Visual Cross-View Metric Localization with Dense Uncertainty Estimates

Zimin Xia[1][0000−0002−4981−9514], Olaf Booij[2], Marco Manfredi[2][0000−0002−2618−2493], and Julian F. P. Kooij[1][0000−0001−9919−0710]

[1] Intelligent Vehicles Group, Technical University Delft, The Netherlands
{z.xia,j.f.p.kooij}@tudelft.nl
[2] TomTom, Amsterdam, The Netherlands
{olaf.booij,marco.manfredi}@tomtom.com

Appendix

The content in this appendix is related to the main paper as follows:

A More details on our used datasets in **Section 4.1 Datasets**.
B Supplemental details on **Section 4.2 Evaluation metrics** about our post-processing on baseline CVR for generating probability estimation.
C Additional results for **Section 4.4 Generalization in the same area / across areas**.
D Additional results for **Section 4.5 Generalization across time**.
E A detailed network diagram to supplement **Section 3.2 Proposed method**.

A Supplemental details on datasets

The original **VIGOR dataset** [6] does not provide a validation set for either the "same-area" nor the "cross-area" split. Therefore, we randomly split 20% of data from the training set as our validation set, used to determine the stopping epoch during training. The test set is kept as is and is used to report the evaluation results. Thereby, in the "same-area" split, there are 42087, 10522, and 52605 ground images in training, validation, and test set respectively. All three sets share the 90618 satellite images. In the "cross-area" setting, there are 41216, 10304, and 53694 ground images in training, validation, and test set respectively, and there are 44055 satellite images for training and validation in the first two cities, and 46563 satellite images for across-city testing.

Note that we do not use our validation set to find one set of hyper-parameters for both the "same-area" and "cross-area" experiments because the "same-area" validation set partially overlaps with the "cross-area" test set and vice versa. Hence, we make use of the "same-area" training image from New York to create a separate "tuning" set (11108 training and 2777 validation) to search for one set of hyper-parameters for all experiments and speed up the hyper-parameter search, see Figure 1. We will release our data splits.

The original **Oxford RobotCar dataset** [1, 2] does not contain satellite images. We make use of the satellite patches provided by [5, 4]. We intend to

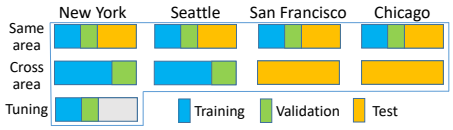

Fig. 1. Overview of our data split on VIGOR dataset

Fig. 2. Overview of our stitched continuous satellite image

conduct data augmentation such that, given a ground-level image, we can crop any satellite patch from a continuous satellite image that contains the ground image. However, the satellite patches from [5, 4] do not provide enough coverage of the target area.

To enable the intended data augmentation, we collect additional satellite patches from Google Maps Static API[1], and stitch all satellite patches to create a continuous satellite map that covers the target area ($\sim 1.5km \times 1.8km$), see Figure 2. The geographical coordinates of each pixel in our continuous satellite image are known. Therefore, given the geographical location of a ground image, we can find its image pixel coordinates on the satellite image. We *will release* the raw data[2] we collected and the code for stitching satellite patches.

[1] `https://developers.google.com/maps/documentation/maps-static/overview`

[2] We release the images in accordance with the "fair use" policy in Google Maps/Google Earth Terms of Service, version: Jan 12, 2022 (`https://www.google.com/help/terms_maps/`). The released data is for the reproducibility of scientific re-

As mentioned in **main paper Section 4.1**, we align the rotation between the ground image and satellite patch. Given a ground image and the heading of the camera we first crop a large satellite patch from the continuous map at the targeted cropping location and then rotate the satellite image such that the up direction in the cropped large patch corresponds to the viewing direction in the ground image. Then we crop the satellite patch with the required resolution from the rotated large patch. Note that, when testing classification of the orientation, **main paper Section 4.5 second to the last paragraph**, the satellite patches are rotated with a fixed set of angles. Thus the orientation-aligned patch will not appear in the set.

B Details on post-processing CVR localization

As we illustrated in **main paper Section 4.2**, to acquire the probability estimation from the baseline CVR method, we assume the regressed location from CVR is the mean of an isotropic Gaussian distribution and we estimate the standard deviation of this Gaussian distribution on the validation set.

Specifically, we record the error distribution on the validation set, and calculate the standard deviation SD using $SD = \sqrt{\frac{\sum^n e^2}{n}}$, in which e is the distance error of each ground image in the validation set and n is the number of ground images. On the VIGOR dataset, the estimated standard deviation is 12.36m on the same-area split and is 11.64m on the cross-area split. On the Oxford RobotCar dataset, the value is 3.36m.

C Additional results, generalization in the same area / across areas

Here, we provide more results for our experiments in **main paper Section 4.4**.

C.1 Extra model variations

We studied two extra variations of our proposed method.
Variation 1 has a multi-layer perceptron on top of our proposed model. It uses the output heat map to regress the relative 2D location offset between the ground camera's location and the center of the satellite patch. Variation 2 drops the whole decoder and selects the location of the highest similarity score out of the 8×8 matching score map. The ground descriptor is then concatenated with the satellite descriptor at the selected location. The concatenated descriptor is used by a multi-layer perceptron to regress the relative 2D location offset between the ground camera's location and the center of the small satellite patch corresponding to the selected location. Note that, both variations only regress a single

search and cannot be used for commercial purposes. Google remains the copyright of images.

Table 1. Localization error on VIGOR satellite patches. Best in bold.

	Same-area				*Cross-area*			
	Positives		Pos.+semi-pos.		Positives		Pos.+semi-pos.	
Error (m)	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Center-only	14.15	14.82	27.78	28.85	14.07	14.07	27.80	28.89
CVR [6]	10.55	9.31	16.64	13.82	11.26	10.02	18.66	16.73
Ours	9.86	**4.58**	13.45	**5.39**	13.06	**6.31**	17.13	7.78
Variation 1	**8.01**	6.29	**12.26**	9.20	**9.62**	7.57	**14.25**	11.42
Variation 2	10.69	5.88	13.40	6.07	12.49	6.92	15.29	**7.30**

location without dense uncertainty estimates. Similar to the post-processing on CVR's output, see Section B, we assumed the regressed location is the mean of an isotropic Gaussian distribution and estimate the standard deviation on the validation set to acquire the probabilistic output.

Table 2. Probability at the ground truth location on VIGOR positive satellite patches. Best in bold.

Prob. at GT, Positives	Same-area		Cross-area	
	Mean	Median	Mean	Median
Uniform	3.81×10^{-6}	3.81×10^{-6}	3.81×10^{-6}	3.81×10^{-6}
CVR [6]	1.55×10^{-5}	1.70×10^{-5}	1.57×10^{-5}	1.72×10^{-5}
Ours	$\mathbf{2.93 \times 10^{-4}}$	$\mathbf{1.17 \times 10^{-4}}$	$\mathbf{1.54 \times 10^{-4}}$	$\mathbf{7.06 \times 10^{-5}}$
Variation 1	1.15×10^{-5}	8.40×10^{-6}	1.13×10^{-5}	9.36×10^{-6}
Variation 2	7.25×10^{-6}	6.89×10^{-6}	7.09×10^{-6}	6.86×10^{-6}

In Table 1 and 2, we provide the comparison of all models on the VIGOR test splits. Variation 1 is better than CVR [6] all-round. This benefit might come from the larger model size and the dense feature before the final regression head. However, the regression-based variation 1 still has a significantly higher median error, which might be due to the regression head picking the midpoint between visually similar locations. Regarding variation 2, on the same-area split, it performs worse than our model. This shows the advantage of having a decoder to process the matching score and satellite feature. However, on the cross-area split, our model performs slightly worse than variation 2. Possibly, given the more uncertain matching score map and features in unseen areas, the decoder does not know how to reduce this uncertainty effectively.

Importantly, variation 1 and 2 are not optimized for higher probability at the ground truth location. Thus, both variations have over one magnitude lower probability at ground truth than our model. We highlight again that probability estimation is crucial in both sensor fusion and temporal filtering. Last but not least, both variations miss the dense output to capture a multi-modal distribution. Thus, we conclude that our proposed model and problem formulation is better.

C.2 Probability evaluation

Supplementary to the cumulative statistics shown in **main paper Figure 5**, we show the noncumulative statistics of the predictions ranked by their probabilities in Figure 3.

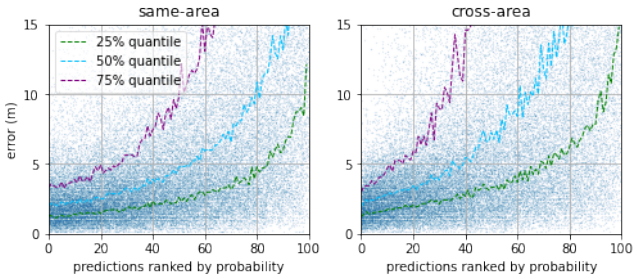

Fig. 3. Ranking the predictions using their probabilities on VIGOR "positives".

C.3 Orientation

Additional to solely evaluating either localization or orientation, we also tested jointly evaluating both on the positive satellite patches. As in **main paper Section 4.4. Orientation**, the ground panorama is rotated by multiples of 22.5° up to 360°. We pair up each of the rotated panoramas with the satellite patch. Hence one mini-batch contains the same satellite image with 16 rotated ground panoramas. We collect all 16 activation maps and pass them into a single softmax operation. The location and corresponding orientation of the peak probability is then the localization and orientation estimation. Under this setting, the mean/median metric localization error increases from 9.86/4.58 (main paper Table 2 Same-area Positives columns) to 13.99/8.28 on the same-area test set, and from 13.06/6.31 (main paper Table 2 Cross-area Positives columns) to 18.18/14.61 meters on cross-area test set. This is a more challenging setting that leads to a noticeable decrease in performance.

C.4 Localization qualitative results

In addition, we provide more qualitative cross-view metric localization results in the same-area and cross-area splits in Figure 4 and 5.

As pointed out in the **main paper Section 4.4.**, our method has slightly more outliers than the CVR. In Figure 6, we show some of the failure cases where our localization error is higher than CVR. Importantly, as we stated in the main paper, instead of regressing to a location closer to the ground truth location as

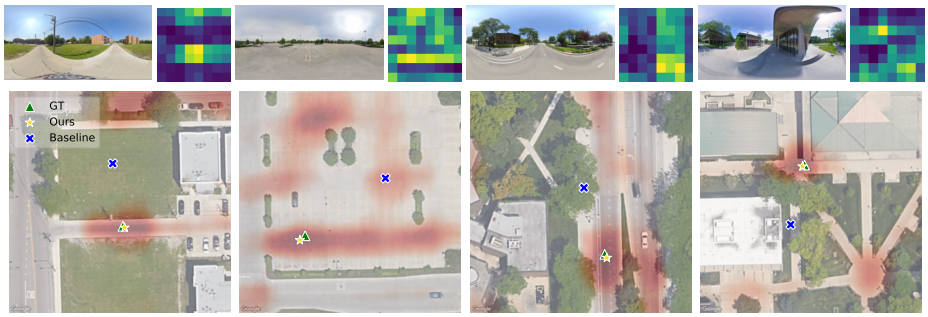

Fig. 4. Qualitative results on VIGOR dataset same-area split. Top: input ground images and matching score maps at the model bottleneck, bottom: input satellite image overlayed with outputs from CVR and our method.

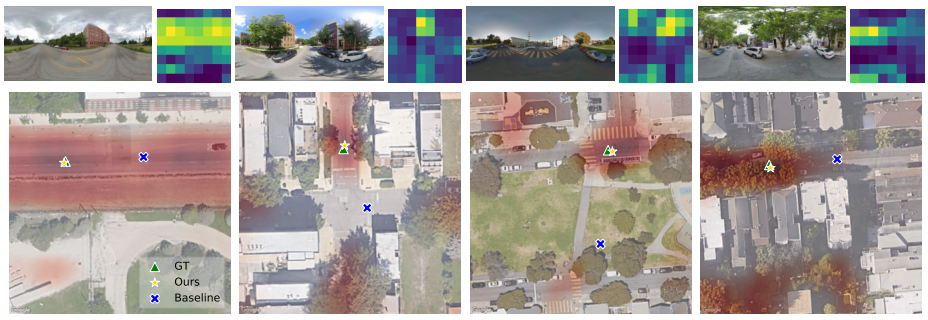

Fig. 5. Qualitative results on VIGOR dataset cross-area split. Top: input ground images and matching score maps at the model bottleneck, bottom: input satellite image overlayed with outputs from CVR and our method.

CVR, our method expresses the underlying uncertainty. This property is desirable in real-world applications since the ground truth location is also captured plus our prediction expresses the model's underlying uncertainty.

Finally, we show qualitative results of localizing the ground image inside both positive and semi-positive satellite patches in Figure 7. The predicted location from our model is more consistent than that from CVR. This confirms our benefit of localizing against semi-positive satellite patches shown in the **main paper Section 4.4 Table 2**.

D Additional results, generalization across time

Next, we provide more results for **main paper Section 4.5**.

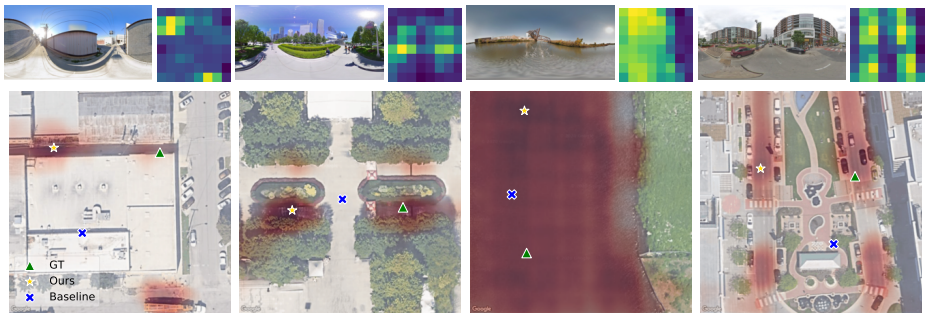

Fig. 6. Failure cases on VIGOR dataset. Top: input ground images and matching score maps at the model bottleneck, bottom: input satellite image overlayed with outputs from CVR and our method.

Table 3. Localization error on Oxford RobotCar test traversals. Last column: Average over all traversals

Error (meters)	Test 1	Test 2	Test 3	Average
CVR mean	1.88	2.64	2.35	2.29 ± 0.31
Ours mean	**1.42**	**1.95**	**1.94**	**1.77±0.25**
CVR median	1.47	1.99	1.71	1.72 ± 0.21
Ours median	**1.10**	**1.33**	**1.29**	**1.24±0.10**

D.1 Metric error and probability evaluation

As a supplement to the average metric localization error and average probability at the ground truth pixel over three Oxford RobotCar test traversals in the main paper, here, we provide the separate test results on each traversal in Table 3 and 4 for reference.

D.2 Orientation

Similar to Section C.3, we also report the localization error when jointly evaluating both localization and orientation estimation. We keep the same setting as in **main paper Section 4.5** that the satellite patches are rotated 16 times with 22.5°, starting at 0° where north points in the vertically up direction. In this case, the mean/median error increases from 1.42/1.10 (Table 3, Test 1) to 5.46/1.98, 1.95/1.33 (Table 3, Test 2) to 6.05/2.44, and 1.94/1.29 (Table 3, Test 3) to 5.77/2.35 meters on three test traversals.

Akin to the evaluation on VIGOR, **main paper Figure 6 right**, we also provide the histogram of orientation classification results on Oxford RobotCar. As shown in Figure 8, the most frequent wrong class is 180°. We show qualitative results of the orientation classification experiment in Figure 9. The model is trained with orientation-aligned satellite patches, which means if the ground image's sight is along the road, the corresponding road in the satellite patch is along the vertical direction. In inference, our model exploits this information and

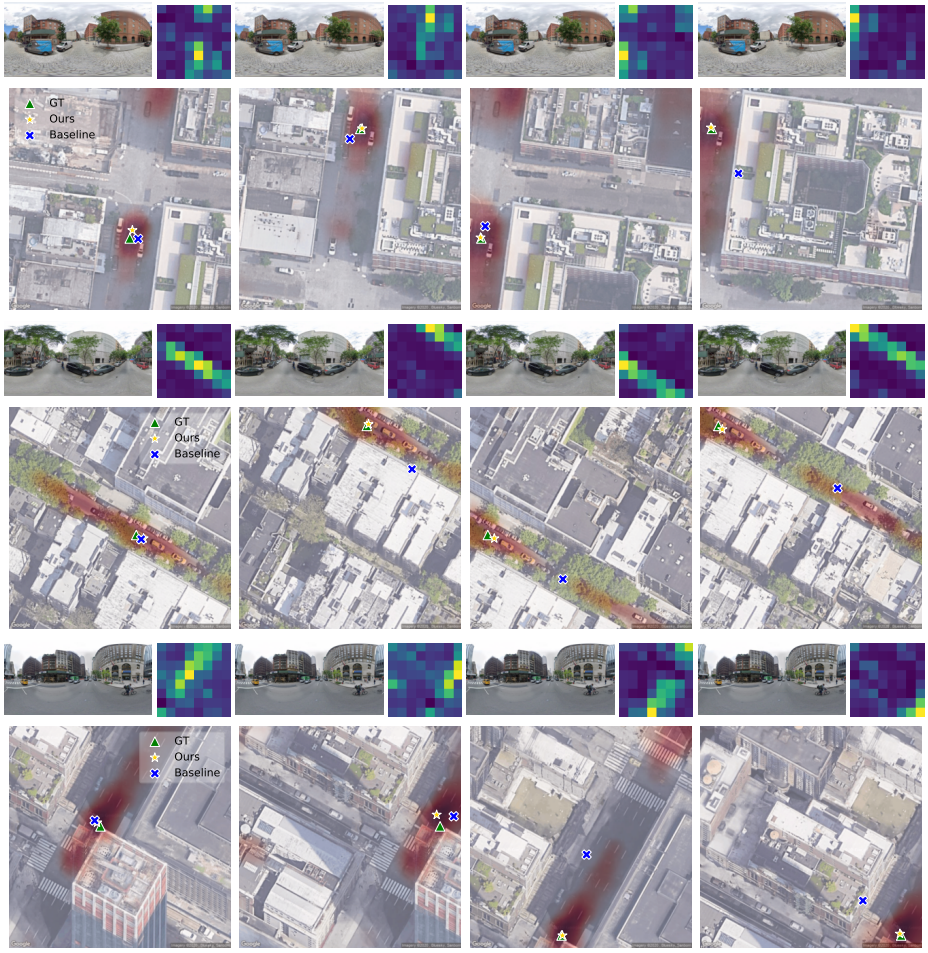

Fig. 7. Matching a ground image to the positive and all semi-positive (shifted) satellite patches on the VIGOR dataset. Each example includes an input ground image, a matching score map at our model bottleneck, and an input satellite image overlayed with outputs from CVR and our method. The first (from left to right) example in each row is the matching between the ground image and positive. The other three show the matching between the same ground image and 3 semi-positive satellite patches.

Table 4. Probabilities at the ground truth pixel on Oxford RobotCar. Last column: Average over all traversals. For reference, the probability in a 512×512 uniform map is 3.81e-6. Best in bold

Prob. at GT	Test 1	Test 2	Test 3	Average
CVR mean	1.78×10^{-4}	1.59×10^{-4}	1.65×10^{-4}	1.67×10^{-4}
Ours mean	$\mathbf{1.76 \times 10^{-3}}$	$\mathbf{1.40 \times 10^{-3}}$	$\mathbf{1.46 \times 10^{-3}}$	$\mathbf{1.54 \times 10^{-3}}$
CVR median	1.96×10^{-4}	1.81×10^{-4}	1.89×10^{-4}	1.89×10^{-4}
Ours median	$\mathbf{1.69 \times 10^{-3}}$	$\mathbf{1.18 \times 10^{-3}}$	$\mathbf{1.28 \times 10^{-3}}$	$\mathbf{1.38 \times 10^{-3}}$

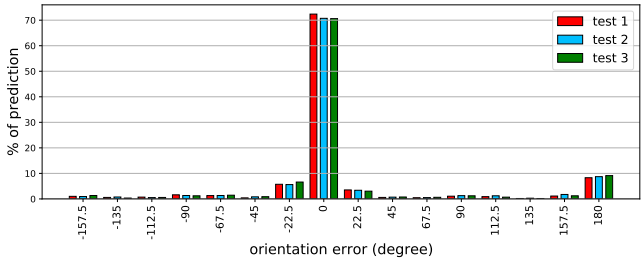

Fig. 8. An overview of orientation classification results on three test traversals from Oxford RobotCar.

assigns low probability when the road is not vertical. Importantly, our model is not a vertical road detector, since it also tries to match other objects, e.g. trees, across views, plus it can also differentiate the satellite patch rotated 180° from the correct orientation class in most cases. As shown in Figure 9, when the ground image indicates the vehicle is under the tree, our model tries to find the vegetation in the satellite view no matter the rotation angle.

D.3 Localization qualitative results

We provide more qualitative results on metric localization for both CVR and our method on three test traversals in Figure 10. Different traversals vary in time, weather, and lighting conditions. In Figure 11, we show the predictions from 3 test traversals at roughly the same location. Note that the headings of the ground camera are slightly different.

E Network Diagram

Finally, we provide our network diagram in Figure 12. We adopt the off-the-shelf SAFA module from [3] for image descriptor construction.

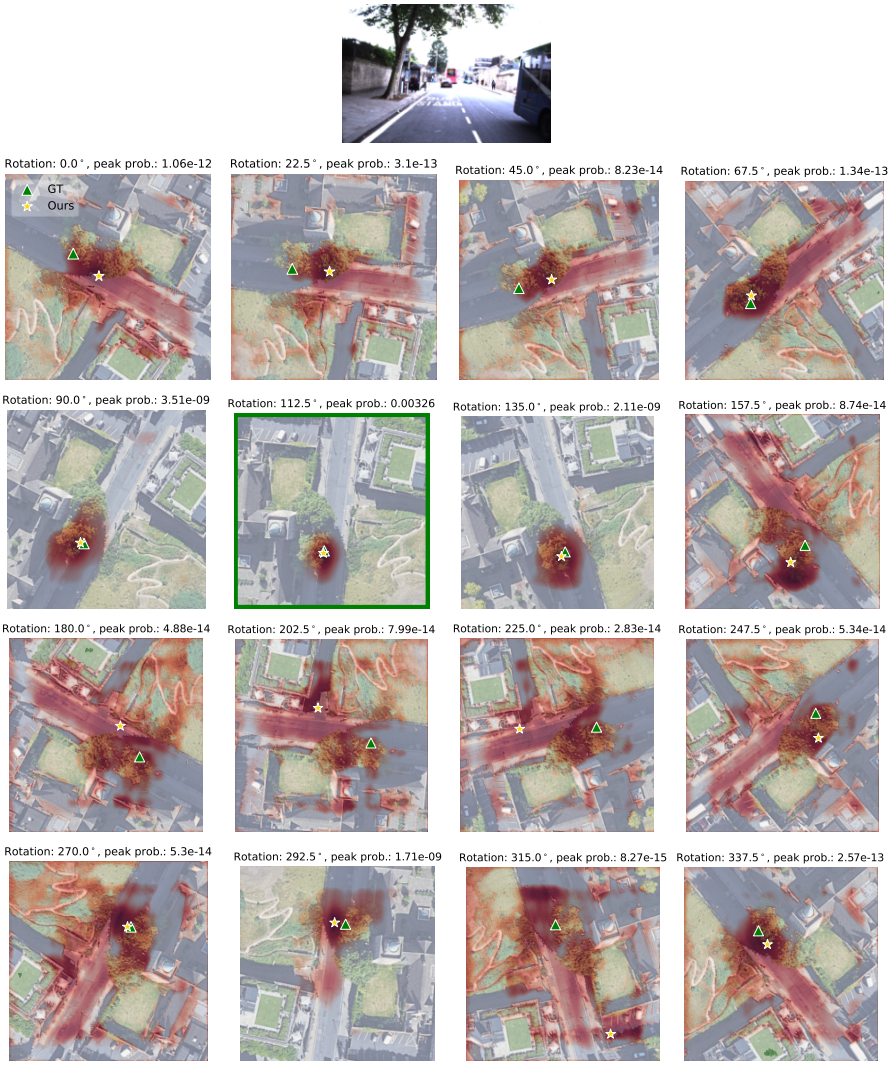

Fig. 9. Oxford RobotCar, the same ground image matched to 16 different rotated satellite patches. The ground truth orientation is 118°. The satellite patch with 0° rotation is the north-aligned satellite patch. The patch with a green box (112.5°) is the one prediction with the highest peak probability. The probabilities are generated by the softmax operation over a batch of inputs. For visualization, we re-scale the probability in each heat map.

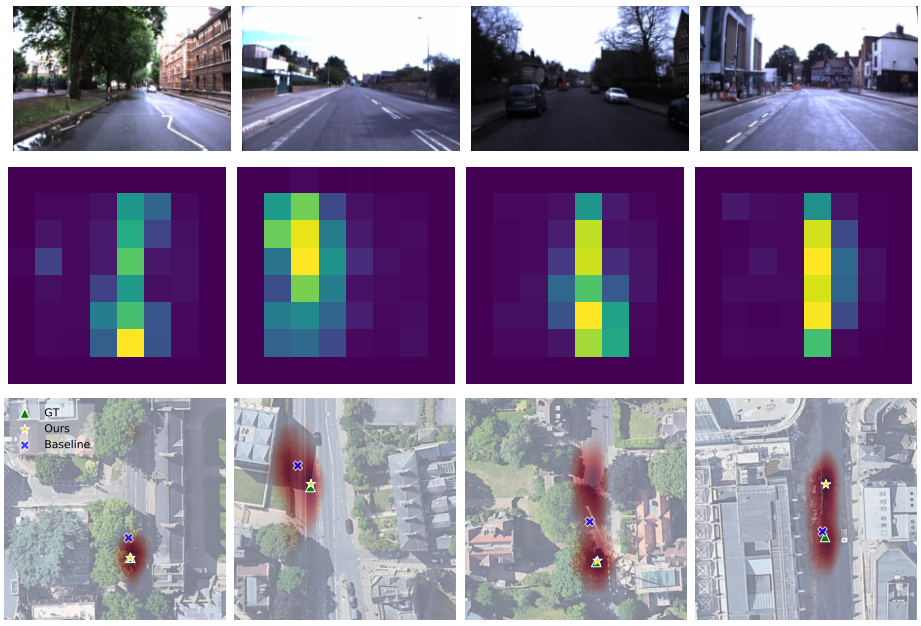

Fig. 10. Qualitative results on Oxford RobotCar dataset. Each column has an input ground image, a matching score map at our model bottleneck, and an input satellite image overlayed with outputs from CVR and our method. Column 1/2/3 (from left to right): good predictions in test traversal 1/2/3. Column 4: a failure case example.

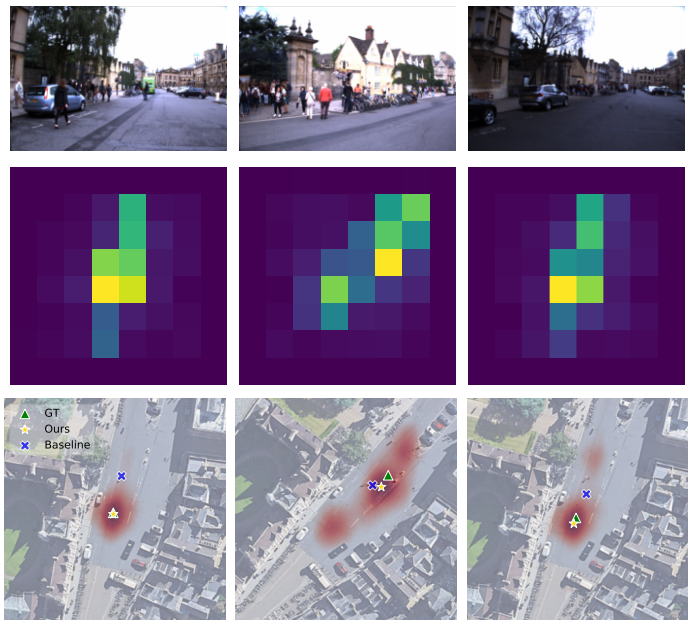

Fig. 11. Examples at a rough same location in 3 test traversals on Oxford RobotCar dataset. Ground images in different test traversals are collected at different times and weather. Each column has an input ground image, a matching score map at our model bottleneck, and an input satellite image overlayed with outputs from CVR and our method. Column 1/2/3 (from left to right): test traversal 1/2/3.

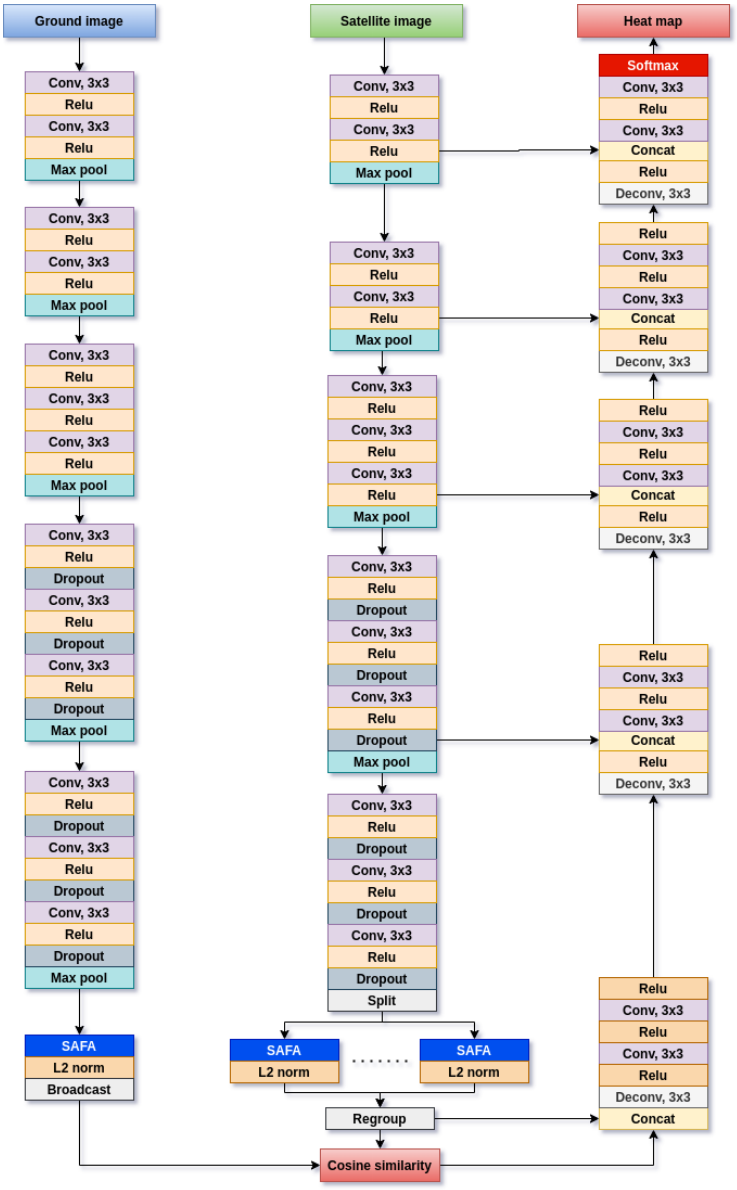

Fig. 12. Network diagram

References

1. Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 year, 1000 km: The oxford robotcar dataset. IJRR **36**(1), 3–15 (2017)
2. Maddern, W., Pascoe, G., et al.: Real-time kinematic ground truth for the oxford robotcar dataset. arXiv preprint: 2002.10152 (2020)
3. Shi, Y., Liu, L., Yu, X., Li, H.: Spatial-aware feature aggregation for image based cross-view geo-localization. In: NeurIPS. pp. 10090–10100 (2019)
4. Xia, Z., Booij, O., Manfredi, M., Kooij, J.F.P.: Cross-view matching for vehicle localization by learning geographically local representations. IEEE Robotics and Automation Letters **6**(3), 5921–5928 (2021). https://doi.org/10.1109/LRA.2021.3088076
5. Xia, Z., Booij, O., Manfredi, M., Kooij, J.F.: Geographically local representation learning with a spatial prior for visual localization. In: ECCV Workshops. pp. 557–573. Springer (2020)
6. Zhu, S., Yang, T., Chen, C.: Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In: Proc. of IEEE/CVF CVPR. pp. 3640–3649 (2021)