CSP-HD: Confidence-Sensitive Progressive Hierarchical Distillation in Cross-View Geo-Localization

Min Xu[1], Zhihong Xu[1], Chaoyu Zhu[1], Xiaochen Li[1], Jing Ye[1], Chen Yang[1*], Guolong Xu[2] and Yan Tian[2*]

[1]China Tower Corporation Limited Zhejiang Branch, Hangzhou, China.

[2]School of Computer Science and Technology, Zhejiang Gongshang University, Hangzhou, China.

*Corresponding author(s). E-mail(s): yangchen@chinatowercom.cn; tianyan@zjgsu.edu.cn;

Abstract

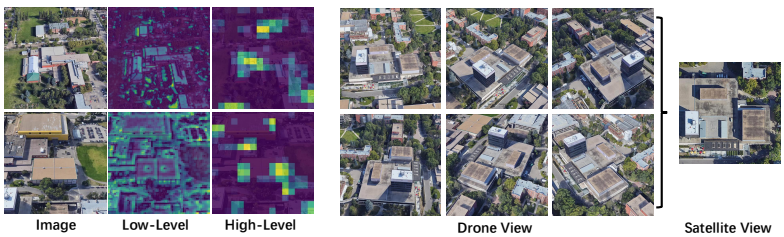
Cross-view geo-localization (CVGL) facilitates drone positioning by correlating overhead visuals with GPS-labeled satellite imagery. However, as data progresses through network layers, the extracted representations become more semantically meaningful but simultaneously discard fine-grained spatial information. Moreover, the vast discrepancy in data volume between drone images and satellite images leads o uneven convergence in the training phrase. Motivated by the knowledge distillation technique, we propose Confidence-sensitive Progressive Hierarchical Distillation (CSP-HD) to improve the effectiveness and efficiency of CVGL. Hierarchical Consistency Distillation (HCD) harnesses the power of inverse self-distillation to simultaneously extract semantic features and spatial details. Progressive Adaptive Loss Weighting (PALW) dynamically regulates the influence of HCD to alleviate short-term fluctuation in the training stage. Confidence-Sensitive Data Alignment (CSDA) dynamically adjusts the learning procedure according to the degree of confidence to alleviate the data imbalance. Experimental results on public datasets demonstrate that our approach improves the effectiveness by a margin of 1.50-1.84% in terms of Recall@1 and 1.50-1.65% in terms of average precision (AP) compared to the prevailing approaches. Project page: https://1781950192.github.io/CSP-HD/.

Keywords: Hard Sample Mining, Data Balance, Cross-View Geo-Localization, Computer Vision

1 Introduction

Cross-view geo-localization (CVGL), correlating and pinpointing image pairs snapped from varying heights and perspectives, such as drone shots from low altitudes versus satellite imagery captured from high above, has been gaining considerable traction due to its real-world applications, including augmented reality [1], autonomous systems [2], and urban navigation [3].

However, CVGL encounters fundamental difficulties that impede implementation in actual environments. Traditional approaches [4, 5, 6, 7, 8, 9, 10] often fail to strike a delicate balance between high-level semantic understanding and maintaining precise spatial accuracy. As data progresses through network layers, the extracted representations become more semantically meaningful but simultaneously discard fine-grained spatial information, such as roof textures and landmark patterns, which is essential for telling apart locations that look similar on the surface. Examples are illustrated in Figure 1(a). Low-level features contain building details such as roof and trees, while high-level features include crucial objectness cues. Moreover, the vast discrepancy in data volume and field variations across various perspectives contributes to uneven convergence, ultimately yielding less than ideal feature alignment. Examples are illustrated in Figure 1(b). Multiple images captured in different views by a drone correspond to a specific image captured by a satellite.



(a) Hierarchical Discrepancy (b) Data Imbalance

Fig. 1 Illustrations of problems in cross-view geo-localization. (a) trade-off between spatial fidelity and semantic abstraction; (b) data imbalance between the different views.

Knowledge distillation [11] is a model compression technique where a small, efficient model (the student) is trained to mimic the behavior of a larger, more accurate model (the teacher), by learning from the teacher's softened output probabilities, which contain valuable "dark knowledge" about the relationships between classes. Inspired by this model compression technique, we can

extract semantic embeddings and spatial knowledge simultaneously. Furthermore, sample confidence can also be obtained as a byproduct in knowledge distillation to select and explore hard samples to improve suboptimal results caused by data imbalance.

We propose Confidence-sensitive Progressive Hierarchical Distillation (CSP-HD) to improve the effectiveness and efficiency of CVGL. Given query image captured by a drone, Hierarchical Consistency Distillation (HCD) is proposed to drive the model's deeper semantic layer to preserve vital spatial insights from earlier processing stages. In addition, a Progressive Adaptive Loss Weighting (PALW) mechanism is proposed to alleviate the short-term fluctuation of HCD in the training phase. Moreover, to balance the effects of images captured by drones and satellites, Confidence-Sensitive Data Alignment (CSDA) intelligently weights learning between drone and satellite channels by exploring uncertainty, which helps cultivate a more reliable and effective training methodology.

The experimental results on the University-1652 [12] and SUES-200 [13] datasets indicate that the proposed approach is competitive with state-of-the-art (SOTA) methods in extensive experiments.

The contributions of this paper are as follows.

- HCD cleverly harnesses the power of inverse self-distillation to craft a streamlined feature extractor that truly hits the mark when it comes to nailing down both semantic and spatial details.
- PALW dynamically regulates the influence of inverse self-distillation, amplifying the penalty of hierarchical inconsistency during the unstable early phase and gradually alleviating short-term fluctuation as training stabilizes.
- CSDA dynamically adjusts the learning procedure according to the degree of confidence the model has about each perspective to alleviate the data imbalance in cross-view geo-localization.

The subsequent sections of this paper are organized as follows. Section 2 reviews studies on cross-view geo-localization, hard sample mining, and data balance. Section 3 presents the proposed approach. Section 4 introduces the experimental results. and Section 5 offers concluding remarks.

2 Related Work

In this section, we present a concise review of the literature on cross-view geo-localization, hard sample mining, and data balance, describing the strengths and drawbacks of each type of approach.

2.1 Cross-View Geo-localization

CVGL identifies a query image's geographical position by comparing it with location-tagged reference imagery. Substantial challenges arise from the drastic visual and geometric disparities between perspectives, which confuse the system locating position accurately [14, 15]. SITE [4] and SRA [5] ensure

that the spatial resolution of the aerial image is adjusted online to match the satellite image, while SMGeo [6] employs a grid-based sparse Mixture-of-Experts (MoE) system that dynamically selects specialized models based on unique characteristics such as content type, scale, and origin—of individual grid elements. This adaptive approach ensures optimal expert activation tailored to specific data requirements. To reduce the negative effects caused by shifting and scale, SDPL [16] uses a dense partition strategy and a shifting-fusion strategy to improve part-based representation learning. MEAN [7] and BEMN [8] explore a progressive cross-domain alignment, global-to-local associations, and a multi-level enhancement strategy to learn robust cross-view consistent features.

Data augmentation remains a largely untapped frontier in cross-view geo-localization, largely due to the sensitivity of the spatial alignment between aerial and terrestrial images—a fragile relationship that can be compromised with even slight disruptions. Some approaches tackle this problem by randomly rotating one perspective while keeping the other stationary. To handle the cost of collecting and annotating cross-view image pairs, UUD [17] uses cross-view projection to produce initial pseudo-labels and then utilizes mutual-matching to refine the pseudo-labels. Alternatively, the panorama-BEV Co-Retrieval Network [18] introduces Bird's Eye View (BEV) and satellite image retrieval branches for collaborative retrieval. Further, Video2BEV [9] transforms the video into a BEV employing 3D Gaussian Splatting to reconstruct the 3D scene based on the multi-view snapshots. P2FCN [10] mitigates stylistic discrepancies between cross-environment images through pixel-level dynamic adjustment.

Higher precision demands have increased processing and resource requirements. MEAN [7] and BEMN [8] use dimension reduction and restoration to decrease computational complexity. To reduce embedding redundancy, DWDR [19] regresses the embedding correlation coefficient matrix to a sparse matrix with dynamic weights to mine more diverse patterns. SMGeo [6] directly predicts object locations by employing an anchor-free detection head for coordinate regression.

2.2 Hard Sample Mining

Various cross-view geo-localization approaches utilize hard data mining to enhance results. Nonetheless, in-batch mining techniques are constrained by the insufficient variety of samples within a single training batch, whereas global mining approaches necessitate extra memory and computational power to sustain the mining pool. To learn discriminant latent representations, GeoDTR+ [20] integrates geometric layout extractor and contrastive hard samples generation for hard sample mining. Game4Loc [21] uses weighted contrastive learning to learn partial matching of drone-satellite image pairs. Recently, Video2BEV[9] generates hard negative samples using a diffusion model.

However, the effect of hard samples is fixed in the training stage. Erratic gradients from inconsistent embeddings trigger short-term fluctuation, leading to suboptimal cross-view alignment.

2.3 Data Balance

Cross-view geo-localization remains challenging since the number of images from different platforms is imbalanced. The vast discrepancy in data volume and field variations across various perspectives contributes to uneven convergence [22], ultimately yielding less than ideal feature alignment. DWDR [19] provide a cross-view symmetric sampling strategy to align the number of the same geo-tag images between different platforms. However, the sampling strategy decreases the number and diversity of samples that joined the learning process, resulting in suboptimal feature alignment due to the weak discrimination ability in feature representations.

3 Our Approach

An approach named CSP-HD is proposed, receiving both advantage in effectiveness and efficiency. The details of our framework are also illustrated in Figure 2. In the training stage, given a query image captured by a drone and a tagged image captured by a satellite, HCD drives the model's deeper semantic layer to preserve vital spatial insights from earlier processing stages. PALW alleviates the short-term fluctuation of HCD in the training phase. CSDA intelligently weights learning between drone and satellite channels by exploring uncertainty.

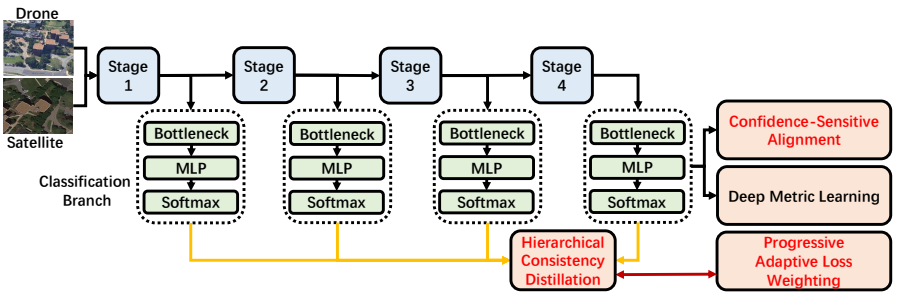


Fig. 2 Illustration of the proposed CSP-HD. In the training stage, given a query image captured by a drone and a tagged image captured by a satellite, Hierarchical Consistency Distillation drives the model's deeper semantic layer to preserve vital spatial insights from earlier processing stages. Progressive Adaptive Loss Weighting alleviates the short-term fluctuation of HCD in the training phase. Confidence-Sensitive Data Alignment intelligently weights learning between drone and satellite channels by exploring uncertainty.