found that this speed up the experiment quite a lot since training every model from scratch or ImageNet initialization take much more time. As shown in Table 3, the pretrained ImageNet model ([I]) can be also be used for retrieval, but not as effective as a model trained for geolocalization task ([L]).

When training with multiple losses, the overall loss will be the weighted sum of all the losses. For [M] model, we use the same weight (1) for all 6 losses.

2. Feature visualization

We show a t-SNE visualization in Figure 9. The feature learnt from GPS supervision seems to be very high level; there's many regions in this visualization with consistent theme such as: sport scene images, people images, beach and sunset images, animal images, landmark type of architecture images, etc. There's a large variety in image appearance within a region.

In Figure 10 we look at some dimensions in the output feature space and show the images whose has a high corresponding feature value. Few activation outputs do correspond to some particularly popular landmarks/architecture; while many correspond to certain type of scene or visual features. Some seems to respond to more than one visual features and some might roughly represent higher level location-based semantics. For example row 5 shows pictures of Disney-like castle and Disney's Mickey mouse even-though they are not visually similar.

We show some more nearest neighbors example result in Figure 11.

Table 3. Performance on Im2GPS3k test set.

Method	Model	Stre.	City	Reg.	Cou.	Cont.
NN	[I]	7.4	17.0	19.6	26.8	41.9
	[L]	7.5	18.9	23.5	32.6	49.5
kNN,σ=1	[I]	7.5	18.3	22.5	30.2	45.8
	[L]	7.8	20.9	27.1	36.8	53.8
kNN,σ=4	[I]	7.0	16.8	22.1	31.9	48.7
	[L]	7.2	19.4	26.9	38.9	55.9
kNN,σ=16	[I]	4.4	10.6	15.4	32.2	51.2
	[L]	5.3	13.8	21.2	39.9	58.9

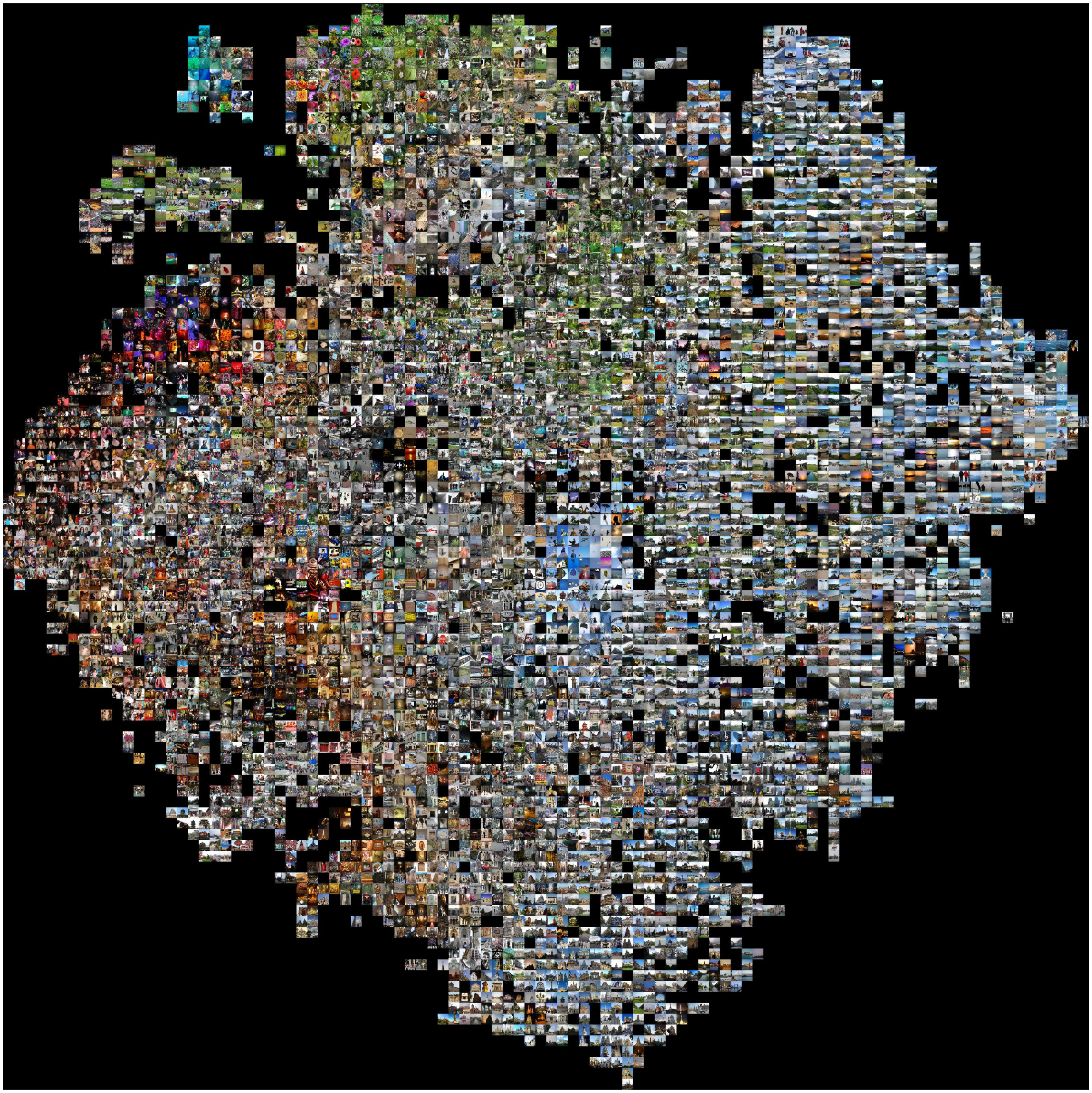

Figure 9. t-SNE visualization

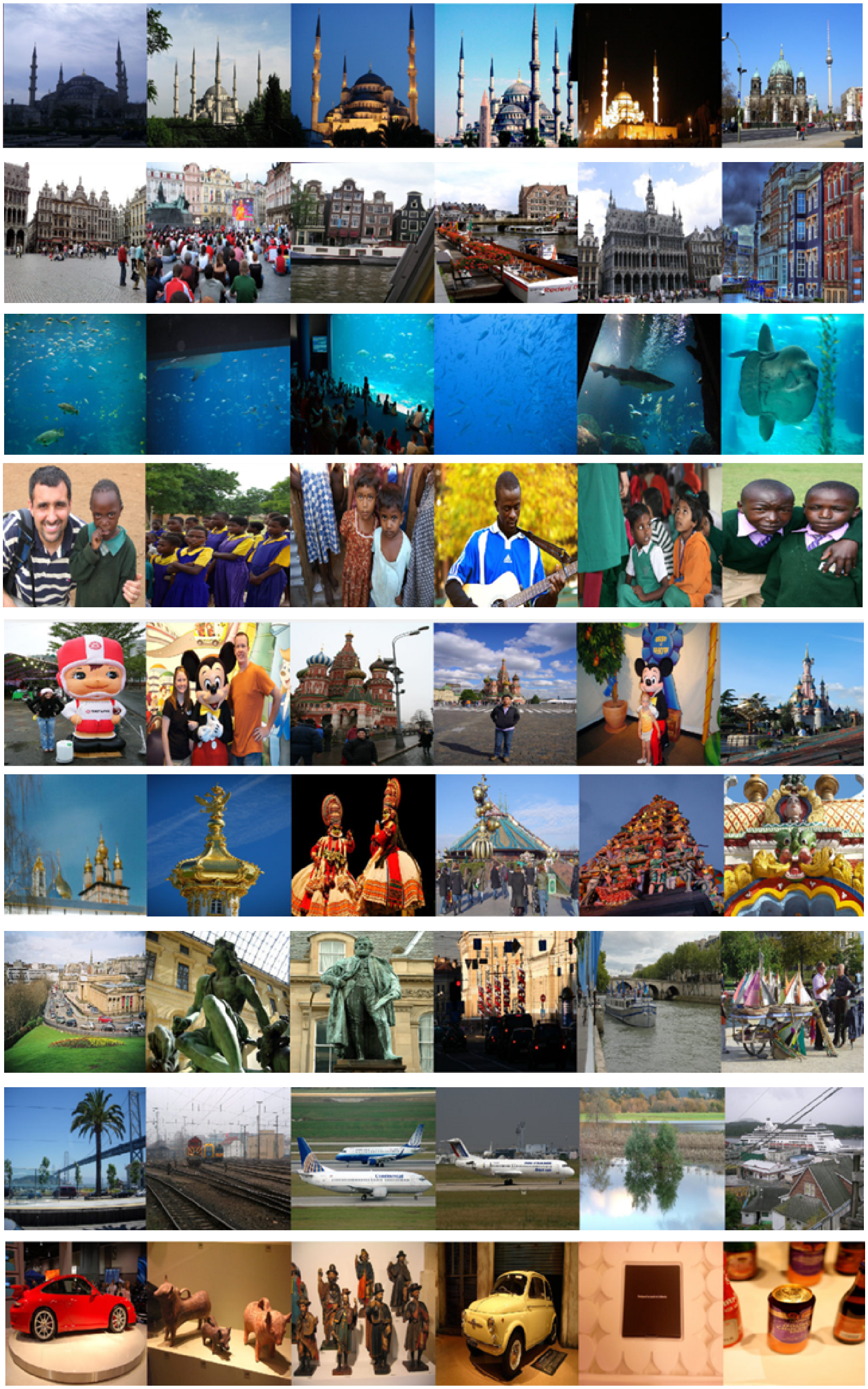

Figure 10. Each row shows a set of images whose feature has a high value at a particular activiation unit (last layer).

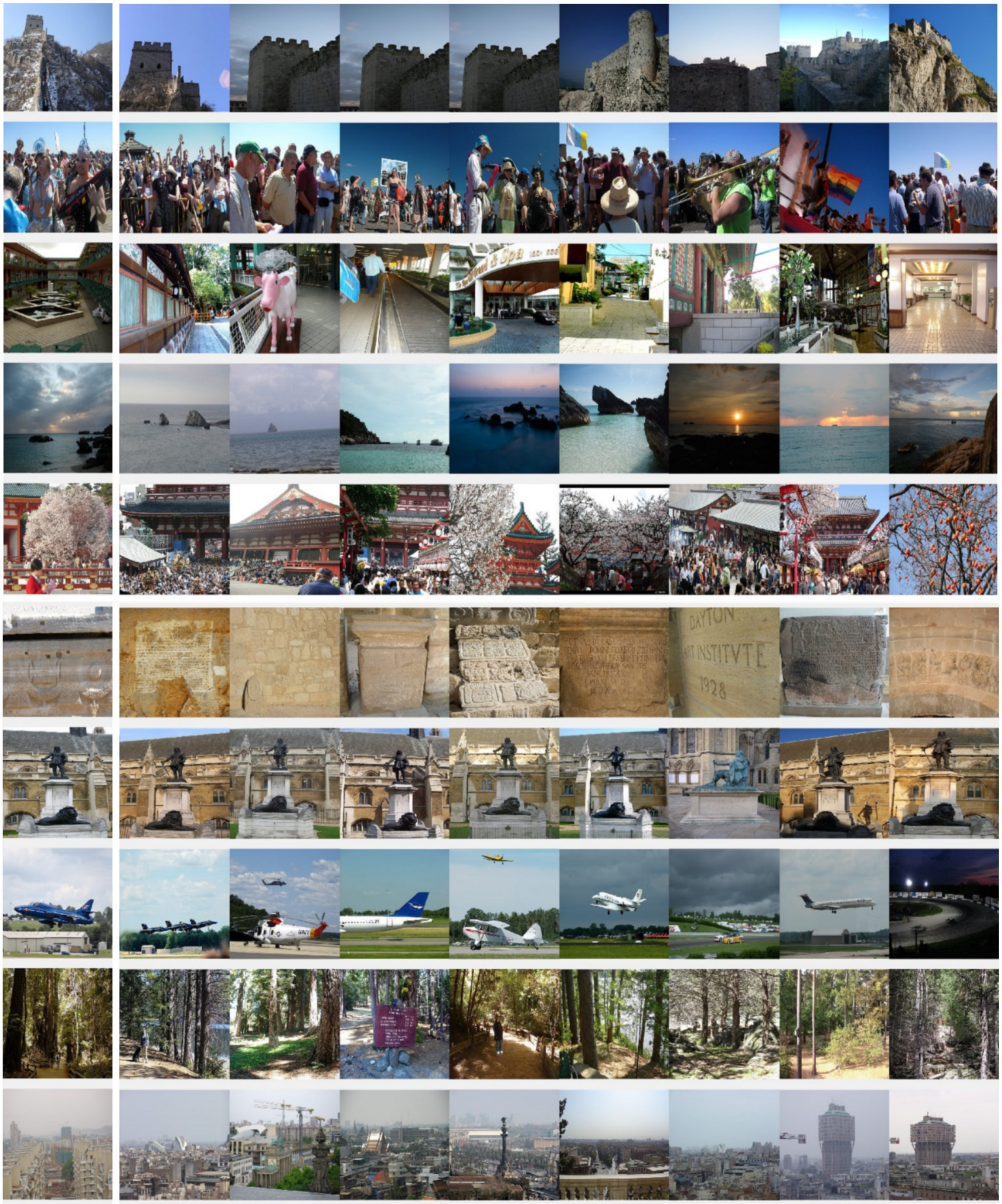

Figure 11. Some qualitative near neighbors result: the images on the left column are query, the other on the same row are its NNs.