GeoDistill: Geometry-Guided Self-Distillation for Weakly Supervised Cross-View Localization

Teacher Param Update	Cross-Area Mean(↓)	Cross-Area Median(↓)	Same-Area Mean(↓)	Same-Area Median(↓)
Fixed	4.65	1.28	4.46	1.43
Prev. Student	5.02	1.37	4.49	1.44
EMA	**4.49**	**1.22**	**4.26**	**1.37**
Baseline	5.20	1.44	4.81	1.61

Table 1. Localization performance comparison with different update strategies for the teacher network in the VIGOR dataset [**?**]. Prev Student means using the student from the last epoch as the teacher for the current epoch. Fixed means the teacher network does not update during training. EMA refers to the exponentially moving average teacher adopted in our method.

1. Teacher-student Parameter Update Strategy

We investigate different strategies for updating the teacher model's parameters, including keeping the teacher model's parameters fixed, denoted as "Fixed", and using the student model's parameters from the last epoch as the teacher model's parameters for the current epoch, denoted as "Prev Student". As shown in Tab. 1, all these different teacher parameters update strategy improves the performance over the baseline model, demonstrating the effectiveness of our key idea: using different FoVs to create a discrepancy between teacher and student models, and this discrepancy works effectively as a learning signal to encourage the model focusing on discriminative local features that are useful for cross-view matching. Compared to Fixed and our EMA parameters update strategy, Prev. Student suffers from abrupt parameters shifts, which causes significant location prediction inconsistency (before and after teacher parameters update) for some examples, resulting in inconsistent supervision which negatively affects the magnitude of the performance improvement. In contrast, our EMA update strategy combines the merits of Fixed and Prev. Student. It inherits the stability of a fixed teacher model while also adaptively integrating the student's refined knowledge, resulting in the most considerable performance improvement.

2. Different Training Objectives for Self-Distillation

To evaluate the impact of different training objectives, we performed an ablation study comparing Cross-Entropy (CE) and Kullback-Leibler Divergence (KLD) as loss functions for our student network. Table 2 shows that CE and KLD achieve a similar localization accuracy.

Loss	Cross-Area Mean(↓)	Cross-Area Median(↓)	Same-Area Mean(↓)	Same-Area Median(↓)
KLD	4.50	1.22	4.25	1.37
CE (ours)	4.49	1.22	4.26	1.37

Table 2. Localization performance comparison with different training objective in VIGOR dataset.

3. Comparison with Fully Supervised Methods in VIGOR Same Area Test Set

Table 3. Localization performance comparison on VIGOR Same Area test set. **Best in bold**. <u>The second-best is underlined</u>. Here, "*" indicates fully supervised methods.

Noise	Method	↓Localization Mean	↓Localization Median	↓Orientation Mean	↓Orientation Median
0°	CVR[39]*	8.82	7.68	-	-
	SliceMatch[18]*	5.18	2.58	-	-
	Boosting[26]*	4.12	<u>1.34</u>	-	-
	CCVPE[33]*	<u>3.60</u>	1.36	10.59	5.43
	HC-Net[31]*	**2.65**	**1.17**	1.92	1.04
	GeoDistll(ours)	4.26	1.37	-	-
±45°	CCVPE[33]*	<u>3.50</u>	<u>1.39</u>	10.56	5.96
	HC-Net[31]*	**2.70**	**1.18**	**2.12**	**1.04**
	GeoDistll(ours)	4.71	1.48	<u>2.90</u>	<u>1.11</u>

Here, we supply the comparison of GeoDistill using G2SWeakly as backbone with state-of-the-art fully supervised methods in the Same-Area setting of the VIGOR dataset in Tab. 3. As anticipated, fully supervised methods generally perform better in the same-area setting than our weakly supervised GeoDistill framework. This is because

fully supervised methods are trained with precisely annotated ground truth data within the same geographic area used for testing, allowing them to effectively learn area-specific features and optimize for performance within the training distribution. In contrast, GeoDistill, trained with weakly supervised noisy GPS data and designed for cross-area generalization, is not explicitly optimized for same-area performance.

For the cross-area evaluation, as highlighted in the main paper, GeoDistill achieves the second-best performance among the compared fully supervised approaches, highlighting its excellent generalization ability compared to fully supervised approaches.

4. Evaluating GeoDistill with Unlabeled Target Domain Data

For completeness, we evaluated GeoDistill under the unlabeled target domain data assumption of [34]. Intriguingly, retraining CCVPE [33] with GeoDistill yielded similar performance using either source (4.05m mean error) or target domain data (3.95m) without GT, aligning with [34]'s weakly supervised distillation (3.85m) for domain adapting. However, target domain data availability is often impractical, limiting real-world applicability. Moreover, while [34] uses reliable teacher predictions as pseudo-labels for retraining, generalization to truly unseen regions remains a concern. Like fully supervised methods, target-domain fine-tuning approaches risk performance degradation when encountering new out-of-distribution data. Conversely, GeoDistill's key advantage is achieving significant generalization gains by retraining solely on source domain data. This enables robust generalization to arbitrary unseen cities, offering a more scalable and practical solution. GeoDistill's ability to match target-domain adaptation performance without requiring target data, while ensuring superior generalization, underscores its practical utility and generalization prowess.