Cross-Attention Between Satellite and Ground Views for Enhanced Fine-Grained Robot Geo-Localization Supplementary Material

Dong Yuan[1], Frederic Maire[1], Feras Dayoub[2]

[1]QUT Centre for Robotics, Queensland University of Technology, Australia {yuand2, f.maire}@qut.edu.au

[2]Australian Institute for Machine Learning (AIML), University of Adelaide, Australia feras.dayoub@adelaide.edu.au

Overview

In this supplementary material, we offer the following components to enhance the comprehension of the paper:

1. More details on the Positional Encoding.
2. Supplemental details on the ASAM optimization method.
3. Additional results on the VIGOR with correct labels.
4. Additional results on the Oxford RobotCar dataset.

1. Details on the Positional Encoding

Positional encoding is one of the essential components of transformer-based models [1, 3, 8], which can preserve spatial information of feature vectors [1] or patch embeddings [3]. In this work, following [1], we employ a fixed 2D sinusoidal positional encoding generated by the sine function:

$$\begin{aligned} \mathrm{PE}_{((x,y),4k)} &= \sin\left(x/10000^{2k/d_{model}}\right) \\ \mathrm{PE}_{((x,y),4k+1)} &= \cos\left(x/10000^{2k/d_{model}}\right) \\ \mathrm{PE}_{((x,y),4k+2)} &= \sin\left(y/10000^{2k/d_{model}}\right) \\ \mathrm{PE}_{((x,y),4k+3)} &= \cos\left(y/10000^{2k/d_{model}}\right) \end{aligned} \tag{1}$$

where (x, y) is the 2D position, d_{model} is the feature channel dimension and k is an index for feature channels. The generated positional encoding provides each element in the feature vectors with unique positional information.

2. Supplemental details on ASAM

Foret *et al*. [4] argue today's models are commonly non-convex with multiple local and global minima [4], which results in varying model generalization abilities. To address this issue, they first introduce Sharpness-Aware Minimization (SAM) for model optimization. SAM procedure can minimize loss sharpness by searching for parameters that lie in a neighborhood with consistently low loss values to improve model generalization. The sharpness of loss function L can be defined as:

$$\max_{\|\epsilon\|_2<\rho} L(w+\epsilon) - L(w), \tag{2}$$

where w is parameter weights and ϵ is the perturbation of w. The maximization region is a l_2 ball with radius ρ [5]. Dinh *et al*. [2] present model parameter re-scaling will not change loss L but generate a difference in the sharpness of loss, namely sharpness scale-dependency problem [5]. To remedy this problem, Kwon *et al*. [5] introduce normalization operators $\{T_w \in \mathbb{R}^k | T_{Aw}^{-1} = T_w^{-1}\}$ and define the adaptive sharpness of the loss L as follows:

$$\max_{\|T_w^{-1}\epsilon\|_2<\rho} L(w+\epsilon) - L(w). \tag{3}$$

The ASAM optimization method has been shown to be highly effective in addressing the overfitting problem in a transformer model [10], and our ablation study results also demonstrate the effectiveness of employing ASAM for accuracy improvement.

3. Results on VIGOR with corrected label

We notice that SliceMatch [6] uses the VIGOR [11] dataset for cross-view pose estimation (location and orientation prediction). Different from our setting, in [6], only positive satellite images are used for training and testing without considering semi-positive satellite images. Additionally, they note that the original VIGOR dataset contains distance errors of the ground truth locations, stemming from resolution mismatches in the satellite images. Therefore, we follow their settings and rerun our approach with corrected-labels. The comparison of localization results are

Model	Same-Area Positives		Cross-Area Positives	
	Mean	Median	Mean	Median
VIGOR [11]	8.99	7.81	8.89	7.73
MCC [9]	6.94	3.64	9.05	5.14
SliceMatch [6]	5.18	2.58	**5.53**	2.55
Ours	**4.26**	**1.86**	5.83	**2.28**

Table 1. The localization results on VIGOR [11] with corrected-label. Following [6], only positive satellite images are used for training and testing. Best performance is in bold.

Model Error	Test 1	Test 2	Test 3	Average
MCC mean	**1.42**	**1.95**	1.94	1.77 ± 0.25
Ours mean	1.50	1.97	**1.83**	$\mathbf{1.77 \pm 0.20}$
MCC median	**1.10**	**1.33**	**1.29**	$\mathbf{1.24 \pm 0.10}$
Ours median	1.22	1.42	1.33	1.32 ± 0.08

Table 2. Experimental results on the Oxford RobotCar [7]. Mean and median errors for three different test traversals are shown. Average and standard deviation values are computed by these data. Best results in bold.

shown in Table 1. As shown in Table 1, in comparison to SliceMatch [6], our approach does not rely on geometric knowledge guidance [6] but still achieves highly competitive localization results.

4. Additional results on Oxford RobotCar

We provide more results on three different test traversals as shown in Table 2. Our approach achieves competitive results on the test dataset of the Oxford RobotCar. However, the performance of our model is negatively impacted by the smaller number of images and the limited field of view.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 1

[2] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *ICLR*, pages 1019–1028. PMLR, 2017. 1

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

[4] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *ICLR*, 2020. 1

[5] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pages 5905–5914. PMLR, 2021. 1

[6] Ted Lentsch, Zimin Xia, Holger Caesar, and Julian FP Kooij. Slicematch: Geometry-guided aggregation for cross-view pose estimation. In *CVPR*, pages 17225–17234, 2023. 1, 2

[7] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, 36(1):3–15, 2017. 2

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 1

[9] Zimin Xia, Olaf Booij, Marco Manfredi, and Julian FP Kooij. Visual cross-view metric localization with dense uncertainty estimates. In *ECCV*, pages 90–106. Springer, 2022. 2

[10] Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *CVPR*, pages 1162–1171, 2022. 1

[11] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *CVPR*, pages 3640–3649, 2021. 1, 2