Supplementary Material

In this supplementary material, we provide the following information to support the main paper:

A. Supervised Finetuning with Noisy Ground Truth

As mentioned in Section 4.5 of the main paper, when the ground truth locations for cross-area supervised finetuning contain errors, the finetuned model has large test errors. We test this by applying random offsets to the ground truth locations.

In our experiments, offsets were sampled randomly and uniformly (in both the north-south and east-west directions) within a defined range for each ground-level image in the cross-area training set prior to finetuning. These offsets were then applied to shift the ground truth locations of the training images. As Figure 1 demonstrates, inaccuracies in the ground truth markedly affect the localization precision of the finetuned model. For the supervised finetuned model to outperform the model trained with our weakly-supervised learning approach in terms of both mean and median test errors, the maximum permissible error in the ground truth for each direction should be under approximately 2.5 m. In practice, acquiring ground truth with this level of accuracy on a large scale is difficult, as standard GNSS positioning does not meet this requirement [2]. Instead, our proposed method requires only ground-aerial image pairs, making it a more scalable solution in practice.

B. Domain Adaptation by Entropy Minimization

As noted in our main paper Section 4.7, we explore entropy minimization [4] as an alternative approach to adapt a model from the source domain to the target domain. Entropy minimization is often used for semi-supervised domain adaptation [10]. In this setting, the model is trained with a combination of samples with ground truth labels from the source domain and unlabeled samples from the target domain. When a source domain sample is presented, the model is

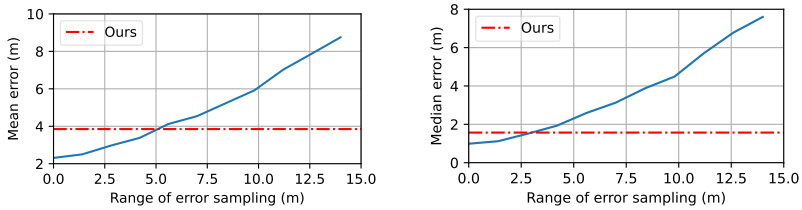


Fig. 1: VIGOR test set errors (vertical axis) of CCVPE models fintuned on noisy ground truth. The horizontal axis denotes the upper bound for error sampling.

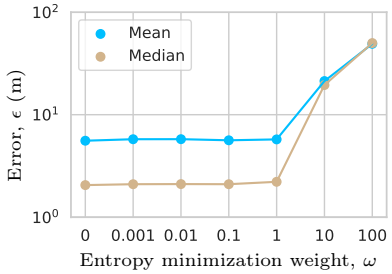


Fig. 2: Errors of CCVPE models with different entropy minimization weights ω on VIGOR validation set.

trained using its default supervised learning loss $\mathcal{L}_{\mathcal{M}}$. When the input is from the target domain, the training objective is to minimize the entropy of the output prediction using an entropy minimization loss $\mathcal{L}_{EM}$.

We train a CCVPE model [12] on a combination of VIGOR source and target domain data using loss $\mathcal{L}_{final}$,

$$\mathcal{L}_{final} = \begin{cases} \mathcal{L}_{\mathcal{M}}(\mathcal{M}(G, A), \hat{y}), & \text{if } \{G, A\} \in \mathbb{I}_{\alpha}, \hat{y} \in \mathbb{Y}_{\alpha}, \\ \omega \cdot \mathcal{L}_{EM}(H_K), & \text{if } \{G, A\} \in \mathbb{I}_{\beta}. \end{cases} \tag{1}$$

In Equation 1, $\mathcal{L}_{\mathcal{M}}$ is the default supervised learning loss of CCVPE [12], H_K is the final output heat map of the model $\mathcal{M}$ on image pair $\{G, A\}$, and ω is a hyperparameter that weighs the entropy minimization loss $\mathcal{L}_{EM}$. As in [10], we calculate the pixel-wise Shannon Entropy [8] in the dense output, and then use the sum of all pixel-wise entropy as our $\mathcal{L}_{EM}$,

$$\mathcal{L}_{EM}(H_K) = -\sum_{u,v} H_K(u, v) \cdot \log(H_K(u, v)), \tag{2}$$

$H_K(u, v)$ denotes the value at each location in the output heat map H_K.

We tuned ω and found that joint training with entropy minimization always hurts the model performance. As shown in Figure 2, the mean and median error on the validation set (target area) increases when the model is trained using a larger weight ω, and the best model appears when $\omega = 0$,

equivalent to direct generalization of a model trained in a supervised manner on only source domain images.

For completeness, we also tried directly finetuning a pre-trained model from the source domain on images from the target domain using entropy minimization (no joint supervised training with source domain samples). Since the model failed completely, we did not include the plots.

Entropy minimization simply encourages the heat map to be sharper in the target area. Therefore, it does not resolve multi-modal uncertainty. As shown in Figure 3, compared to direct generalization, training with entropy minimization makes the red region in the heat map smaller, but the peak of the heat map stays in the same mode in the multi-modal distribution. Instead, our proposed knowledge self-distillation adapts the model to the target domain by explicitly encouraging the model to disambiguate multiple modes using the proposed single-modal pseudo ground truth. As a result, our proposed method can correct the wrong mode and also reduce uncertainty.

C. Domain Adaptation by Other Pseudo Label-based Approaches

Our proposed Coarse-only Supervision uses the model's highest resolution output to supervise low-resolution ones. Alternatively, we also studied fusing the outputs at different levels to generate supervision signals.

Similar to [5], we fuse information in both top-down and bottom-up directions to generate pseudo ground truth at each level for the student model. We achieved this by up/downsampling teacher's matching volumes at different levels and fusing them with averaging. The error of the resulting student (4.49 m) is larger than ours (3.85 m) and the teacher model (4.38 m). We hypothesize that for localization, fine-grained high-resolution heat maps can help supervise low-resolution maps, but not vice versa, which may be why [5]'s top-down + bottom-up approach does not work for our task.

As an alternative to our proposed outlier filtering, we also tried an uncertainty-based outlier filtering approach while keeping other proposed modules unchanged. Similar to [6,11,13], we use the entropy of teacher's output heat maps as a measure of their uncertainty. The teacher's heat maps are ranked based on their entropy and we use the most certain T% for student training. For a fair comparison, CCVPE uses top 80% and GGCVT uses top 70% (same as in our proposed outlier detection). The resulting models have higher errors (CCVPE/GGCVT: 4.17/4.52 m) than ours (3.85/4.34 m). Entropy-based methods do not consider the spatial order of classes, e.g. a two-mode heat map with 1 m between two modes will have the same entropy as a two-mode heat map with 10 m between modes. However, the latter results in larger errors.

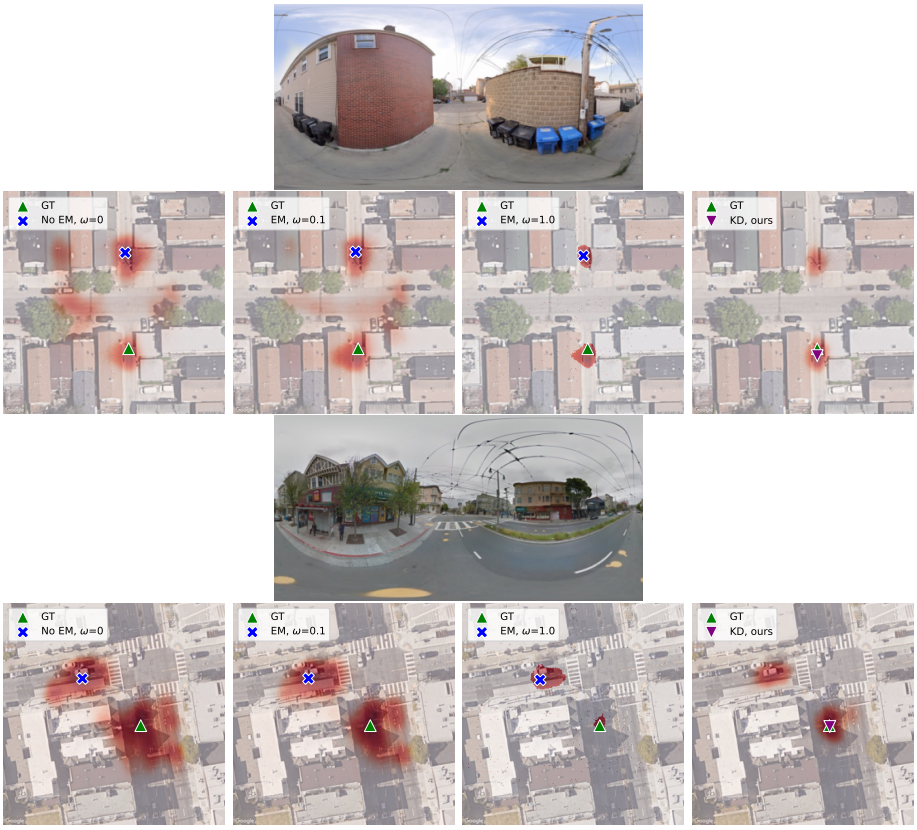


Fig. 3: Adapting a CCVPE model to the target domain with different methods. Results on the VIGOR test set. Comparison between direct generalization (No EM, $\omega = 0$), different entropy minimization weights (EM, $\omega = 0.1$ and EM, $\omega = 1.0$), and our proposed knowledge self-distillation (KD, ours). The red color denotes the localization probability (a darker color means a higher probability).

D. Analysis of Prediction Errors after KD for GGCVT

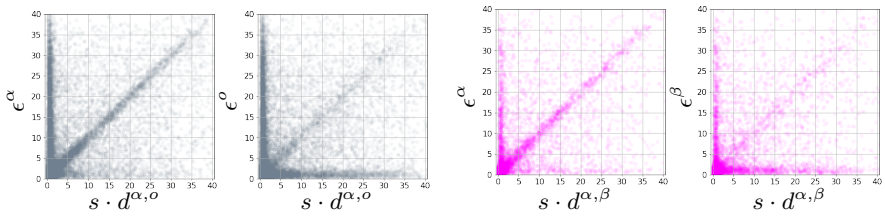
Similar to the analysis of the predictions of CCVPE in our main paper Section 4.6, we here provide the overall statistical relation between the GGCVT's prediction errors and the change in its predicted locations after knowledge distillation. **Overall, we observe the same trend for GGCVT as we had for CCVPE in the main paper, see Figure 4.**

First, a strong correlation between the teacher model's prediction errors and the amount of difference between the predicted locations of a teacher and its auxiliary student model is observed from the diagonal line in Figure 4a left. This again confirms that the outliers in the teacher's prediction can be identified by measuring the changes in the predicted location after knowledge self-distillation,

no matter what the localization backbone is, demonstrating the effectiveness of our proposed outlier filtering.

Figure 4a right plot has many scattered points along the horizontal axis, representing the predictions that are corrected by the auxiliary student model. The diagonal line in this plot then shows the samples in which the auxiliary student model introduced an error in its predictions, i.e. the correct teacher's predictions being moved to a wrong location or the wrong teacher's predictions being moved to another wrong location.

On the VIGOR test set $\mathbb{I}_{test}$, Figure 4b validated that the final GGCVT student model reduces the error of its teacher, as shown by the less prominent diagonal line and more points along the horizontal axis in the right plot compared to those in the left plot.



(a) Teacher (left) *vs.* Auxiliary student (right) models on $\mathbb{I}_{\beta}$

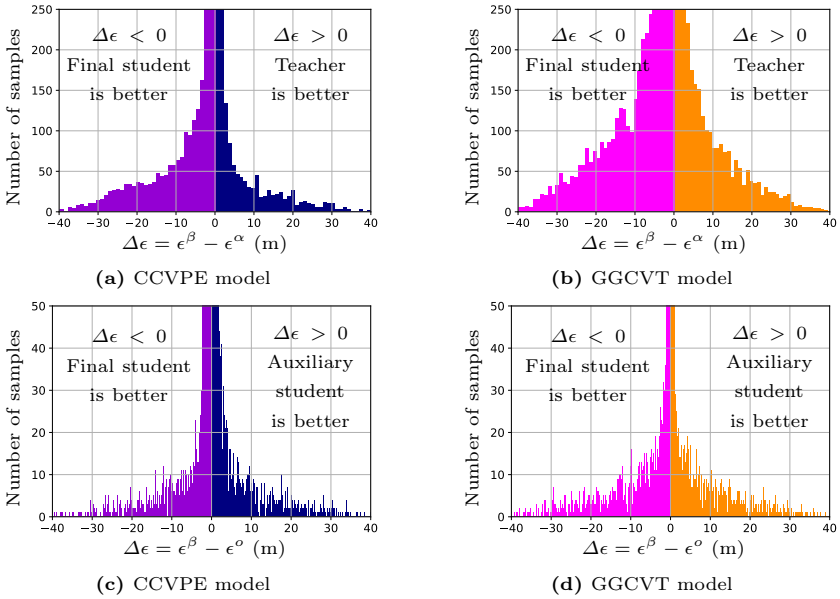
(b) Teacher (left) *vs.* Final student (right) models on $\mathbb{I}_{test}$

Fig. 4: GGCVT model, relation between error ϵ and change d in predicted locations from teacher and student models on VIGOR. ϵ^{α} / ϵ^{o} / ϵ^{β}: errors (m) of teacher model's / auxiliary student model's / final student model's predictions. $s \cdot d^{\alpha,o}$ / $s \cdot d^{\alpha,\beta}$: the difference (m) between predicted locations of teacher and auxiliary / final student.

E. Error Distribution of Teacher and Student Models

Next, we compare the error in predictions of the teacher model and that of the final student model for both CCVPE and GGCVT on the VIGOR test set $\mathbb{I}_{test}$. We calculate the error change after weakly-supervised knowledge self-distillation and visualize the statistics. In Figure 5 (a) and (b), the left part of the two histograms (in purple and magenta) shows the samples that have a smaller error in the final student model's prediction. Similarly, the right part of the two histograms (in navy and orange) denotes the samples that the teacher model has a more accurate prediction. Overall, we see that, for both CCVPE and GGCVT, there are more samples located in the left part. **It demonstrates that the final student model reduces the error for the majority of samples.**

For completeness, we also compare the final student model to the auxiliary student model, see Figure 5 (c) and (d). As expected, for CCVPE, it is difficult to see from the histograms a clear performance gap between the auxiliary student model and the final student model, since the improvement of the final student



(a) CCVPE model **(b)** GGCVT model

(c) CCVPE model **(d)** GGCVT model

Fig. 5: Change in errors. Plots (a) and (b): Comparison between prediction errors of the teacher model $\mathcal{M}^{\alpha}$ and prediction errors of the final student model $\mathcal{M}^{\beta}$ on VIGOR test set $\mathbb{I}_{test}$. Purple and Magenta region: The final student model has smaller errors. Navy and Orange region: The teacher model has smaller errors. Plots (c) and (d): Comparison between prediction errors of the auxiliary student model $\mathcal{M}^{o}$ and prediction errors of the final student model $\mathcal{M}^{\beta}$ on VIGOR test set $\mathbb{I}_{test}$. Purple and Magenta region: The final student model has smaller errors. Navy and Orange region: The auxiliary student model has smaller errors.

model over the auxiliary student model is relatively small, *i.e.* ~ 0.1 m mean error reduction. For GGCVT, the final student model outperforms the auxiliary student model by ~ 0.4 m in mean error. It can be seen in Figure 5 (d) that the magenta area has slightly more samples in the -40 to -20 m region compared to the 20 to 40 m region in the orange area. This shows that the final student model has fewer outliers than the auxiliary student.

F. T-SNE Feature

To study if the extracted features by the teacher and final student models differ, we use t-SNE [7] to map the features to a two-dimensional space for visualization. We collected CCVPE's ground features and the aerial features at the GT locations at the model bottleneck. Figure 6 shows their t-SNE plots before (teacher model) and after adaptation (final student model). For the teacher model, ground and aerial samples are disjoint in the feature space, complicating matching across views. For our student the plot shows more overlap between the

two views, indicating better alignment. This result supports that the quantitative improvement of our approach results from adaptation to the target domain.

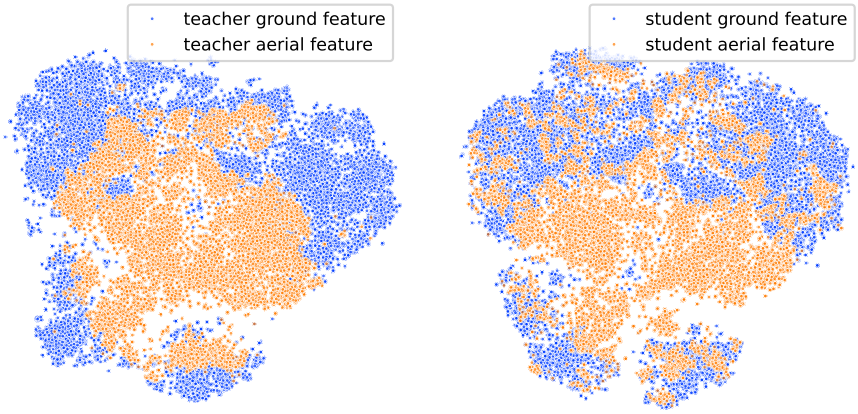


Fig. 6: t-SNE, VIGOR test set: CCVPE teacher model (left) and final student model (right).

G. Assumption on Orientation

Our practical objective is to adapt a trained fine-grained cross-view localization model to a new target area with easy-to-collect data, which includes orientation measurements (see footnote 1 in the main paper). The known orientation is used for lifting the 2D pseudo ground truth heat map to 3D for the CCVPE method, see main paper Section 4.4. As noted in Equation 5 in the main paper, our loss at output level k is a weighted sum of infoNCE losses. Each infoNCE loss is defined as,

$$\mathcal{L}_{\text{infoNCE}}(H_k^\beta \mid (m,n)) = -log\frac{\exp(H_k^\beta(m,n)/\tau)}{\sum_{m',n'}\exp(H_k^\beta(m',n')/\tau)}. \tag{3}$$

The infoNCE loss interprets the output heat map H_k^β in the target area β as similarity scores. (m',n') is the location index in H_k^β and (m,n) is the given location index of the positive class. τ is a hyperparameter and we use $\tau = 10$ as in [12]. When the output heat map and pseudo ground truth are 3D, *i.e.* including the orientation channels, the pseudo ground truth becomes a 3D mask that weighs the infoNCE loss at each location-orientation combination, and the infoNCE loss contrasts the given positive location-orientation combination to all other location-orientation combinations.

Since the ground truth orientation is used in the weakly-supervised learning in the target area, we are interested in whether this information largely contributes to the superior performance of the learned student model. We perform two additional experiments to show that **known orientation is not the factor that boosts the student model's performance.** (1) We train the CCVPE student model with only orientation loss using ground truth orientation. For various learning rates, we find the localization test error increases compared to the teacher model. Thus, simply exposing the student model to the target domain orientation does not improve its localization performance. (2) We simplify the CCVPE method into a localization-only method by removing its orientation-related components and then use this setting for both teacher and student models. Our method achieves a similar performance boost from 4.33 to 3.87 m (original with orientation: from 4.38 to 3.85 m). So, its gains are not due to the student's orientation supervision.

H. Extra qualitative results of teacher and student models

Then, we visualize more teacher and student models' predictions. In Figure 7, examples (a) to (e) show the situation where the teacher model's prediction contains multi-modal uncertainty, and the predicted location is in the wrong mode. After knowledge distillation, our student model assigns a higher probability at the correct mode. In example (f), the teacher's prediction is accurate, and the student model maintains this accurate prediction. Lastly, we showcase challenging scenarios where there is a lack of discriminative features. In Figure 7 (g), the buildings in the aerial view mostly contain repetitive patterns. Although the teacher model picks a location close to the ground truth and the student has a higher error in this example, the inherent uncertainty in both the teacher's and student's heat maps is large. In example (h), the teacher model focuses on the street in the middle, and our student model explores more streets. The appearance of the vegetation in the aerial view looks similar, and both teacher and student models output a wrong location. We expect both challenging scenarios can be addressed by using a sequence of ground-level images, and we will explore this in future work.

I. Potential Negative Impact

Our paper proposed a weakly-supervised learning technique that enhances the localization accuracy of pre-trained fine-grained cross-view localization models. Fine-grained cross-view localization techniques raise the risk of exposing precise location information of individuals. For instance, mobile phone images, such as those from iPhones, often include a GNSS geo-tag in their metadata. This approximate location can be utilized to identify a local aerial image patch, thereby allowing fine-grained cross-view localization to pinpoint the exact location where the image was captured. Consequently, hackers could exploit this method to track individuals, such as social media influencers, by accessing the images they share

online. This presents security and privacy concerns. To counter these risks, social media platforms should alert users to the potential for location data leakage and provide features that enable the removal of geo-tags from images upon upload.

J. Limitations

In knowledge self-distillation, it is often required that the initial model is at a "good enough" starting point, otherwise, it will not converge to a better solution. This requirement also applies to the method we propose. We conducted experiments where a teacher model, trained on one dataset such as KITTI [3], was used to generate pseudo ground truth to train a student model on a different dataset, for instance, the Ford dataset [1]. In this case, the teacher's predictions on the target dataset were not much better than random guesses, making our method not applicable. When the training and test sets are from different datasets, the teacher fails in the target area since the domain gap comes not only from different areas, but also from different sensors, and different resolutions of aerial images. In our work, we target the domain gap between different areas but for the same sensor setup.

References

1. Agarwal, S., Vora, A., Pandey, G., Williams, W., Kourous, H., McBride, J.: Ford multi-av seasonal dataset. International Journal of Robotics Research **39**(12), 1367–1376 (2020)
2. Ben-Moshe, B., Elkin, E., et al.: Improving accuracy of GNSS devices in urban canyons. In: Proceedings of the 23rd Annual Canadian Conference on Computational Geometry. pp. 511–515 (2011)
3. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. International Journal of Robotics Research (2013)
4. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. Advances in Neural Information Processing Systems **17** (2004)
5. Ji, M., Shin, S., Hwang, S., Park, G., Moon, I.C.: Refine myself by teaching myself: Feature refinement via self-knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10664–10673 (2021)
6. Litrico, M., Del Bue, A., Morerio, P.: Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7640–7650 (2023)
7. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
8. Shannon, C.E.: A mathematical theory of communication. ACM SIGMOBILE mobile computing and communications review **5**(1), 3–55 (2001)
9. Shi, Y., Wu, F., Perincherry, A., Vora, A., Li, H.: Boosting 3-dof ground-to-satellite camera localization accuracy via geometry-guided cross-view transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21516–21526 (2023)

10. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2517–2526 (2019)
11. Wang, Y., Peng, J., Zhang, Z.: Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9092–9101 (2021)
12. Xia, Z., Booij, O., Kooij, J.F.P.: Convolutional cross-view pose estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
13. Zhou, Q., Feng, Z., Gu, Q., Cheng, G., Lu, X., Shi, J., Ma, L.: Uncertainty-aware consistency regularization for cross-domain semantic segmentation. Computer Vision and Image Understanding **221**, 103448 (2022)

Fig. 7: Teacher and student models' predictions on VIGOR test set. The red color denotes the localization probability, and a darker color means a higher probability. (a)-(f): success cases. (g) and (h): failure cases.