RHO: Robust Holistic OSM-Based Metric Cross-View Geo-Localization

Supplementary Material

A. More Implementation Details

A batch size of 36 was employed when training exclusively with a single **CV-RHO** variation. In contrast, a larger batch size of 96 was utilized for training involving all noisy variations or the complete set of data variations. We limit the maximal training epoch to 20, and the best model checkpoint is selected with early stopping based on validation loss. The evaluation batch size is set to 3. This value represents the minimal batch size for the RHO model, as the SUM module requires the input of at least three pinhole images for correct operation.

B. Prompt for Data Generation of CV-RHO

To achieve a realistic simulation of real-world scenarios under diverse weather conditions without destroying the original buildings and streets' structure in the image, we carefully design prompts and test parameters listed in Table 9 for generating **CV-RHO** images. All weather variations are generated with seed value 42 and true cfg scale 1.2, in inference steps 50. The guidance scale value differs among the 4 variations.

Utilizing the SOTA and open-source image generation model FLUX.1 Kontext, we edit images to produce photo-realistic variations under diverse weather conditions. Fig. 5 presents additional samples of generated images. The generated weather effects are various, rather than being static or uniform, which corresponds to the randomness in the real world.

C. Data Collection of CV-RHO

We selected sequences captured after 2020 using panoramic cameras. Fig. 6 shows OSM covered by the selected sequences. Images of each location were split into disjoint training and validation sets; the ratio between them is approximately 8:2, resulting in **2.16M** training and **540K** validation views. Data collected in Mount Vernon is used for cross-region evaluation. In the **Sim2Real** dataset, the Detroit snow area does not overlap with the area where clean data is collected. Data collected in Hasselt is utilized for sensor noise in **Sim2Real**.

D. Detail of Fusion Strategies

Table 8 of our main paper shows the performance of different fusion strategies. The **POF** fusion strategy outperforms $Prior_{uv}$ fusion and $Prior_{\theta}$ fusion, which are components of **POF** itself. These two fusion strategies are illustrated in Fig. 7. In $Prior_{uv}$ fusion, the orientation dimension is marginalized out, and the positional probability distribution $Prior_{uv}$ is derived from the panoramic score volume S_{pano}. This distribution is then directly added to the score volume S_1 to refine its position prediction. In contrast, $Prior_{\theta}$ fusion marginalizes out the positional dimensions to extract an orientation prior from S_1, which is then incorporated into S_{pano}. If no fusion strategy is employed, we use S_{pano} directly to estimate the final pose.

E. Limitations and Future Work

While our model demonstrates remarkable performance on the OSM-based MCVGL task, our proposed SUM module and cross fusion logic are specifically designed for panoramic imagery processing. When applied to pinhole images alone, these mechanisms fail to boost performance compared to pinhole-branch architectures. This limitation restricts the model's applicability to scenarios where only pinhole images are available. When training on the complete CV-RHO dataset containing over $2.7M$ images across all variations, our model achieves optimal check-

Table 9. Prompts and Parameters configuration for generating images with Flux.1 Kontext.

Variation	Prompt	Hyperparameter
Rain	Change the background into a heavy rainy day, change the sky into gray overcast sky, add visible rain streaks and puddles.	Guidance scale: 7.5
Snow	Change the background into a snowy day, add visible falling snowflakes	Guidance scale: 7.7
Fog	Turn the background into a foggy day, add visible smog on the street	Guidance scale: 7.7
Nighttime	Change the background into evening, illuminate the scene brighter with streetlights.	Guidance scale: 5.0
Over Exposure	-	Brightness factor: 2.5
Under Exposure	-	Brightness factor: 0.25
Motion Blur	-	Kernel size: 10

Figure 5. More samples of generated images. Images with red borders are original images.

points within early steps, despite iterative optimization of learning rates and optimizer configurations. This suggests that the current model architecture may not fully exploit the rich diversity of training samples.

Future work should focus on developing more sophisticated model architectures capable of processing multi-scale datasets and extending the two-branch framework to accommodate diverse input modalities. The evaluation across various datasets should be conducted to test our model's generalization capability.

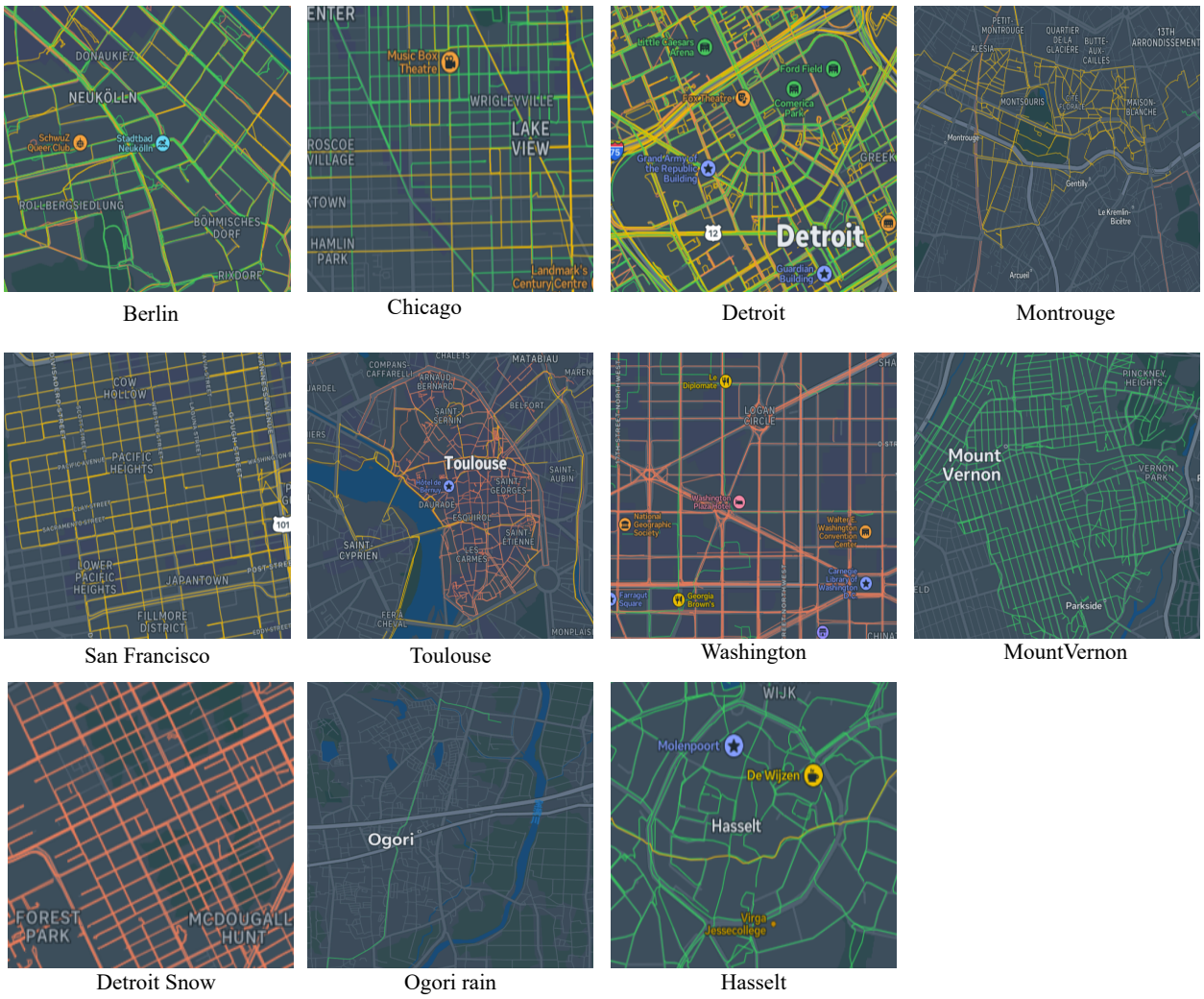
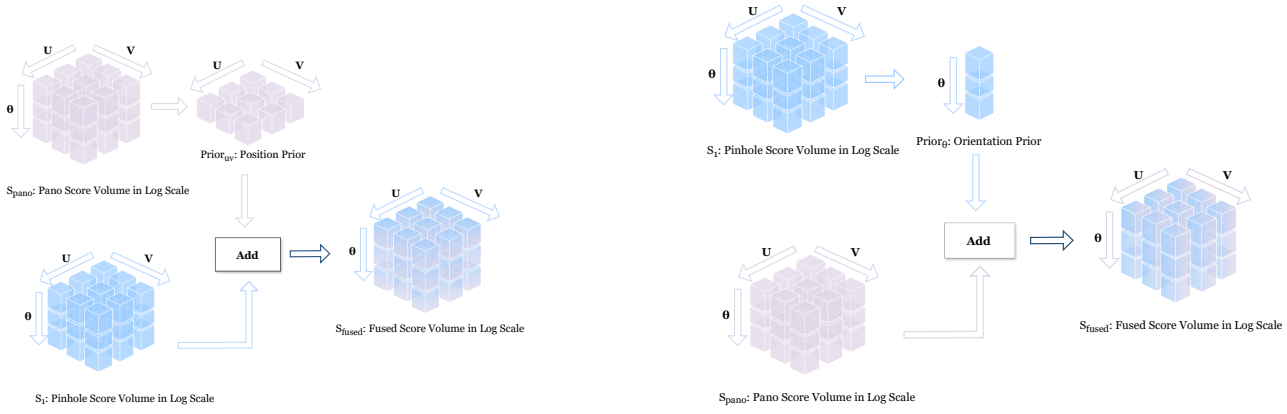


Figure 6. Selected data points of our CV-RHO dataset and Sim2Real across 11 cities. Color from red to green indicates capture date.



(a) $Prior_{uv}$ fusion (b) $Prior_{\theta}$ fusion

Figure 7. One-way fusion strategy: (a) $Prior_{uv}$ fusion: fuse S_1 with position prior $Prior_{uv}$. (b) $Prior_{\theta}$ fusion: fuse S_{pano} with orientation prior $Prior_{\theta}$.