TU Delft

Delft University of Technology

Document Version
Final published version

Citation (APA)
Xia, Z., Shi, Y., Li, H., & Kooij, J. F. P. (2025). Adapting Fine-Grained Cross-View Localization to Areas Without Fine Ground Truth. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXXI* (pp. 397-415). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 15089 LNCS). Springer. https://doi.org/10.1007/978-3-031-72751-1_23

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' - Taverne project

https://www.openaccess.nl/en/you-share-we-take-care

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.

Adapting Fine-Grained Cross-View Localization to Areas Without Fine Ground Truth

Zimin Xia[1(✉)], Yujiao Shi[2], Hongdong Li[3], and Julian F. P. Kooij[4]

[1] École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
zimin.xia@epfl.ch
[2] ShanghaiTech University, Shanghai, China
[3] Australian National University, Canberra, Australia
[4] Delft University of Technology, Delft, The Netherlands

Abstract. Given a ground-level query image and a geo-referenced aerial image that covers the query's local surroundings, fine-grained cross-view localization aims to estimate the location of the ground camera inside the aerial image. Recent works have focused on developing advanced networks trained with accurate ground truth (GT) locations of ground images. However, the trained models always suffer a performance drop when applied to images in a new target area that differs from training. In most deployment scenarios, acquiring fine GT, *i.e.* accurate GT locations, for target-area images to re-train the network can be expensive and sometimes infeasible. In contrast, collecting images with noisy GT with errors of tens of meters is often easy. Motivated by this, our paper focuses on improving the performance of a trained model in a new target area by leveraging only the target-area images without fine GT. We propose a weakly supervised learning approach based on knowledge self-distillation. This approach uses predictions from a pre-trained model as pseudo GT to supervise a copy of itself. Our approach includes a mode-based pseudo GT generation for reducing uncertainty in pseudo GT and an outlier filtering method to remove unreliable pseudo GT. Our approach is validated using two recent state-of-the-art models on two benchmarks. The results demonstrate that it consistently and considerably boosts the localization accuracy in the target area.

1 Introduction

Visual localization, a fundamental task in vision and mobile robotics, aims to identify the location of a camera only from the images it takes. Commonly, the image is compared to a pre-constructed map. However, constructing a suitable map with traditional survey-grade mapping vehicles (often equipped with cameras, LiDAR, and high-precision GNSS sensors) is both laborious and expensive. On the other hand, aerial or satellite images provide global coverage and become more easily accessible, making them promising map sources. In this

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-72751-1_23.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Leonardis et al. (Eds.): ECCV 2024, LNCS 15089, pp. 397–415, 2025.
https://doi.org/10.1007/978-3-031-72751-1_23

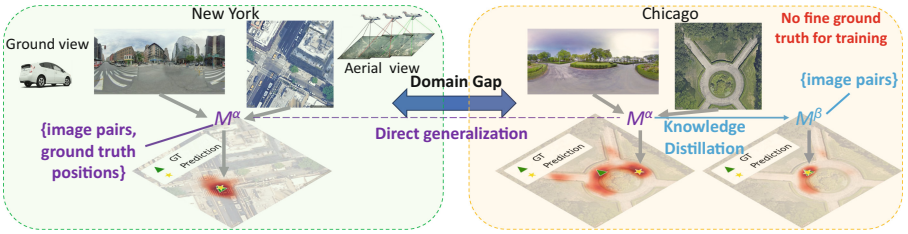


Fig. 1. Learning-based cross-view localization models often perform well when test images are from the same area used in training, as shown in the green box. When inference in a new target area where no fine ground truth is available, the standard practice (in purple) directly deploys a model trained in a different area, leaving an obvious domain gap. Due to this domain gap, the direct generalization often results in a performance drop, causing uncertain or erroneous predictions. To address this, we propose a knowledge self-distillation-based weakly-supervised learning approach (in cyan) to adapt the model to the target area using only ground-aerial image pairs without fine ground truth locations. This leads to better localization performance (Color figure online).

work, we focus on the task of fine-grained cross-view localization to pinpoint the precise geospatial location of a ground camera within a geo-referenced aerial image patch covering local surroundings. The key underlying assumption of this task [10,22,37,39,51,56,59] is that although we *do not* have an accurate fine-grained location of the ground camera, we *do* have a *noisy* localization prior available at inference time to identify the aerial image that covers the ground camera's location. For applications such as autonomous driving, fine-grained cross-view localization is a viable supplement to traditional positioning sensors, such as GNSS, especially in urban canyons where the GNSS positioning error can reach tens of meters [3].

As shown in Fig. 1, there are two main scenarios in cross-view localization. **(1)** Same-area testing (Fig. 1, green box): When the fine ground truth, *i.e.* the accurate location of the ground camera, is available in the target area, a cross-view localization model can be trained on this data and then deployed for inference on new test images. **(2)** Cross-area testing (Fig. 1, yellow box, left): When there is no fine ground truth in the target area, it is common to train the model on images from a different area for which fine ground truth is available, and then the trained model is directly deployed in the target area. Because of the domain gap between the two areas, the predicted location becomes less reliable. Although many works [10,22,37,39,56,59,71] have been proposed for fine-grained cross-view localization, they all suffer from this performance drop when directly deploying in a new target area. Nevertheless, this cross-area scenario is more realistic for real-world use cases, since collecting fine ground truth of every region is expensive and sometimes infeasible. Recent works [10,22,37] even found errors in ground truth locations in popular datasets [1,12,53,71]. Therefore, an alternative to fully-supervised training on fine ground truth is needed to scale cross-view localization models to larger areas.

We propose to address this problem of cross-area localization by relying on the exact same key assumption in the fine-grained cross-view localization task. Namely, it is straightforward to collect ground images with noisy ground truth, *i.e.* the rough location of the ground camera, at a new area to identify the local aerial image patch. For instance, inaccurate GNSS measurements in urban canyons are unreliable as fine ground truth [3], but can still be used as noisy localization prior. Then, our goal is to *improve a pre-trained model's localization performance in the target area by leveraging only the ground-aerial image pairs in the target area, without associated fine ground truth locations*[1].

For this goal, we adopt knowledge self-distillation [11,49] to finetune a fine-grained cross-view localization model in a weakly-supervised manner in which only rough location is used for pairing the ground and aerial images. We use a model pre-trained from another area as the teacher model to generate pseudo ground truth for the target-area images and use it to train a student model, which is initialized as a copy of the teacher model. Since the teacher's output can be uncertain in the target area, directly using it as pseudo ground truth might reinforce incorrect localization estimates and lead to sub-optimal results. We address this by introducing methods to reduce the uncertainty and filter out the outliers in the pseudo ground truth. Concretely, our contributions are[2]:

(1) We propose a knowledge self-distillation-based weakly-supervised learning approach that considerably improves models' localization performance in a new area by only leveraging the ground-aerial image pairs without ground truth locations. The proposed approach is validated using two state-of-the-art methods on two benchmarks. **(2)** For methods with coarse-to-fine outputs, we investigate how to reduce the uncertainty and suppress the noise in teacher model's predictions. Using our proposed single-modal pseudo ground truth leads to a better student model than using the multi-modal heat maps from the teacher model. **(3)** We design a simple but effective method for filtering outliers in the pseudo ground truth. Training with filtered pseudo ground truth further improves the localization accuracy of the student model.

2 Related Work

Cross-View Localization is formulated differently depending on the use case. For large-scale coarse localization, a common formulation is image retrieval [18, 23,25,33,38,40,46,54,61,70]. In this setting, the continuous aerial imagery is divided into small patches. The ground query image's location is approximated by the retrieved patch's geolocation. However, for fine-grained localization, image retrieval methods need to sample the patch densely [57,58], and it increases both computation and storage usage.

[1] Recent models need the ground camera's orientation for training. We assume the camera orientation is known since it can be acquired easily, *e.g.* by the digital compass in a mobile phone or a vehicle.

[2] Our code is available at: https://github.com/tudelft-iv/Adapting_CVL.

Recently, there have been increasing attempts to estimate the precise location directly, sometimes together with the orientation, of the ground camera on a known aerial image patch. In [71], the location offset between the ground query and the aerial image is regressed based on their image descriptors. Instead of regression, [59] formulated the localization task as a dense classification problem to capture the multi-modal localization uncertainty. Later, this idea is extended by [56] to include coarse-to-fine predictions and build orientation equivariant ground image descriptors. Several works [37,41,51] explored the geometry transformation between ground and aerial views. [37] estimated the ground camera pose using the iterative Levenberg–Marquardt algorithm and [51] made use of a deep homography estimator [6] to infer the ground camera pose. In [10,35,36,39], the ground camera pose is estimated by densely comparing a Bird's Eye View (BEV) representation constructed using ground images to an aerial representation. SliceMatch [22] took an efficient generative testing approach to select the most probable pose from a candidate set. Commonly, the localization output is represented as a heat map [10,22,39,51,56,59], where the value at each location (*i.e.* pixel in the aerial image) denotes how likely the ground camera locates there, and state-of-the-arts [39,56] construct the heat map in a coarse-to-fine manner. Despite extensive methodological consideration, the performance of the above approaches dropped considerably when directly generalizing to images collected in an area that differs from the training set. We aim to bridge this gap.

Unsupervised Domain Adaptation (UDA) is a well-studied problem in many other vision tasks [50,66]. The objective is to adapt a model trained in the source domain to the target domain without labels from the target domain, such that the adapted model can perform well on the test samples from the target domain. More specifically, UDA can be categorized as source-free [20,24,28,67] and non-source-free [7,16,19,21,45,52,55,69] depending on if the source domain labels are used during adaptation. To minimize the discrepancy between features from the source and target domain, some works [13,26,44] use manually crafted metrics to measure this discrepancy. Adversarial methods [27,47,65] deploy a discriminator to achieve this. [4,42,48] observed that predictions in the target domain often contain more uncertainty than those in the source domain. Hence, additional objectives, e.g. entropy minimization [15], are included for training the model in a semi-supervised manner using images from both the source and target domain. Another promising type of domain adaptation is based on pseudo labels [60,64]. It bears similarities to **knowledge distillation** (KD) [5]. KD's primary objective is to transmit the knowledge acquired by a more comprehensive teacher model to a smaller student model [14,49]. Knowledge self-distillation, in which the teacher and student share the same architecture, is a special branch of KD pioneered by Born-Again Networks [11]. The key idea is to use the model from the previous step to generate pseudo labels for training the model at the current step. Recent works [2,8,17,19,43,62,63] also tried to use the information from deeper layers to supervise the shallower layers inside the model. To apply KD for UDA, the teacher model generates pseudo labels in the target domain

to adapt the student model [9,30,68]. Since the pseudo labels are not always reliable, uncertain ones (*e.g.* with high entropy [24,52,69]) are often removed in student learning. However, such uncertainty measures are developed for purely categorical tasks and do not consider any spatial ordering between classes, as is needed for localization methods that produce heat maps: 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, but the latter has more localization uncertainty.

3 Methodology

The most desirable real-world setup is to adapt a pre-trained model to the target area without requiring access to (perhaps licensed or high-volume) source-domain data. Our scope is thus source-free UDA. We first formalize the fine-grained cross-view localization task. Then, we introduce our proposed approach.

3.1 Task Definition

Given a ground-level image G and an aerial image A that covers the local surroundings of G, the task of fine-grained cross-view localization is to determine the image coordinates $\hat{y} = (\hat{u}, \hat{v})$ of the ground camera within aerial image A, where $\hat{u} \in [0, 1]$ and $\hat{v} \in [0, 1]$. Recent methods [10,22,39,56,59] achieve this task by training a deep model $\mathcal{M}(G, A)$ which predicts a *heat map* H to capture the underlying localization confidence over spatial locations, and the most confident location can be used as predicted location y,

$$H = \mathcal{M}(G, A), \quad y = \arg\max_{u,v}(H(u, v)). \tag{1}$$

To optimize the model's parameters θ_α with respect to a model specific loss functions $\mathcal{L}_\mathcal{M}$, an annotated dataset of a set of N_α ground-aerial image pairs, $\mathbb{I}_\alpha = \{\{G_1, A_1\}, ..., \{G_{N_\alpha}, A_{N_\alpha}\}\}$, and their corresponding fine ground truth $\mathbb{Y}_\alpha = \{\hat{y}_1, ..., \hat{y}_{N_\alpha}\}$ is used,

$$\theta_\alpha = \arg\min_\theta \mathbb{E}_{\{G,A\}\in\mathbb{I}_\alpha, \hat{y}\in\mathbb{Y}_\alpha} \left[\mathcal{L}_\mathcal{M}(\mathcal{M}(G, A \mid \theta), \hat{y})\right]. \tag{2}$$

The training image set $\mathbb{I}_\alpha$ consists of samples drawn from a true distribution $\mathcal{D}_\alpha$ representing a specific geographic area α, *i.e.* $\mathbb{I}_\alpha \overset{\text{i.i.d.}}{\sim} \mathcal{D}_\alpha$. When the model is deployed, the test image set $\mathbb{I}_{test}$ can either come from the *same area* α, or a new environment β. As motivated before, we focus on the *cross-area* setting, namely $\mathbb{I}_{test}$ is from the target area β, *i.e.* $\mathbb{I}_{test} \overset{\text{i.i.d.}}{\sim} \mathcal{D}_\beta$. Because of the domain gap, $\mathcal{D}_\beta \neq \mathcal{D}_\alpha$, directly deploying the trained model $\mathcal{M}^\alpha := \mathcal{M}(\cdot \mid \theta^\alpha)$ on test set $\mathbb{I}_{test}$ as in current practice is sub-optimal.

It is important to note that standard fine-grained cross-view localization [10, 22,39,56] assumes the pairing between ground and aerial images is known during inference, as collecting ground-level images with rough location estimates in the

target area is often easy. Therefore, we propose to consider the easily available pairing information for weakly-supervised learning by collecting another set of images $\mathbb{I}_\beta = \{\{G_1, A_1\}, ..., \{G_{N_\beta}, A_{N_\beta}\}\}$ from the target area β, $\mathbb{I}_\beta \overset{\text{i.i.d.}}{\sim} \mathcal{D}_\beta$, without corresponding fine ground truth $\mathbb{Y}_\beta$. As noted before, the orientation of the ground camera is assumed known.

Our objective is then to adapt a fine-grained cross-view localization model $\mathcal{M}^\alpha$ to the target area β by leveraging the image set $\mathbb{I}_\beta$ without fine ground truth locations such that the model performance on $\mathbb{I}_{test}$ can be improved.

3.2 UDA for Cross-View Localization

So far, no prior work addressed the task of adapting fine-grained cross-view localization to new areas without fine ground truth. To decide on a suitable UDA approach, we first note that heat maps of state-of-the-art models reflect more uncertainty for cross-area samples than for same-area samples [22,39,56]. The higher uncertainty results in more small positional errors, but also more modes in the heat map, yielding more outliers with large positional errors.

We therefore consider UDA techniques that can help reduce the uncertainty. One option is *entropy minimization* [15], *i.e.* to directly deploy the trained model $\mathcal{M}^\alpha$ on the image set $\mathbb{I}_\beta$ and encourage the final output heat map H to be more certain by minimizing its entropy. However minimizing the entropy does not necessarily encourage the model to converge towards the correct location for $\{G, A\} \in \mathbb{I}_\beta$, as the model may just as well become more confident about the outliers. Our experiments shall validate entropy minimization's shortcomings.

We instead propose to pursue *knowledge self-distillation* [62] for our task. The trained model $\mathcal{M}^\alpha$ from the source area α can be used as the teacher model to generate *pseudo ground truth* X for image set $\mathbb{I}_\beta$ to train a student model $\mathcal{M}^\beta$. Here, we consider X as a target heat map with the same spatial resolution as the aerial image A. The student model has the same architecture as the teacher model and is initialized using the teacher model's weights θ_α. Encouraging the outputs of the student model to mimic X can improve the accuracy of the student model on images from β if we control the generation of pseudo ground truth to suppress unwanted modes and select reliable samples.

Finally, we point out that the recent state-of-the-art methods [39,56] have K coarse-to-fine heat map outputs, *i.e.* $\mathbb{H} = \mathcal{M}(G, A)$ and $\mathbb{H} = \{H_1, ..., H_K\}$. The spatial resolution of the next level heat map is higher than that of the previous level, *i.e.* $\text{res}(H_{k+1}) > \text{res}(H_k)$, where k is the index for the level and res() returns the spatial resolution. The final predicted location then becomes $y = \arg\max_{u,v}(H_K(u, v))$. For other applications with coarse-to-fine models, encouraging shallower layers' activation to mimic deeper layers' activation can bootstrap model performance [62]. Similarly, knowledge self-distillation for cross-view localization may also exploit such coarse-to-fine maps.

3.3 Proposed Approach

Usually, the deeper layers in the model have access to more information than the shallower layers, *e.g.* the fine-grained scene layout information passed by the skipped connections, as in UNet [34]. Hence, the output from deeper layers can be more precise than that from shallower layers. We therefore propose to follow the "Best Teacher Distillation" paradigm [62] and generate pseudo ground truth X from only the highest-resolution heat map predicted by the teacher model.

A naive approach is, for any $\{G, A\} \in \mathbb{I}_\beta$, using simply $X := H_K^\alpha$ from teacher output[3] $\{H_1^\alpha, \cdots, H_K^\alpha\} = \mathcal{M}^\alpha(G, A)$. Then, this high-resolution pseudo ground truth X is down-sampled to create a set of pseudo ground truth heat maps $\mathbb{P} = \{P_1, ..., P_K\}$ to supervise the student model at all levels,

$$P_k = \text{downsample}_k(X) \quad \text{s.t.} \quad \text{res}(P_k) = \text{res}(H_k). \tag{3}$$

The set $\boldsymbol{\mathcal{P}}_\beta = \{\mathbb{P}_1, ..., \mathbb{P}_{N_\beta}\}$ is the complete pseudo ground truth for image set $\mathbb{I}_\beta$ in the target area for training the student model, where N_β is the number of the ground-aerial image pairs in $\mathbb{I}_\beta$.

However, since the pseudo ground truth X contains errors, directly following this naive approach might propagate the errors to the student model $\mathcal{M}^\beta$. Thus, we present several strategies to reduce the teacher's uncertainty, and deal with noise and large outliers in X. Our proposed designs are highlighted in bold in the overview of our weakly-supervised learning approach in Fig. 2.

Coarse-Only Supervision: Standard Best Teacher Distillation [62] suggests supervising heat maps at all levels of the student model using the pseudo ground

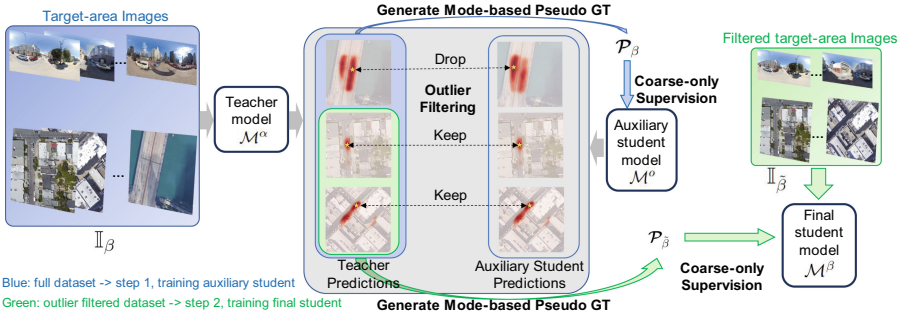


Fig. 2. Overview of our approach. We first employ a teacher model trained on data from another area to generate pseudo GT, $\boldsymbol{\mathcal{P}}_\beta$, on target-area images (in blue). The pseudo GT is then used to train an auxiliary student model $\mathcal{M}_o$. After that, we compare the predictions from the teacher model and those from the auxiliary student model to filter out unreliable teacher predictions (the middle grey box). The remaining samples with their pseudo GT, $\boldsymbol{\mathcal{P}}_{\tilde{\beta}}$, are used to train our final student model $\mathcal{M}_\beta$ (in green) (Color figure online).

[3] Note that we use superscript α to indicate output from model $\mathcal{M}^\alpha$.

truth. However, the spatial accuracy of X is limited, and using X to supervise the high-resolution outputs of the student model might propagate this noise. We note that the down-sampling in Eq. 3 suppresses such positional noise at the lower resolution P_k. Thus using only the lower level P_k might lead to a better student model. We therefore consider to only compute the loss on student model's outputs $\mathbb{H}^\beta = \mathcal{M}^\beta(G, A)$ up to a certain level $K' \leq K$,

$$\mathcal{L}(\mathbb{H}^\beta, \mathbb{P}) = \frac{1}{K'} \sum_{k=1}^{K'} \mathcal{L}_k(H_k^\beta, P_k). \tag{4}$$

Here, K' is a hyperparameter. $\mathcal{L}_k(H_k^\beta, P_k)$ is a weighted sum of infoNCE losses [31], similar to the training in [56,59], except we use pseudo ground truth P_k as weight,

$$\mathcal{L}_k(H_k^\beta, P_k) = \frac{1}{\sum P_k} \sum_{m,n} P_k^{m,n} \cdot \mathcal{L}_{\text{infoNCE}}(H_k^\beta \mid (m, n)). \tag{5}$$

$\mathcal{L}_{\text{infoNCE}}(H_k^\beta \mid (m, n))$ is an infoNCE loss interpreting H_k^β as similarity scores, location (m, n) as the positive class, and all other locations as negative classes.

Mode-Based Pseudo Ground Truth: Rather than using H_K^α directly as pseudo ground truth X, we propose to create a "clean" pseudo ground truth X that only represents its mode $y^\alpha = \arg\max(H_K^\alpha)$. We thus provide the student with a training objective that represents less uncertainty for the target domain input than its teacher. Still, it is common when training fine-grained cross-view localization models, to apply Gaussian label smoothing [10,59] even with reliable ground truth to aid the learning objective and increase robustness to remaining errors in the annotation [29]. We similarly apply Gaussian label smoothing centered at y^α,

$$X(u, v) = \mathcal{N}((u, v) \mid y^\alpha, I_2\sigma^2), \ \text{res}(X) = \text{res}(A). \tag{6}$$

The standard deviation σ is a hyperparameter and I_2 is a 2D identity matrix.

Outlier Filtering: Recent deep learning advances [32] highlighted the importance of using curated data. Motivated by this principle, we prefer having fewer but more reliable samples of the target domain, over having more samples but with potentially large errors in the pseudo ground truth. The *Mode-based Pseudo Ground Truth* could force a sample's ground truth to commit to a wrong (outlier) location, therefore we seek to filter out such samples.

We here make another observation: samples where the predicted locations y^α of a teacher and y^β of a student greatly differ, the teacher's predictions were more likely to be outliers compared to samples where the teacher and student's predicted locations are more consistent, as we will demonstrate in our experiments. Thus, we propose to first train another auxiliary student model $\mathcal{M}^o$ on all target domain data, and compare its predictions to the teacher's to identify stable ones with little change in the predicted location. Then, we only use those

reliable non-outlier samples to train the final student model $\mathcal{M}^{\beta}$. Concretely, we first optimize the auxiliary student model $\mathcal{M}^{o}$ on all $\mathbb{I}_{\beta}$ with $\boldsymbol{\mathcal{P}}_{\beta}$ using,

$$\theta_o = \arg\min_{\theta} \mathbb{E}_{\{G,A\}\in\mathbb{I}_{\beta},\mathbb{P}\in\boldsymbol{\mathcal{P}}_{\beta}} \left[\mathcal{L}(\mathcal{M}(G, A \mid \theta), \mathbb{P})\right]. \tag{7}$$

Then, we calculate the L2-distance $d^{\alpha,o} = \|\mathbf{y}^{\alpha} - \mathbf{y}^{\mathbf{o}}\|_2$ between the image coordinates predicted by $\mathcal{M}^{\alpha}$ and $\mathcal{M}^{o}$ to find the potential unreliable $\mathbb{P}$. The resulting distance set $\mathbb{D} = \{d_1^{\alpha,o}, ..., d_{N_\beta}^{\alpha,o}\}$ is used to keep the top-T% samples in $\mathbb{I}_{\beta}$ that have the smallest T% distance $d^{\alpha,o}$. Denoting the filtered image set as $\mathbb{I}_{\tilde{\beta}}$ and corresponding pseudo ground truth as $\boldsymbol{\mathcal{P}}_{\tilde{\beta}}$, the final student model $\mathcal{M}^{\beta}$ is optimized using Eq. 7 by substituting $\mathbb{I}_{\beta}$ with $\mathbb{I}_{\tilde{\beta}}$ and $\boldsymbol{\mathcal{P}}_{\beta}$ with $\boldsymbol{\mathcal{P}}_{\tilde{\beta}}$.

4 Experiments

We first introduce the two used datasets and our evaluation metrics. Then, we discuss two state-of-the-art methods [39,56], based on which the proposed weakly-supervised learning is evaluated, followed by implementation details. After this, we provide the test results and a detailed ablation study.

4.1 Datasets

We focus on the cross-area split of VIGOR [71] and KITTI [12] datasets.

VIGOR dataset contains ground-level panoramic images and their corresponding aerial images collected in four US cities. In its cross-area split, the training set contains images from two cities, and the test set is collected from two other cities. We use the training set to train the teacher model and focus on the cross-area setting in our experiments. To compare direct generalization and our proposed weakly-supervised learning, we conduct a 70%, 10%, and 20% split on the original cross-area test set to create our weakly-supervised training set (no ground truth locations), validation set, and test set. We use the validation set for finding the stopping epoch during training, as well as for conducting the ablation study. Our test set is used for benchmarking our method. We use the improved VIGOR labels provided by [22].

KITTI dataset contains ground-level images with a limited field of view. We use the aerial images provided by [37] and adopt their cross-area setting, where the training and test images are from different areas. Similar to our settings on the VIGOR dataset, we use the training set to train the teacher model and then split the original cross-area test set into 70%, 10%, and 20% for weakly-supervised training of the student model, validation, and testing.

4.2 Evaluation Metrics

We measure the displacement error ϵ in meters between the predicted location and the ground truth location, *i.e.* $\epsilon = s\|y - \hat{y}\|_2$, where s is the scaling factor

from image coordinates to real-world Euclidean coordinates. Then, mean and median displacement errors over all samples are reported as our evaluation metrics. On the KITTI dataset, we further decompose the displacement errors into errors in the longitudinal direction (along the camera's viewing direction, typically along the road), and errors in the lateral direction (perpendicular to the viewing direction), following the common evaluation protocol [22,37,56].

4.3 Backbone State-of-the-Art Methods

Two state-of-the-arts, Convolutional Cross-View Pose Estimation (CCVPE) [56] and Geometry-Guided Cross-View Transformer (GGCVT) [39] are used to test our proposed weakly-supervised learning approach. Both methods were proposed for fine-grained cross-view localization and orientation estimation, and have a coarse-to-fine architecture. CCVPE has two separate branches for localization and orientation prediction. GGCVT uses an orientation estimation block before its location estimator. In this work, we use them for localization only. CCVPE has seven levels of heat map outputs, in which the first six heat maps are 3D, with the first two dimensions for localization and the third dimension for orientation. The last heat map is 2D. GGCVT has three levels of 2D heat map outputs.

4.4 Implementation Details

We use the official code of CCVPE [56] and GGCVT [39] for model implementations. Auxiliary and final student models are trained following our proposed approach. For CCVPE's 3D heat map output, we simply lift the pseudo ground truth heat map P_k to 3D using the known orientation as done in [56].

The hyperparameters K', T, and σ are tuned on the VIGOR validation set. For CCVPE, we find that including the first two levels of losses, *i.e.* $K' = 2$, and $T\% = 80\%$ gives the lowest mean localization error. For GGCVT, we use all three levels of losses, *i.e.* $K' = 3$, and $T\% = 70\%$. We tested $\sigma = 1, 4, 8, 12, 20$ pixels, and one-hot pseudo ground truth. Because of $\sigma = 4$ gave the best validation result, it is used for both methods. The same setting is directly applied to KITTI.

4.5 Results

We compare the trained student models to teacher models (baselines) on the cross-area test set of VIGOR and KITTI datasets. Previous state-of-the-art was set by directly deploying CCVPE and GGCVT teacher models to the target area. On the VIGOR dataset, Table 1 top, the performance of student models trained using proposed weakly-supervised learning surpasses baselines by a large margin. For CCVPE, our approach reduces the mean and median error by 20% and 15% when the orientation of test ground images is unknown. GGCVT only released its code and models for orientation-aligned setting for the VIGOR dataset. Thus, we follow the same setting. In this case, our approach reduces 16% and 5% mean and median error for GGCVT. Without extra hyperparameter tuning, we directly

Table 1. Test results on VIGOR and KITTI. **Best in bold.** Baselines are teacher models (previous state-of-the-art). "Student" denotes models trained using our weakly-supervised learning without ground truth labels. On VIGOR, we provide test results for both known and unknown orientation cases. On KITTI, we test with known orientation.

VIGOR, cross-area test	Known orientation		Unknown orientation	
	Mean (m)	Median (m)	Mean (m)	Median (m)
CCVPE [56]	4.38	1.76	5.35	1.97
CCVPE student (ours)	**3.85** (↓ 12%)	**1.57** (↓ 11%)	**4.27** (↓ 20%)	**1.67** (↓ 15%)
GGCVT [39]	5.19	1.39	-	-
GGCVT student (ours)	**4.34** (↓ 16%)	**1.32** (↓ 5%)	-	-
KITTI, cross-area test	Longitudinal error		Lateral error	
	Mean (m)	Median (m)	Mean (m)	Median (m)
CCVPE [56]	6.55	2.55	1.82	**0.98**
CCVPE student (ours)	**6.18** (↓ 6%)	**2.35** (↓ 8%)	**1.76** (↓ 3%)	**0.98** (↓ 0%)
GGCVT [39]	9.27	4.66	2.19	0.85
GGCVT student (ours)	**8.56** (↓ 8%)	**4.35** (↓ 7%)	**1.90** (↓ 13%)	**0.79** (↓ 7%)

use our proposed approach to train models on KITTI, and it again improves the overall localization performance for both models, see Table 1 bottom.

We also study the gap between each student model to an Oracle, *i.e.* the same method using supervised finetuning on fine ground truth at the target area. Even though the Oracles still achieve lower errors (CCVPE: Oracle 2.31 m *vs.* student 3.85 m; GGCVT: Oracle 2.91 m *vs.* student 4.34 m), we emphasize again that in practice such reliable fine ground truth is generally not available. Importantly, we also find that when the ground truth does contain errors, using supervised finetuning leads to large test errors, see additional results in our Supplementary Material. Instead, our weakly-supervised learning approach scales well because it boosts performance at a low cost: First, there are no extra requirements on the accuracy of localization prior in the target area over previous fine-grained cross-view localization works [10,22,39,51,56,59], as only ground-aerial image pairs are needed. Second, since student models are initialized from their teacher, the training time is short. For example, on VIGOR, using a single V100 GPU our weakly-supervised learning for CCVPE only adds $\sim$ 6 hours of training time (including pseudo ground truth generation and outlier filtering) on top of the direct generalization, which has training time of $\sim$ 16 hours.

Next, we visualize samples where the student model improves over the teacher model. A typical case is shown in Fig. 3 top, in which the teacher model has a multi-modal prediction, and the peak is located in a wrong mode. The student model learned to weigh the modes better after adapting to the target environment. As shown in Fig. 3 bottom, sometimes, even though the teacher model's heat map does not capture the correct location, the student model can still identify it. In this case, the student model might learned discriminative features from

other samples in this area to localize the ground camera. This demonstrates the effectiveness of adapting the student model to the target area by our knowledge distillation process.

4.6 Analysis of Prediction Errors After KD

Following the visual examples, we now analyze the overall statistical relation between the model prediction errors and the change in predicted locations after knowledge distillation. Figure 4 plots this relation for CCVPE. The results for GGCVT are included in our Supplementary Material.

First, we confirm that potential outliers can indeed be identified by the amount of difference between the predicted locations of a teacher and its auxiliary student model in Fig. 4a left. We see there is a large portion of samples located around the diagonal line, *i.e.* $\epsilon^{\alpha} = s \cdot d^{\alpha,o}$. Most samples in $\mathbb{I}_{\beta}$ with large change $d^{\alpha,o}$ in predicted location indeed obtained a large error ϵ^{α} for the teacher model's prediction. Next, Fig. 4a right shows how the difference in location correlates with the prediction error of the auxiliary student. There are more samples being scattered at the bottom of the plot, implying many wrong predictions of the teacher model have already been corrected. Still, our ablation study will demonstrate that using the auxiliary student model directly as a new teacher for a final student model does not work as well as using it for outlier detection. Note that the (less prominent) diagonal line now indicates errors introduced by the auxiliary student model. Lastly, we validate that the final student model reduces the localization error compared to the teacher model on the target test set $\mathbb{I}_{test}$ in Fig. 4b. Comparing the left plot to the right plot, we observe a similar trend

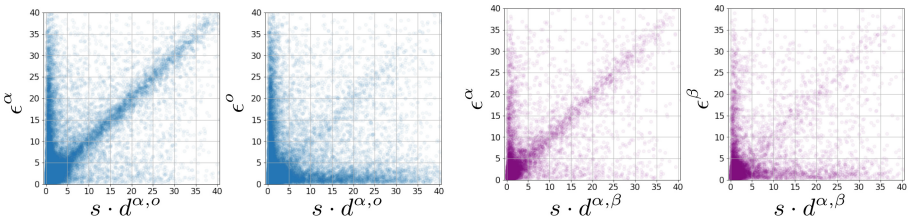


Fig. 3. CCVPE teacher and student model's predictions on VIGOR test set. The red color denotes the localization probability (a darker color means a higher probability) (Color figure online).

as for the auxiliary student model before, namely that many samples with high teacher error in the left plot now obtain low student error in the right plot.

4.7 Entropy Minimization

We also tested entropy minimization (EM) [15] for the CCVPE model on the VIGOR dataset as an alternative domain adaptation technique. We tuned the strength of EM on predicted heat maps of training samples from the target area but found that stronger EM always leads to higher localization errors. The best performance appears when no EM is applied. Therefore, simply exposing the model to the images from the target area and enforcing the confidence of outputs is not sufficient for improving cross-view localization across areas. We also observe that EM makes all heat maps sharper than direct generalization, but does not help the model resolve wrong modes. Our proposed knowledge self-distillation instead reduces uncertainty by filtering out unreliable samples.



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

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

Fig. 4. CCVPE 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.

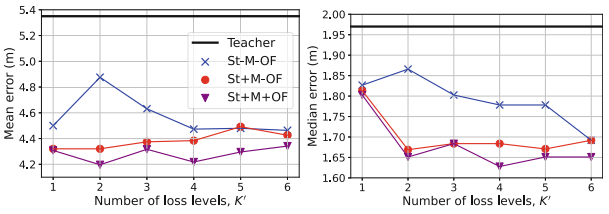


Fig. 5. Ablation study on the proposed mode-based pseudo ground truth, outlier filtering, and different levels for coarse-only supervision in our teacher-student KD using CCVPE.

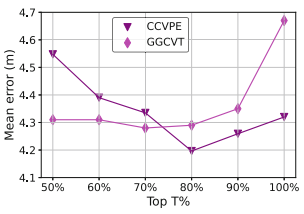


Fig. 6. Effect of T in the proposed outlier filtering. 100% means no outlier filtering.

4.8 Ablation Study

An extensive ablation study is conducted to validate the effectiveness of our proposed designs. We denote the following: **Teacher** (baseline): directly deploy the teacher model $\mathcal{M}^{\alpha}$ in the target area. **St-M-OF**: student model trained using teacher's heat maps, no mode-based pseudo ground truth, no outlier filtering. **St+M-OF**: student model trained using mode-based pseudo ground truth, no outlier filtering. **St+M+OF** (proposed): student model trained using mode-based pseudo ground truth with outlier filtering, *i.e.*the model $\mathcal{M}^{\beta}$.

The performance of these ablation variants when supervising different levels of student predictions of the CCVPE is shown in Fig. 5. It can be seen that the proposed mode-based pseudo ground truth (+M) and outlier filtering (+OF) both improve the performance and the final version, St+M+OF, achieves the best results, no matter how many prediction levels of the student model are supervised. For CCVPE student models, supervising the first $K' = 2$ and $K' = 4$ levels have similar localization performance overall. Since $K' = 2$ gives the lowest mean error, we use it in our final setting. We also tuned K' for GGCVT and found that supervising all three levels, *i.e.* $K' = 3$ gives the best results. The effectiveness of the proposed mode-based pseudo ground truth (+M) and outlier filtering (+OF) on GGCVT is verified in Table 2. When not using any of the proposed designs, *i.e.* GGCVT student model follows Best Teacher Distillation [62], the student's performance (5.34 m) is worse than the Teacher's (5.16 m). This highlights the importance of reducing uncertainty and removing outliers in teacher's predictions. Additionally, we also tried directly using the predictions of the auxiliary student as pseudo ground truth to train the final student model (similar to iterative knowledge self-distillation [11]), denoted as St+M+A in Table 2. However, it does not perform better than using the auxiliary student model for outlier filtering.

Figure 6 shows the ablation study results on different percentage values $T\%$ in our outlier detection. The best CCVPE and GGCVT student models appear at $T\% = 80\%$ and $T\% = 70\%$. In general, there is a trade-off between the quality and quantity of data. When too little data is kept, there is a risk of model overfitting. Filtering out some detected outliers ($20\% \sim 30\%$) improves the quality of the data and can result in better model performance. This suggests that, in practice, blindly increasing the data amount without guaranteeing its quality might negatively influence models' performance.

Table 2. Ablation study for GGCVT. **Best in bold.**

Error (m)	Teacher	St-M-OF	St+M-OF	St+M+A	St+M+OF
Mean	5.16	5.34	4.67	4.54	**4.28**
Median	1.40	1.48	1.32	1.55	**1.28**

5 Conclusion

This paper focuses on improving the localization performance of a pre-trained fine-grained cross-view localization model in a new target area without any fine ground truth. We have proposed a knowledge self-distillation-based weakly-supervised learning approach that only requires ground-aerial image pairs from the target area. Extensive experiments were conducted to study how to generate appropriate pseudo ground truth for student model training. We found that selecting the predominant mode in the teacher model's predictions is better than directly using the output heat maps. Furthermore, supervising coarse-level predictions of a student model using the down-sampled teacher's high-resolution predictions can suppress the positional noise and might lead to a slight boost in the student model's performance. We demonstrated that unreliable target domain samples can be filtered out by comparing predicted locations from teacher and student models, which motivates using an auxiliary student model to curate the data. Training a final student model on the filtered data further improves the localization accuracy. Our proposed approach has been validated on two state-of-the-art methods on two benchmarks. It achieves a consistent and considerable performance boost over the previous standard that directly deploys the trained model in the new target area.

Acknowledgements. This work is part of the research programme Efficient Deep Learning (EDL) with project number P16-25, which is (partly) financed by the Dutch Research Council (NWO).

References

1. Agarwal, S., Vora, A., Pandey, G., Williams, W., Kourous, H., McBride, J.: Ford multi-AV seasonal dataset. Int. J. Robot. Res. **39**(12), 1367–1376 (2020)
2. An, S., Liao, Q., Lu, Z., Xue, J.H.: Efficient semantic segmentation via self-attention and self-distillation. IEEE Trans. Intell. Transp. Syst. **23**(9), 15256–15266 (2022)
3. 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)
4. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: MixMatch: a holistic approach to semi-supervised learning. Adv. Neural Inf. Process. Syst. **32** (2019)
5. Buciluǎ, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 535–541 (2006)
6. Cao, S.Y., Hu, J., Sheng, Z., Shen, H.L.: Iterative deep homography estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1879–1888 (2022)
7. Cardace, A., Spezialetti, R., Ramirez, P.Z., Salti, S., Di Stefano, L.: Self-distillation for unsupervised 3D domain adaptation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 4166–4177 (2023)

8. Ding, Y., Zhu, Q., Liu, X., Yuan, W., Zhang, H., Zhang, C.: KD-MVS: knowledge distillation based self-supervised learning for multi-view stereo. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. ECCV 2022. Lecture Notes in Computer Science, vol. 13691. Springer, Cham. (2022). https://doi.org/10.1007/978-3-031-19821-2_36
9. Feng, H., et al.: KD3A: unsupervised multi-source decentralized domain adaptation via knowledge distillation. In: International Conference on Machine Learning, pp. 3274–3283 (2021)
10. Fervers, F., Bullinger, S., Bodensteiner, C., Arens, M., Stiefelhagen, R.: Uncertainty-aware vision-based metric cross-view geolocalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21621–21631 (2023)
11. Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., Anandkumar, A.: Born again neural networks. In: International Conference on Machine Learning, pp. 1607–1616. PMLR (2018)
12. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: the KITTI dataset. Int. J. Rob. Res. **32**(11) (2013)
13. Gholami, B., Sahu, P., Rudovic, O., Bousmalis, K., Pavlovic, V.: Unsupervised multi-target domain adaptation: an information theoretic approach. IEEE Trans. Image Process. **29**, 3993–4002 (2020)
14. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: a survey. Int. J. Comput. Vis. **129**, 1789–1819 (2021)
15. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. Adv. Neural Inf. Process. Syst. **17** (2004)
16. Guan, D., Huang, J., Xiao, A., Lu, S., Cao, Y.: Uncertainty-aware unsupervised domain adaptation in object detection. IEEE Trans. Multimedia **24**, 2502–2514 (2021)
17. Hou, Y., Ma, Z., Liu, C., Loy, C.C.: Learning lightweight lane detection CNNs by self attention distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1013–1021 (2019)
18. Hu, S., Feng, M., Nguyen, R.M., Lee, G.H.: CVM-Net: cross-view matching network for image-based ground-to-aerial geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7258–7267 (2018)
19. 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)
20. Jing, M., Zhen, X., Li, J., Snoek, C.G.: Order-preserving consistency regularization for domain adaptation and generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 18916–18927 (2023)
21. Lai, Z., et al.: PADCLIP: pseudo-labeling with adaptive debiasing in clip for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16155–16165 (2023)
22. Lentsch, T., Xia, Z., Caesar, H., Kooij, J.F.P.: SliceMatch: geometry-guided aggregation for cross-view pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17225–17234 (2023)
23. Lin, T.Y., Cui, Y., Belongie, S., Hays, J.: Learning deep representations for ground-to-aerial geolocalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5007–5015 (2015)

24. 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)
25. Liu, L., Li, H.: Lending orientation to neural networks for cross-view geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5624–5633 (2019)
26. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: International Conference on Machine Learning, pp. 97–105. PMLR (2015)
27. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. Adv. Neural Inf. Process. Syst. **31** (2018)
28. Lu, Z., Li, D., Song, Y.Z., Xiang, T., Hospedales, T.M.: Uncertainty-aware source-free domain adaptive semantic segmentation. IEEE Trans. Image Process. **32** (2023)
29. Müller, R., Kornblith, S., Hinton, G.E.: When does label smoothing help? Adv. Neural Inf. Process. Syst. **32** (2019)
30. Nguyen-Meidine, L.T., Belal, A., Kiran, M., Dolz, J., Blais-Morin, L.A., Granger, E.: Unsupervised multi-target domain adaptation through knowledge distillation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1339–1347 (2021)
31. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
32. Oquab, M., et al.: DINOv2: learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
33. Regmi, K., Shah, M.: Bridging the domain gap for ground-to-aerial image matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 470–479 (2019)
34. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
35. Sarlin, P.E., et al.: OrienterNet: visual localization in 2D public maps with neural matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21632–21642 (2023)
36. Sarlin, P.E., Trulls, E., Pollefeys, M., Hosang, J., Lynen, S.: SNAP: self-supervised neural maps for visual positioning and semantic understanding. arXiv preprint arXiv:2306.05407 (2023)
37. Shi, Y., Li, H.: Beyond cross-view image retrieval: highly accurate vehicle localization using satellite image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17010–17020 (2022)
38. Shi, Y., Liu, L., Yu, X., Li, H.: Spatial-aware feature aggregation for image based cross-view geo-localization. Adv. Neural Inf. Process. Syst. **32** (2019)
39. 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)
40. Shi, Y., Yu, X., Liu, L., Campbell, D., Koniusz, P., Li, H.: Accurate 3-dof camera geo-localization via ground-to-satellite image matching. IEEE Trans. Pattern Anal. Mach. Intell. **45**(3), 2682–2697 (2022)

41. Shi, Y., Yu, X., Wang, S., Li, H.: CVLNet: cross-view semantic correspondence learning for video-based camera localization. In: Wang, L., Gall, J., Chin, TJ., Sato, I., Chellappa, R. (eds.) Computer Vision – ACCV 2022. ACCV 2022. Lecture Notes in Computer Science, vol. 13841, pp. 123–141. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-26319-4_8
42. Sohn, K., et al.: FixMatch: simplifying semi-supervised learning with consistency and confidence. Adv. Neural. Inf. Process. Syst. **33**, 596–608 (2020)
43. Song, K., Xie, J., Zhang, S., Luo, Z.: Multi-mode online knowledge distillation for self-supervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11848–11857 (2023)
44. Sun, B., Saenko, K.: Deep CORAL: correlation alignment for deep domain adaptation. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9915, pp. 443–450. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_35
45. Tarvainen, A., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. Adv. Neural Inf. Process. Syst. **30** (2017)
46. Toker, A., Zhou, Q., Maximov, M., Leal-Taixé, L.: Coming down to earth: satellite-to-street view synthesis for geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6488–6497 (2021)
47. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7167–7176 (2017)
48. 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)
49. Wang, L., Yoon, K.J.: Knowledge distillation and student-teacher learning for visual intelligence: a review and new outlooks. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6) (2021)
50. Wang, M., Deng, W.: Deep visual domain adaptation: a survey. Neurocomputing **312**, 135–153 (2018)
51. Wang, X., Xu, R., Cui, Z., Wan, Z., Zhang, Y.: Fine-grained cross-view geo-localization using a correlation-aware homography estimator. arXiv preprint arXiv:2308.16906 (2023)
52. 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)
53. Wilson, B., et al.: Argoverse 2: next generation datasets for self-driving perception and forecasting. arXiv preprint arXiv:2301.00493 (2023)
54. Workman, S., Souvenir, R., Jacobs, N.: Wide-area image geolocalization with aerial reference imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3961–3969 (2015)
55. Xia, Y., Yun, L.J., Yang, C.: Transferable adversarial masked self-distillation for unsupervised domain adaptation. Complex Intell. Syst. **9**(6), 6567–6580 (2023)
56. Xia, Z., Booij, O., Kooij, J.F.P.: Convolutional cross-view pose estimation. IEEE Trans. Pattern Anal. Mach. Intell. **46**(5)(2023)
57. Xia, Z., Booij, O., Manfredi, M., Kooij, J.F.P.: Geographically local representation learning with a spatial prior for visual localization. In: Bartoli, A., Fusiello, A. (eds.) ECCV 2020. LNCS, vol. 12536, pp. 557–573. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-66096-3_38

58. Xia, Z., Booij, O., Manfredi, M., Kooij, J.F.P.: Cross-view matching for vehicle localization by learning geographically local representations. IEEE Rob. Autom. Lett. **6**(3), 5921–5928 (2021). https://doi.org/10.1109/LRA.2021.3088076
59. Xia, Z., Booij, O., Manfredi, M., Kooij, J.F.P.: Visual cross-view metric localization with dense uncertainty estimates. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, pp. 90–106. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19842-7_6
60. Xie, S., Zheng, Z., Chen, L., Chen, C.: Learning semantic representations for unsupervised domain adaptation. In: International Conference on Machine Learning, pp. 5423–5432. PMLR (2018)
61. Yang, H., Lu, X., Zhu, Y.: Cross-view geo-localization with layer-to-layer transformer. Adv. Neural. Inf. Process. Syst. **34**, 29009–29020 (2021)
62. Zhang, L., Bao, C., Ma, K.: Self-distillation: towards efficient and compact neural networks. IEEE Trans. Pattern Anal. Mach. Intell. **44**(8), 4388–4403 (2021)
63. Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., Ma, K.: Be your own teacher: improve the performance of convolutional neural networks via self distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3713–3722 (2019)
64. Zhang, W., Ouyang, W., Li, W., Xu, D.: Collaborative and adversarial network for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3801–3809 (2018)
65. Zhang, Y., Tang, H., Jia, K., Tan, M.: Domain-symmetric networks for adversarial domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5031–5040 (2019)
66. Zhang, Y.: A survey of unsupervised domain adaptation for visual recognition. arXiv preprint arXiv:2112.06745 (2021)
67. Zheng, Z., Yang, Y.: Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. Int. J. Comput. Vis. **129**(4), 1106–1120 (2021)
68. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain adaptive ensemble learning. IEEE Trans. Image Process. **30**, 8008–8018 (2021)
69. Zhou, Q., et al.: Uncertainty-aware consistency regularization for cross-domain semantic segmentation. Comput. Vis. Image Underst. **221**, 103448 (2022)
70. Zhu, S., Shah, M., Chen, C.: TransGeo: transformer is all you need for cross-view image geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1162–1171 (2022)
71. Zhu, S., Yang, T., Chen, C.: VIGOR: cross-view image geo-localization beyond one-to-one retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3640–3649 (2021)