Supplementary Material Sample4Geo: Hard Negative Sampling For Cross-View Geo-Localisation

1. Hyperparameter

In order to determine the best hyperparameter for our Dynamic Similarity Sampling (DSS) we choose different k on the VIGOR SAME split during sampling, where k is the actually selected amount of neighbours from the nearest neighbour pool of size K. As can be seen in table 1 our approach performs best with $k = 64$ and also does not collapse to simple solutions, when all hard negatives are sampled if $k = K$ without a random factor.

Neighbours k	R@1	R@5	R@10	R@1%	Hit Rate
16	76.84	95.24	96.98	99.68	88.55
32	77.39	95.61	97.19	99.65	89.46
64	**77.86**	**95.66**	**97.21**	**99.61**	**89.82**
128	77.84	95.42	97.07	99.58	89.69

Table 1: Comparison between the different neighbours we select during our similarity sampling on the VIGOR same split, with pool size $K = 128$. Based on this analysis we decided to use $k = 64$ for all our experiments.

2. Symmetric InfoNCE Loss

In the analysis of our approach, we also examined to what extent the symmetrical InfoNCE loss contributes to the performance. As shown in table 2, the direction of the loss calculation plays a significant role. The performance suffers, when the similarity of the satellite-view as query is calculated against the street-view as reference. Whereas if we use the street-view as query and the satellite-view as reference, the performance is almost on par with the symmetric loss calculation in both directions. An explanation for this behaviour comes from the data in VIGOR and our similarity sampling. Since in our DSS we are looking for street-view images whose distance is minimal to satellite images, the other direction cannot benefit from our sampling. To avoid this and to benefit from both directions, we formulated the loss symmetrically to achieve the best performance.

Loss Direction	R@1	R@5	R@10	R@1%	Hit Rate
Sat $\rightarrow$ Street	74.69	93.80	95.81	99.40	86.27
Street $\rightarrow$ Sat	77.49	95.66	97.18	99.58	89.57
Street $\leftrightarrow$ Sat	**77.86**	**95.66**	**97.21**	**99.61**	**89.82**

Table 2: Comparison between unidirectional loss calculation and symmetric loss caluclation on the VIGOR same split. Our model profits from a symmetric loss the most.

3. Loss comparison

The triplet loss is very prone to model collapsing when using only hard negatives within a batch. A proposed extension to overcome this is the soft-margin triplet loss. In Table 3 we compare the two triplet losses with the InfoNCE loss. As we show the model collapses when using the triplet loss without extension.

Dataset	R@1	R@5	R@10	R@1%
Triplet Loss	0.00	0.06	0.10	1.26
Soft-Margin Triplet	91.83	97.87	98.75	99.67
InfoNCE	**98.68**	**99.68**	**99.78**	**99.87**

Table 3: **Loss function comparison for CVUSA.**

4. Sampling Strategies

We additionally perform the comparison of the sampling strategies on the CVUSA and CVACT dataset and come to similar results as for VIGOR. Our experiments show that regardless of the dataset, the combination of GPS sampling and DSS leads to the best results. Considering the two sampling strategies in isolation leads to different results, which strongly depend on the dataset. When the samples are from geographically close areas, as in CVACT (area around Canberraast) and VIGOR (four different cities), GPS sampling performs similarly to DSS. When there are a large number of distant locations, as in CVUSA, GPS sampling is no more advantageous than random sampling.

Sampling	R@1	R@5	R@10	R@1%
CVUSA				
Random	97.83	99.63	99.75	99.89
GPS	97.83	99.53	99.72	99.87
DSS	98.51	99.69	99.79	99.85
GPS + DSS	**98.68**	**99.68**	**99.78**	**99.87**
CVACT$_{\text{test}}$				
Random	60.57	89.50	92.99	98.92
GPS	71.13	90.28	92.47	98.09
DSS	71.04	92.26	94.33	98.58
GPS + DSS	**71.51**	**92.42**	**94.45**	**98.70**

Table 4: **Comparison of sampling for CVUSA/CVACT.**

5. Visualisation

To provide a better visualisation for the alignment of the activation maps we additionally apply a inverse polar transformation to the activation maps of the street view image. With this transformation it is easier to visualise the correspondence between both views more clearly.

5.1. Generalisation CVACT → CVUSA

As we have shown in our previous results, our model also performs well on new regions compared to previous work. However, this performance is lower as when the same region is used, were the model is trained on. For this transferability we provide some insights for a model trained on CVACT predicting on CVUSA. We used randomly selected false predictions and compare them with each other visually as well as by the cosine similarity. The results are shown in 1 for two examples.

5.2. Correct predictions

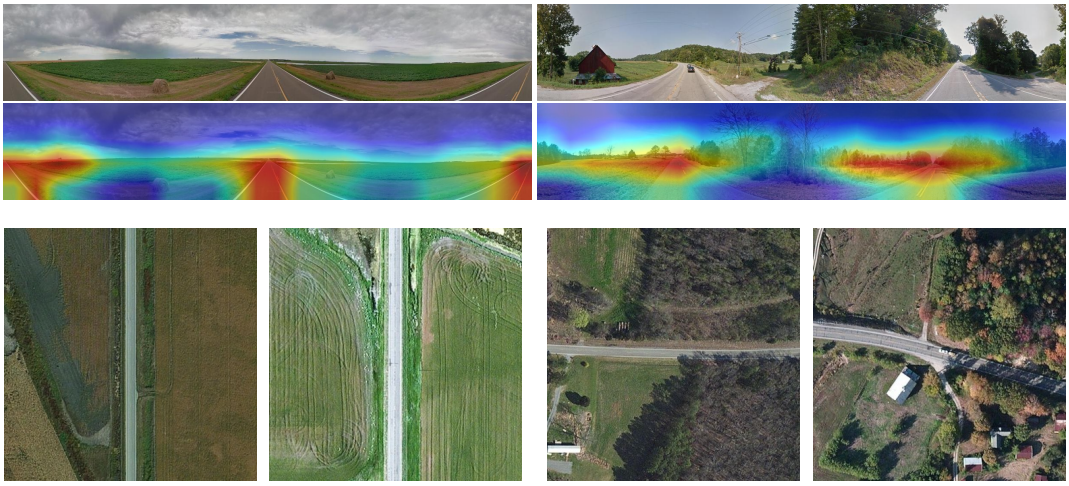
For VIGOR, our model has learned to pay attention particularly to vegetation and also on road markings, as shown in Figure 2 for a correct predicted sample.

We visualise the activation maps for some correct predictions from a in-domain trained model on CVUSA in figure 3. Here, the models mainly focus on the road features such as the course and the position of intersections.

VIGOR covers more urban locations in its training data compared to CVUSA, so in rural settings the road itself plays a more dominant role.

6. Acknowledgement

The authors gratefully acknowledge the computing time granted by the Institute for Distributed Intelligent Systems and provided on the GPU cluster Monacum One at the University of the Bundeswehr Munich.



(a) GT: similarity 0.4311 (b) Pred: similarity 0.670 (c) GT: similarity 0.4708 (d) Pred: similarity 0.4965

Figure 1: **False predictions from the CVUSA dataset predicted with a model trained on CVACT.** In the left image the street course is very similar, but the street-view image seems to be taken at another season than the satellite image. On the left the course of the road differs but the vegetation matches better.

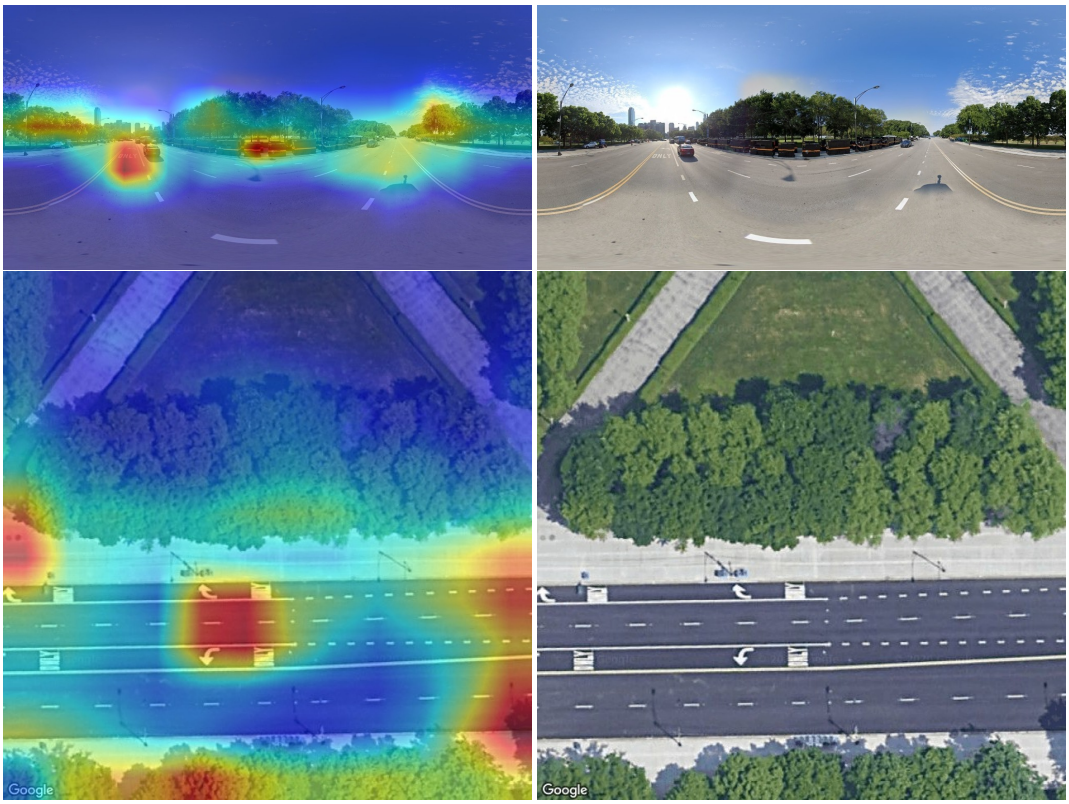


Figure 2: **Heatmap visualisation of a correct prediction on the VIGOR SAME split.** The model focuses here mainly on the road markings and the vegetation on both sides of the street.

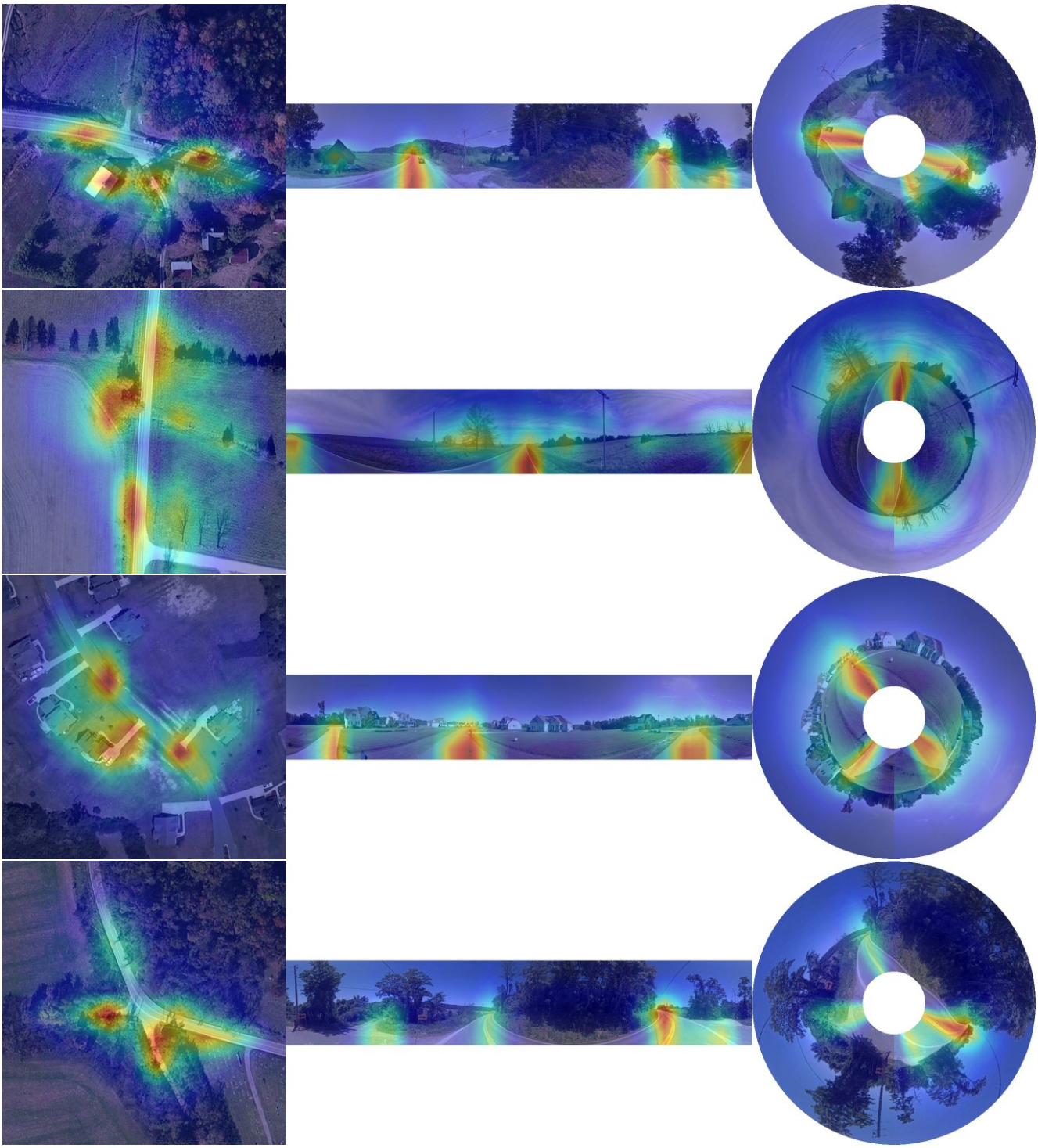


Figure 3: **Heatmap visualisations for correct predictions on the CVUSA dataset.** We use the inverse polar transformation to show the correspondence of the activation maps in satellite and street-views. According to the activation maps the most significant regions are the street course and position of intersections on that samples.