Dynamic Contrastive Learning for Hierarchical Retrieval: A Case Study of Distance-Aware Cross-View Geo-Localization

Suofei Zhang, Xinxin Wang, Xiaofu Wu, Quan Zhou, and Haifeng Hu

***Abstract*—Existing deep learning-based cross-view geo-localization methods primarily focus on improving the accuracy of cross-domain image matching. Less attention is paid to ensuring that models can comprehensively capture contextual information around the target and minimize the cost of localization errors. To support quantitative research into this Distance-Aware Cross-View Geo-Localization (DACVGL) problem, we construct Distance-Aware Campus (DA-Campus), the first benchmark that pairs multi-view imagery with precise distance annotations across three spatial resolutions. Based on DA-Campus, we formulate DACVGL as a hierarchical retrieval problem across different domains. In this setting, we further identify that due to the inherent complexity of spatial relationships among buildings, conventional metric learning lacks a unified semantic hierarchy to guide the organization of the latent feature space. To tackle this challenge, we propose Dynamic Contrastive Learning (DyCL), a novel framework that progressively aligns feature representations according to spatial margins. Extensive experiments demonstrate that DyCL can serve as a strong baseline for the DACVGL task, yielding substantial improvements in both hierarchical retrieval performance and overall geo-localization accuracy. Our code and benchmark are publicly available at https://github.com/anocodetest1/DyCL.**

***Index Terms*—Cross-view geo-localization, Hierarchical retrieval, Contrastive learning, Re-ranking.**

I. Introduction

CROSS-VIEW Geo-Localization (CVGL) has emerged as a fundamental and challenging task in the fields of remote sensing and computer vision. Given a query image captured from one viewpoint (such as the ground), it aims to identify the corresponding images of the same geographic location in a reference database, where candidate images are acquired from a different viewpoint (such as satellite) [1]. CVGL can serve as a critical complement to Global Navigation Satellite Systems (GNSS) especially when the signals are unavailable. Thus it is pivotal for a wide range of location-based applications, including autonomous driving, drone navigation, and disaster rescue. From a theoretical perspective, CVGL can be formulated as a cross-domain image retrieval task [2], [3]. Within the deep learning literature, metric learning [4], [5] has emerged as a widely adopted approach to address this challenge, enabling neural networks to learn view-invariant feature representations. These models construct an embedding space where visual features from disparate domains are aligned, allowing for quantitative similarity comparisons.

Based on feature embeddings, existing CVGL methods [6]–[8] are designed to optimize the final ranking so that images precisely matching the query appear at the top of the retrieval list. In this paper, we revisit this paradigm and argue that there exists a fundamental limitation of this strategy. It treats geo-localization as a purely semantic label matching task, rather than prompting the model to capture the geographic relationship between the query and candidate images. Such relationships, however, are essential in realistic scenarios.

For example, as shown in Fig. 1, given a query image of a specific building, *exact-match* localization successfully places the true matches at the top of the retrieval list. However, when no other exact matches exist, it tends to randomly assign high ranks to distant and unrelated buildings, simply due to the absence of better alternatives. By taking spatial relationships into account, a more rational localization should prioritize not only the exact matches but also images that are geographically close to the query. Such ranking strategy offers practical advantages: even when the model makes mistake, the top-ranked images are more likely to contain potential clues or contextual information about the target location, thus resulting in more pertinent geo-localization. While recent works have implicitly considered spatial constraints [7] or neighborhood consistency [9] in CVGL, there still exists neither a unified theoretical framework to explicitly model geographic relationships in retrieval rankings, nor a dedicated benchmark to quantitatively assess how well algorithms capture spatial proximity. Established benchmarks, such as CVUSA [10], CVACT [11], and University-1652 [12], [13], predominantly employ a standard retrieval evaluation protocol.

To address this gap, in this paper we introduce the concept of *Distance-Aware Cross-View Geo-Localization* (DACVGL). Differing from conventional CVGL task, DACVGL reformulates the problem from single-scale exact matching into a hierarchical retrieval framework, where image relevance is continuously defined by geographic distance rather than discrete semantic labels. The resulting rankings are also quantitatively evaluated using distance-aware metrics [14]–[16]. This formulation requires the model to explicitly incorporate spatial proximity into the retrieval process, generating rankings that are both geographically meaningful and semantically

This work was supported in part by the National Natural Science Foundation of China under Grants 62371245 and 62476139, in part by the Enterprise Innovation and Development Joint Fund of the National Natural Science Foundation of China under Grant U24B20187 and in part by the Natural Science Foundation of Jiangsu Province under Grant BK2024023.

Suofei Zhang is with the School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: zhangsuofei@njupt.edu.cn).

Xinxin Wang, Xiaofu Wu, Quan Zhou and Haifeng Hu are with the National Engineering Research Center of Communications and Networking, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mails: 1022010111@njupt.edu.cn; xfuwu@ieee.org; quan.zhou@njupt.edu.cn; huhf@njupt.edu.cn).

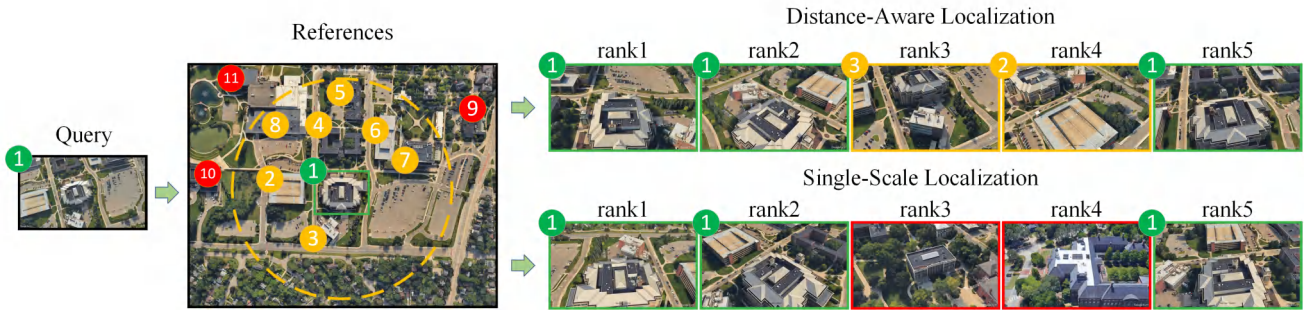


Fig. 1. Illustration of cross-view geo-localization retrieval process. Given a query image of a specific building marked in green, the system ranks candidates from a reference set that includes both the target and surrounding buildings. Two types of retrieval strategies are compared. The *exact-match localization* refers to the conventional retrieval paradigm, which focuses on searching precisely matched samples in reference images. To the contrary, the proposed *distance-aware localization* explicitly incorporates spatial proximity into the ranking process. In such ranking, despite that rank 3 and rank 4 candidates (marked in yellow) are mismatched samples, it is worth noting that they contain partial views or contextual cues of the true target in the background, thus providing potential information for geo-localization.

consistent. We believe this capability is crucial, as empowering the model to comprehensively capture contextual information rather than merely performing instance matching aligns better with the intrinsic demands of real-world CVGL applications.

DACVGL is reminiscent of the "dynamic range" concept in the context of HR [14], [15], where the identification of instances dynamically depends on the specific scale. However, it is important to note that there exists an essential difference between two tasks. In DACVGL, the building identity at the smallest scale is consistent with standard HR. By contrast, the spatial relationships between buildings can only be measured by geographic distances from a specific building taken as an anchor. As a result, higher-level annotation can only be provided in the form of an anchor-specific ground-truth ranking derived from these distances, rather than a unified semantic hierarchy as in standard HR. We discuss this difference in spatial annotation in more detail in Section III-B. Such anchor-dependent variability in supervision motivates us to formulate DACVGL within a contrastive learning framework. Hence, we propose the *Dynamic Contrastive Learning* (DyCL) approach, which explicitly integrates spatial distances into the design of loss functions to promote the learning of hierarchical retrieval.

In addition, to facilitate systematic evaluation of how different localization methods understand spatial relationships, we construct the Distance-Aware Campus (DA-Campus) dataset. Analogous to University-1652 [12], [13], DA-Campus features a carefully designed spatial distribution of buildings, with multi-view images collected from both drone and satellite perspectives. Furthermore, every image in DA-Campus is geo-referenced with precise GPS tags, enabling spatial relationships between buildings to be quantified by distance and thus supporting hierarchical evaluation at multiple scales.

To sum up, the main contributions of this paper are summarized as follows:

- We introduce the DACVGL task, which is formulated as a hierarchical retrieval problem based on geographic distance. To support the setting, we construct the DA-Campus benchmark with quantitative distance annotations. DA-Campus features densely sampled buildings organized as spatial clusters, better reflecting real-world CVGL scenarios and providing a more challenging benchmark for evaluating distance-aware retrieval.
- We propose the DyCL framework with scale-dependent margin control to mitigate optimization conflicts across hierarchical spatial ranges and better preserve distance-aware structure in the embedding space.
- We develop a multi-scale re-ranking algorithm, which adaptively refines neighborhood sizes across hierarchical levels to further improve retrieval performance across different scales of spatial granularity.
- Extensive experiments on multiple datasets demonstrate that DyCL can serve as a strong baseline for DACVGL. Meanwhile, the performance of existing methods on DA-Campus indicates that DACVGL remains a challenging problem, leaving ample room for further investigation.

II. Related Work

A. Cross-View Geo-Localization

Recent progress in CVGL has largely been driven by deep learning models with increasingly sophisticated network architectures and optimization strategies. The introduction of NetVLAD [2] into siamese frameworks [9] enabled the extraction of descriptors robust to large viewpoint variations. Subsequent research on network architectures introduced several key enhancements, including orientation encoding [11], spatial layout modeling [17], domain alignment and spatial attention [18], as well as dynamic similarity matching modules [19], all of which have contributed to improving cross-view feature representation. From the perspective of model training, specialized loss functions such as ranking losses [20] and instance loss [12] are exploited to achieve competitive retrieval performance. Also, Sample4Geo [7] explored hard negative sampling strategies, demonstrating that effective negative mining significantly improves model discrimination in cross-view matching. Building on these strong baselines, recent studies prioritized enhancing feature granularity and model efficiency. Specifically, strategies involving domain alignment and position-aware partitioning [21], [22] are proposed to capture fine-grained local details. In parallel, lightweight ar-

chitectures [23] are developed to balance retrieval accuracy with computational costs via multi-level embedding.

More recently, research in CVGL has focused on the design and adaptation of backbone architectures, either by developing specialized backbone networks tailored for this task or by leveraging large-scale vision foundation models for higher performance. Zhu et al. [8] proposed a streamlined and generalizable backbone architecture, achieving strong performance across various geo-localization tasks. The CV-Cities dataset [6] introduced a large-scale benchmark covering major world cities, supporting research on CVGL in complex urban environments. Leveraging this dataset, they demonstrated that large-scale vision foundation models can achieve strong performance for CVGL. Despite these advances, current methods still paid little attention to the error cost or spatial implications of incorrect localizations.

B. Metric Learning

Metric learning underpins CVGL by constructing an embedding space where images from different domains can be compared meaningfully, thereby enhancing generalization beyond the training set [24]. To achieve this, a variety of loss functions, such as CosFace [4], ArcFace [5], Triplet Loss [24], [25], Contrastive Loss [26], N-pair Loss [27] and Instance Loss [28], [29], have been developed to encourage a geometrically well-structured feature space. These approaches have been widely adopted in various tasks such as face recognition, person re-identification, and vehicle re-identification, contributing to substantial improvements in retrieval accuracy.

Recently, DyML has emerged as a promising direction of retrieval tasks, aiming to learn a unified metric space that adapts to multiple semantic scales. Differing from hierarchical classification [30], [31], DyML is formulated in an open-set regime, where the set of classes during test is disjoint from those during training. Sun et al. [15] formally introduced the DyML problem and established three multi-scale retrieval datasets as standard evaluation benchmarks. They also proposed the Cross-Scale Learning (CSL) framework as a strong baseline in this task. The HAPPIER approach [14], which currently represents the state of the art in DyML, directly optimizes hierarchical average precision by employing differentiable surrogate functions. Note that although these methods excel at multi-scale retrieval, they are not inherently tailored for geo-localization. To bridge this gap, we adapt dynamic metric learning to the DACVGL setting, proposing a novel contrastive learning framework to enhance HR performance across different scales.

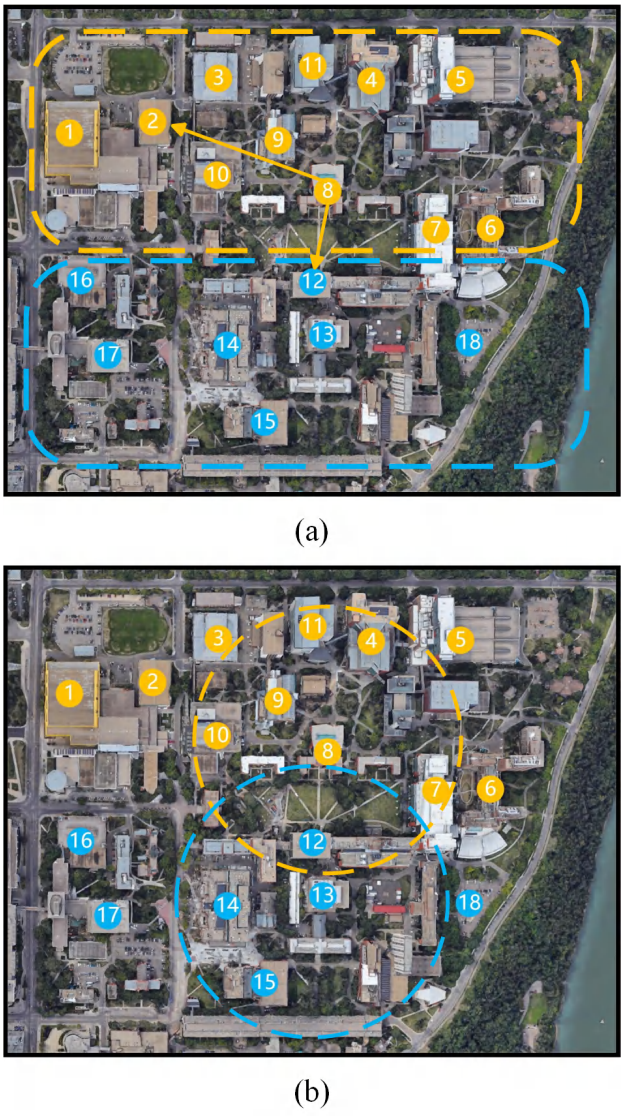


Fig. 2. Spatial annotation schemes. (a) Unified tree structure as in DyML. The area is divided into non-overlapping subregions according to geographic location. Buildings within each region are assigned multi-scale labels according to membership. A hierarchical label tree is constructed to represent correspondences across scales. (b) Distance-based spatial annotation scheme. Each building serves as an anchor, and surrounding buildings are ranked by geographic distance to form the ground truth for retrieval. Different distance thresholds (circles) are adopted to define multiple relevance levels for hierarchical retrieval evaluation.

III. Distance-Aware Campus Dataset

A. Data Collection

Existing benchmarks for CVGL are limited by their reliance on semantic location labels, which fail to capture hierarchical relationships or spatial continuity between locations. From the perspective of DACVGL, it is also desirable for image content to provide rich contextual information beyond the primary target. Compared to ground-level imagery, drone imagery is better suited to this requirement, as it naturally encompasses broader backgrounds and more environmental details. Motivated by these considerations, we establish the DA-Campus dataset, which features multi-source, multi-view imagery with explicit hierarchical and distance-aware labels.

DA-Campus is a large-scale dataset consisting of satellite-view images collected from Google Maps at its highest spatial resolution (approximately 0.3 m per pixel in urban regions) and synthetic drone-view images generated from Google Earth 3D models. We mainly follow the data construction protocol of University-1652 [12], [13], where each building corresponds to one satellite-view image and multiple drone-view images. In total, 750 buildings from universities around the world are selected. A total of 27.4k images are sampled from 450 buildings

as the training set, while 18.3k images are sampled from 300 buildings as the test set. Specifically, we collected metadata for these buildings from Wikipedia, including their names and respective university affiliations. The building names were then geocoded into precise geographic coordinates (latitude and longitude), which were subsequently used to extract corresponding satellite-view images from Google Maps.

To acquire drone-view imagery without incurring the cost of real-world drone flights, we leveraged 3D models from Google Earth to simulate drone footage. Consistent with University-1652 [12], a virtual drone was controlled to follow a 360-degree circular path in a Point-Of-Interest (POI) orbit mode around each building. Specifically, the drone maintained a relative flight altitude of approximately 170 m. The pitch angle of camera was fixed at an oblique view of $50°$, with the heading continuously adjusted to ensure the view always aligned with the target building. Finally, all images captured along this trajectory are assigned the geographic coordinate of the target building, which serves as the anchor for subsequent calculation of spatial relationships. This configuration not only eliminates orientation ambiguity but also naturally encompasses neighboring buildings within the field of view due to the oblique perspective. Furthermore, we adopted a higher image resolution (1920×1080 pixels) than that of University-1652. While the flight geometry ensures that spatial context is included, this high resolution renders background structures with sufficient fidelity as contextual cues.

B. Spatial Annotation Scheme

To precisely annotate the relative spatial relationships among the collected images, we considered two alternative schemes, as illustrated in Fig. 2. The first scheme, shown in Fig. 2(a), follows the standard approach commonly used in DyML tasks, where buildings are grouped into discrete regions. A hierarchical label tree is constructed to represent the correspondence between different scales. Despite this method facilitates clear semantic organization, there is a notable problem arising with buildings located near region boundaries. These buildings often share related visual features with those in adjacent regions, yet are labeled as negative samples solely due to region assignments. For instance, Building 8 and Building 12 in Fig. 2 are geographically close and share many visual features from their surrounding environment. However, according to semantic correspondence, they are assigned to different regions. In contrast, more distant buildings, such as Building 2 and Building 8, are grouped into the same region.

Due to this issue of geographic spatial continuity, we ultimately adopted the distance-based spatial annotation scheme illustrated in Fig. 2(b). Although this scheme requires calculating the spatial relevance between every pair of buildings, it is better suited for capturing true geographic relationships and supporting the study of DACVGL. Specifically, we manually annotated each building with GPS coordinates and computed the distance between all building pairs. For each building, a ranking list of all other buildings was constructed according to geographic distance. The ranking list serves as the ground truth for evaluating retrieval performance in the subsequent tasks.

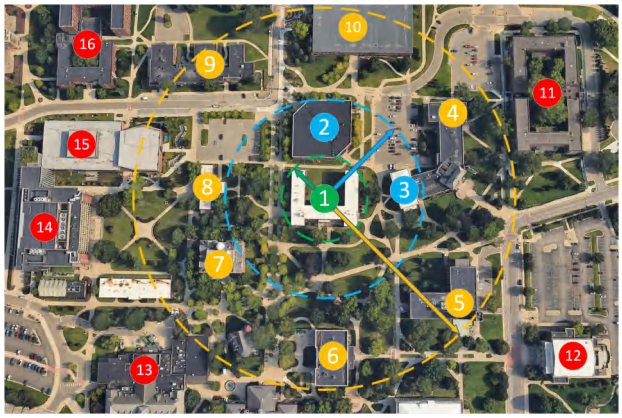


Fig. 3. Illustration of the spatial partitioning strategy in DACVGL. Buildings surrounding the anchor (Building 1) are partitioned into different geographic scales according to distance thresholds, as indicated by concentric colored circles. The green circle represents the anchor building itself (distance = 0), while the blue, yellow, and red circles denote neighboring buildings at increasing distance ranges respectively.

C. Evaluation Protocol

From the perspective of CVGL, DA-Campus supports two standard tasks: drone navigation (Satellite→Drone) and drone-view target localization (Drone→Satellite). Leveraging the available spatial information, both tasks enable evaluation under an HR protocol. Specifically, for each query, buildings in the ranking list are categorized into different relevance levels according to distance thresholds of 0 m, 200 m, and 500 m, corresponding to small, medium, and large retrieval scales. At each scale, conventional metrics such as Recall@K (R@K) and mean Average Precision (mAP) are computed.

Furthermore, we employ three standard metrics commonly used in HR tasks [15], i.e., Hierarchical Average Precision (H-AP) [14], Average Set Intersection (ASI) [32], and Normalized Discounted Cumulative Gain (NDCG) [16], to provide a comprehensive evaluation of hierarchical retrieval quality. H-AP generalizes the standard AP to graded relevance, offering a severity-aware measure of ranking consistency across hierarchical levels. ASI quantifies the overlap between retrieved sets at different scales, capturing the stability of retrieval results across multiple spatial levels. NDCG assesses the ranking of highly relevant items, emphasizing fine-grained precision at the top of the list. Detailed mathematical definitions of these metrics are provided in the Supplementary Material.

In summary, compared with existing CVGL datasets, DA-Campus was constructed by densely sampling buildings across multiple geographic regions with relatively even local coverage, forming coherent spatial clusters. Such a cluster-centric organization produces realistic spatial relationship among buildings, allowing both nearby and distant structures to coexist within each local region. Supported by comprehensive annotations of the spatial relationship, DA-Campus better aligns with real-world CVGL scenarios and offers a more challenging yet informative benchmark for evaluating distance-aware retrieval.

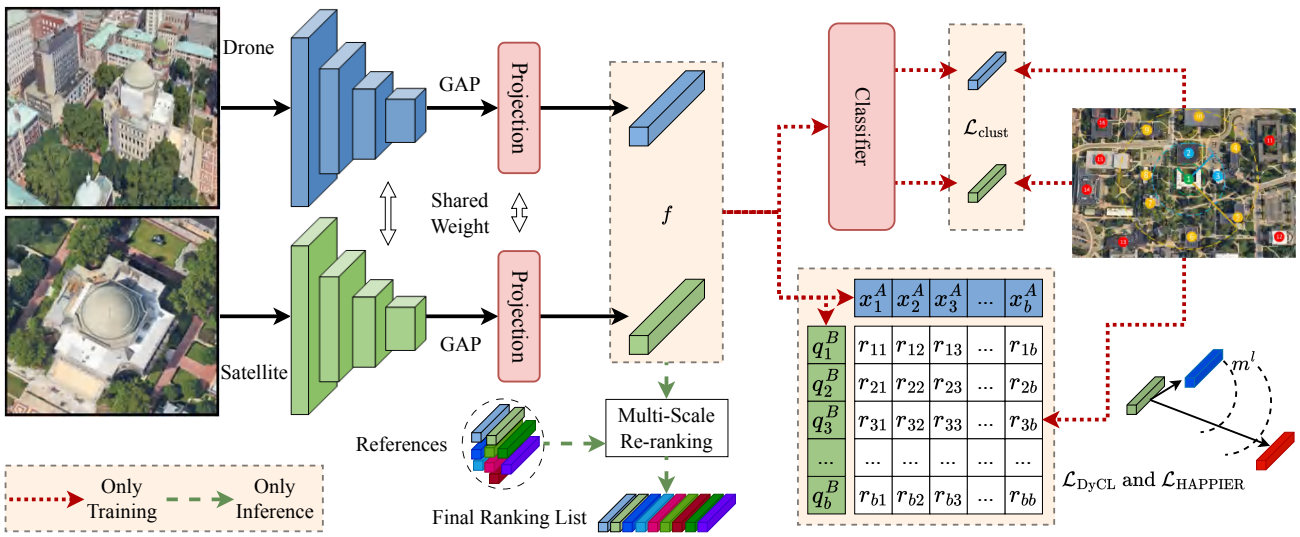


Fig. 4. Overview of the proposed DyCL framework, which integrates hierarchical supervision and Multi-Scale Re-ranking (MSRerank). Images from different views are processed by a shared-weight backbone, followed by Global Average Pooling (GAP) and a shared projection layer to generate 512-dimensional feature embeddings. During training, each batch consists of b image pairs (x_i^A, q_i^B), where each pair is captured from the same building y_i with different views. The embedding output f is optimized under hierarchical supervision via a hybrid objective: a clustering loss $\mathcal{L}_{\text{clust}}$ for discriminative representation, and two multi-scale metric-learning losses, $\mathcal{L}_{\text{DyCL}}$ and $\mathcal{L}_{\text{HAPPIER}}$, to enhance hierarchical retrieval performance. During inference, cross-view similarities are computed from the learned embeddings to produce retrieval ranking. Optionally, the proposed MSRerank algorithm can also be applied as a post-processing module to further improve the multi-scale consistency of the result.

IV. Distance-Aware Cross-View Geo-Localization

A. Task Formulation

We formulate the DACVGL task in a manner analogous to HR as follows. Let us assume a retrieval set $\Omega^A = \{I_1^A, I_2^A, \ldots, I_C^A\}$ comprising images from C buildings. Each I_c^A denotes the set of images of the c-th building captured from viewpoint A. As illustrated in Fig. 3, for a randomly selected building, e.g., Building 1, the set of its own instances forms the smallest geographic scale, denoted as $\mathcal{S}_c^0$. Its nearest neighbors, such as Buildings 2 and 3, are grouped into the next geographic scale $\mathcal{S}_c^1$. The distances between Building 1 and these buildings are less than the smallest positive threshold. By comparison, buildings that are farther away, such as Buildings 4–10, constitute the larger scale, $\mathcal{S}_c^2$, and so on. By applying a set of increasing distance thresholds, we define a series of nested geographic ranges $\mathbb{S}_c = \{\mathcal{S}_c^l | l \in [0, L]\}$. Each $\mathcal{S}_c^l$ contains a total of N_c^l labeled images of buildings from a subset C_c^l of C, i.e.,

$$\mathcal{S}_c^l = \{(x_j^A, y_j) | j = 1, 2, \ldots, N_c^l, y_j \in C_c^l\}, \tag{1}$$

where x_j^A indicates that images are captured from viewpoint A.

Given a query image q_i^B of the c-th building captured from viewpoint B, the goal of DACVGL is to learn a unified feature space $\mathcal{F}^{A \leftrightarrow B}$ in which cross-view images can be directly compared across all geographic scales. Specifically, the similarity between q_i^B and candidate image x_j^A can be computed in $\mathcal{F}^{A \leftrightarrow B}$ and denoted as $r_{ij} = \text{sim}(q_i^B, x_j^A)$. At each scale $\mathcal{S}_c^l$, we define $\mathcal{S}_c^{\leq l} = \bigcup_{m=0}^{l} \mathcal{S}_c^m$, and its complement $\mathcal{S}_c^{>l} = \mathbb{S}_c \setminus \mathcal{S}_c^{\leq l}$. Formally, the objective function of DACVGL can be formulated as:

$$\max(r_{ij} - r_{ik}), \quad \forall \mathcal{S}_c^l, \ (x_j^A, y_j) \in \mathcal{S}_c^{\leq l}, \ (x_k^A, y_k) \in \mathcal{S}_c^{>l}. \tag{2}$$

This learning objective ensures that, at each geographic scale, images more relevant to the query are assigned higher similarity scores, and vice versa. As a result, the model is encouraged to capture the hierarchical structure of spatial relevance in cross-view retrieval. In the final ranking, images of the same building or nearby buildings are consistently prioritized over those that are more distant, yielding a distance-aware retrieval paradigm. Note that the proposed anchor-specific definition of scales $\mathbb{S}_c$ highlights the essential distinction between DACVGL and standard HR: in DACVGL, visual and spatial correlations are always defined dynamically and continuously with respect to each specific anchor. It is therefore impossible to organize all images into a fixed, mutually exclusive semantic hierarchy as in standard HR.

We adopt a threshold-based partitioning strategy instead of direct distance-based ranking for two main reasons. First, in practical scenarios, it is often uninformative to impose a strict ordering of buildings by precise distance, especially when they are distributed in different directions. Grouping buildings by distance thresholds better reflects local visual context and feature correlations. Second, this threshold-based partitioning aligns with established evaluation protocols in HR, thereby facilitating a more rigorous and interpretable assessment of retrieval performance across multiple scales.

B. Overall Framework

As illustrated in Fig. 4, our proposed DyCL framework extends the standard cross-view retrieval pipeline with ex-

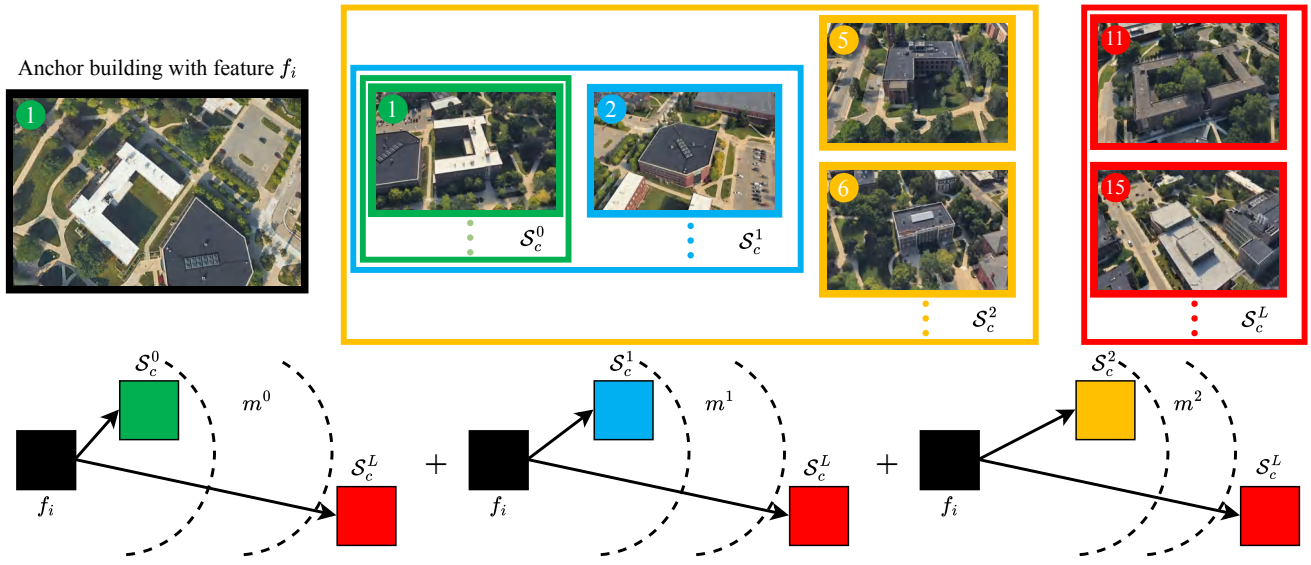


Fig. 5. Illustration of the margin control mechanism in the dynamic contrastive loss $\mathcal{L}_{\text{DyCL}}$. Given an anchor representation f_i, the green, blue, and yellow squares represent reference samples at increasing distances, corresponding to different geographic scales. The red squares denote pure negative samples, i.e., images of buildings whose distances exceed the largest threshold. Each sample is annotated with a consistent building identity as in Fig. 3, enabling a direct comparison of their spatial relationships. In this diagram, a shorter visual distance between squares indicates a higher feature similarity. During learning, $\mathcal{L}_{\text{DyCL}}$ explicitly controls the similarities between the anchor (black square) and the reference samples (colored squares). Notably, the similarity between f_i and $\mathcal{S}_c^L$ serves as a direct reference, while the similarities to other samples decrease gradually according to the margins m^l at different scales.

plicitly designed hierarchical supervision to learn a distance-aware retrieval model. Given input images captured from different viewpoints, a siamese backbone network is exploited to extract feature representation f_i for each image. The backbone can be instantiated with various pre-trained models, with the original classification head removed. After Global Average Pooling (GAP), a fully connected projection layer is appended to map all features into an embedding space with unified dimensionality. Differing from conventional cross-view retrieval frameworks, our approach incorporates hierarchical supervision during training. Specifically, each batch consists of b pairs of images (x_i^A, q_i^B), where each pair is captured from the same building y_i with different views. For $\mathcal{S}_c^0$, the embedding feature f_i is optimized by a proxy-based clustering loss under the supervision of y_i. For larger scales, each sample in the batch serves as an anchor. Contrastive learning is applied to similarities r_{ij} between samples at each scale to comprehensively capture hierarchical spatial relationships. During inference, the learned embeddings are directly compared across views to compute similarity scores between query and reference images, generating a distance-aware ranking that jointly captures semantic correspondence and geographic proximity.

C. Dynamic Contrastive Learning

A typical challenge in HR is the inherent conflict that arises when metric learning is performed across different scales within a unified feature space. As can be seen in Eq. (2), without any extra constraints, optimizing $\max(r_{ij} - r_{ik})$ at scale $\mathcal{S}_c^l$ causes the samples in $\mathcal{S}_c^l$ to be included in r_{ij}. However, at scale $\mathcal{S}_c^{l-1}$, the same samples are included in r_{ik}, resulting in opposing optimization objectives at adjacent scales. Our experiments in Section V-B also empirically verify this observation. To address this issue in the DACVGL scenario, we propose a specialized dynamic contrastive learning loss function, denoted as $\mathcal{L}_{\text{DyCL}}$. This loss independently optimizes the similarity among the anchor, positive, and pure negative samples at each scale. By explicitly controlling the similarity margins for different scales as a series of decreasing values, we mitigate the contradictions in Eq. (2) and enhance the generalization ability of the model across multiple scales.

As illustrated in Fig. 5, for a given query image with output feature vector f_i as the anchor, we take cosine similarity, $r_{ij} = f_i^T f_j$, as an instance of the metric between different samples. At each scale $\mathcal{S}_c^l$, the similarity between the anchor and positive samples, namely $(f_j, y_j) \in \mathcal{S}_c^{\leq l}$, is denoted as $r_{p,j}^l$. The similarity between the anchor and pure negative samples is denoted as $r_{n,k}^L$. Here pure negatives $(f_k, y_k) \in \mathcal{S}_c^L$ correspond to images of buildings whose geographic distances to the anchor exceed the largest threshold. $\mathcal{L}_{\text{DyCL}}$ seeks to enforce a margin between $r_{p,j}^l$ and $r_{n,k}^L$ as:

$$r_{p,j}^l - r_{n,k}^L \geq m^l, \quad l = 0, 1, \ldots, L-1, \tag{3}$$

where m^l denotes the margin associated with each scale. Intuitively, we set $m^0 > m^1 > \cdots > m^{L-1} > 0$ to explicitly reflect the variation in geographic distance across scales. Therefore, the loss can be formally defined as:

$$\mathcal{L}_{\text{DyCL}} = \sum_{l=0}^{L-1} \log(1 + \sum_{j=1}^{|\mathcal{S}_c^{\leq l}|} \sum_{k=1}^{|\mathcal{S}_c^L|} \exp \tau(r_{n,k}^L - r_{p,j}^l + m^l)), \tag{4}$$

where τ is a scaling factor and $|\cdot|$ denotes set cardinality. Finally, the entire loss function follows a symmetric cross-view retrieval paradigm: images from either viewpoint can serve as anchors, with their cross-view counterparts in $\mathbb{S}_c$ serving as reference samples. This symmetric formulation

allows the loss to be applied in both directions. Since each batch is constructed by pairing images from different views, all required triplets for Eq. (4) can be efficiently generated within a single batch.

The design of dynamic contrastive learning takes inspiration from the idea of CSL [15], but with explicit difference. CSL adopts a metric learning approach similar to CosFace [4], in which the classification branch serves as proxies for the class centers during training. The similarity is calculated as the inner product between the output feature and the classifier weights. In contrast, $\mathcal{L}_{\text{DyCL}}$ adopts an explicit contrastive learning approach, directly performing anchor-based similarity comparisons between samples. This design is apparently more suitable for the DACVGL scenario, given the spatial complexity of building distributions. On the other hand, compared with HAPPIER [14], another multi-scale metric learning loss, our design instead optimizes the geometric structure of the feature space within a standard contrastive learning framework, rather than optimizing the ranking metric via surrogate functions. Our comparisons in Section V-G empirically validate that two approaches exhibit strong complementarity. As a result, their combination yields the best performance.

Loss functions. Besides $\mathcal{L}_{\text{DyCL}}$, at the scale $\mathcal{S}_c^0$, we additionally employ a loss to ensure that instances from the same building are closely clustered:

$$\mathcal{L}_{\text{clust}} = -\log\left(\frac{\exp(w_i^T f_i)}{\sum_{j=1}^{C}\exp(w_j^T f_i)}\right), \tag{5}$$

where w_i is the normalized proxy corresponding to the fine-grained class of feature f_i. This normalized variant of cross-entropy loss [33] effectively enforces clustering of instances across the entire training set according to their labels. It compensates for the limitation that in $\mathcal{L}_{\text{DyCL}}$ all sample comparisons are restricted within a single batch. Finally, the overall training objective of the proposed DyCL framework is formulated as a combination of three complementary loss terms:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{DyCL}} + \lambda_2\mathcal{L}_{\text{clust}} + \lambda_3\mathcal{L}_{\text{HAPPIER}}, \tag{6}$$

where λ_1, λ_2, and λ_3 are weighting coefficients that control the relative contributions of each component. This objective integrates the contrastive supervision from $\mathcal{L}_{\text{DyCL}}$, the proxy-based clustering loss $\mathcal{L}_{\text{clust}}$, and the hierarchical ranking optimization term $\mathcal{L}_{\text{HAPPIER}}$ [14], enabling robust learning across multiple spatial scales.

D. Multi-Scale Re-ranking

As depicted in Fig. 4, beyond the optimization losses, we also explore re-ranking algorithms within the HR framework. A Multi-Scale Re-ranking (MSRerank) scheme is proposed as an integral component of the overall solution, further refining retrieval results during inference. This comprehensive pipeline ultimately constitutes a strong baseline for addressing the DACVGL problem. Re-ranking is a prevailing post-processing step in image retrieval. It is designed to refine the initial ranking by leveraging structural relationships within the retrieval set. In the literature, discussions of re-ranking algorithms are normally based on distance metric $d(q_i^B, x_j^A)$ defined on the output feature embedding. For consistency, we adopt the same notation in this section, instead of the inversely related feature similarity in previous discussions. In practice, it is straightforward to convert between the two metrics when integrating the re-ranking module. Given the original distance $d(q_i^B, x_j^A)$, standard re-ranking algorithm fuses it with a smoothed Jaccard distance, resulting in the re-ranked distance $d^\star(q_i^B, x_j^A; k)$ [34], where k is a hyper-parameter controlling the size of the reciprocal neighborhood.

Algorithm 1 Multi-Scale Re-ranking Algorithm.

Input: Distance matrix $D \in \mathbb{R}^{(|\Omega^A|+1)\times(|\Omega^A|+1)}$ between q_i^B and each entry in Ω^A, parameter list $\{k^l\}_{l=0}^{L-1}$.
Output: Final re-ranked distances $d^\star(q_i^B, x_j^A)$ for $x_j^A \in \Omega^A$.
Initialise $d^\star(q_i^B, x_j^A) \leftarrow 0$, Mask $\mathcal{M} \leftarrow \Omega^A$.
for $l = 0$ **to** $L-1$ **do**
  (a) Standard re-ranking:

$$d_l^\star(q_i^B, x_j^A) \leftarrow d^\star(q_i^B, x_j^A; k^l).$$

  (b) Update active entries:

$$\forall\, x_j^A \in \mathcal{M},\ d^\star(q_i^B, x_j^A)\ \mathrel{+}= d_l^\star(q_i^B, x_j^A).$$

  (c) Select top-k^l samples:

$$\mathcal{M}^l \leftarrow \text{Top-}k^l \text{ samples in } \Omega^A \text{ sorted by } d^\star(q_i^B, x_j^A).$$

  (d) Mask update:

$$\mathcal{M} \leftarrow \mathcal{M} \setminus \mathcal{M}^l.$$

end for
return $d^\star(q_i^B, x_j^A)$.

Existing re-ranking approaches typically fix $k = 20$ to match the expected number of positive samples in standard benchmarks. A suitable k should cover the main distribution of positive samples, thereby ensuring that the reciprocal sets used for Jaccard distance computation are more discriminative. However, in multi-scale retrieval, the number of relevant samples increases significantly at larger scales, making a fixed k suboptimal. Our experiments in Section V-D also reveal that the impact of re-ranking is highly sensitive to the choice of k at different scales. To address this, MSRerank repeatedly applies the standard re-ranking module with different k parameters. The k values are selected based on prior knowledge of the building distribution in the training set as:

$$k^l = \max\left(20,\ \frac{\mu}{C}\sum_{c=1}^{C}\left|\mathcal{S}_c^{\leq l}\right|\right), \tag{7}$$

where μ is an empirical hyper-parameter. In our experiments, we simply fix $\mu = 0.1$. Building on these results, we introduce a segmented accumulative algorithm to generate the final distance as described in Algorithm 1.

The proposed MSRerank algorithm iteratively accumulates the results of standard re-ranking at each scale and masks out

the top-ranked samples after each step. This strategy prevents the re-ranking operations at larger scales from interfering with the refined orderings already established at smaller scales. To the best of our knowledge, MSRerank is the first re-ranking method specifically designed for hierarchical retrieval. It can be seamlessly integrated into various cross-domain and standard HR frameworks. Our experiments demonstrate that it consistently yields performance gains across all evaluation settings.

V. Experimental Results

A. Implementation Details

We conducted extensive comparisons on the DA-Campus dataset to benchmark the performance of various methods for DACVGL. For $\mathbb{S}_c$, we categorize reference buildings into four relevance levels based on their geographic distance to the anchor building, i.e., $L = 3$. Two primary tasks are considered: drone navigation (Satellite $\rightarrow$ Drone) and drone-view target localization (Drone $\rightarrow$ Satellite). For our proposed DyCL framework, we adopt ResNet-50 [35], [36] as the backbone and train the feature extraction model. During training, the Adam optimizer is employed to minimize the combination of losses in Eq. (6), with a batch size of 128 and a total of 20 epochs. All images are resized to 256×256 pixels. Basic data augmentation is applied, including horizontal flipping. For satellite-view images, random rotations are also performed. During inference, DACVGL follows the standard retrieval protocol without additional operations. The trained CNN is used to extract 512-dimensional embeddings for all query and reference images. The features are L2-normalized, and cosine similarity is computed to obtain the ranking list. Then resulting rankings are evaluated by the aforementioned metrics, including H-AP, ASI, NDCG, R@1, and mAP at each scale. For experiments that adopt the proposed MSRerank, an additional post-processing step is applied to refine the ranking based on reciprocal neighborhood consistency.

B. Single-scale Learning vs. Multi-scale Learning

As the first part of our experiments, we empirically validate the efficacy of multi-scale learning over single-scale learning on the DA-Campus dataset. Specifically, for the drone navigation task, we first perform deep metric learning independently at each scale $\mathcal{S}_c^l$ and evaluate the resulting models across all scales. These single-scale models are then compared with a multi-scale learning model, which learns a unified embedding space encompassing all scales. For a fair comparison, all models are trained exclusively using the conventional Triplet Loss [24], [25]. The results are illustrated in Fig. 6. Here, we use the terms small, middle, and large scales to refer to S_c^0, S_c^1, and S_c^2, respectively. It can be observed that: 1) Each single-scale metric demonstrates relatively high accuracy within its specific scale. For example, the model trained only on S_c^0 achieves the highest performance in small-scale testing. 2) However, single-scale metric models do not generalize well to other spatial scales. For instance, the S_c^0 model achieves only 32.34% mAP on large-scale testing, which is 30.74% lower than the S_c^2 model. 3) In contrast, the multi-scale metric learning model consistently outperforms the single-scale models, especially in terms of overall accuracy. It surpasses the S_c^0, S_c^1, and S_c^2 models by 12.28%, 4.43%, and 1.43%, respectively, in the overall accuracy evaluation. These results indicate that, even without employing targeted strategies such as DyCL, multi-scale metric learning is essential for improving overall retrieval accuracy.

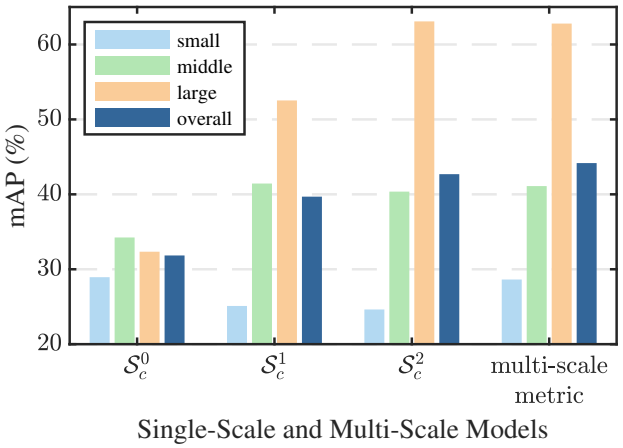


Fig. 6. Comparison of performance between single-scale models and a multi-scale model on the drone navigation (Satellite $\rightarrow$ Drone) task. Each single-scale model is trained individually at scale $\mathcal{S}_c^l$, while the multi-scale model is jointly trained across all scales. For each model, mean Average Precision (mAP) is evaluated at the small, middle, and large scales on the test set. The term *overall* denotes the average of three mAP values as the final comprehensive evaluation metric.

C. Main Results of DACVGL on DA-Campus

We then benchmark various methods on the DA-Campus dataset and present a comprehensive comparison with our proposed DyCL framework in Table I. The candidate methods include both prevailing metric learning approaches and conventional CVGL methods. For single-scale methods [1], [2], [37], models are trained at the small scale using building labels as supervision. Regarding the state-of-the-art CVGL methods [7], [21]–[23], we utilized their official open-source models to conduct validation directly on the DA-Campus dataset. For CV-Cities [6], since the official model weights are unavailable, we followed their released training pipeline and learned a model on DA-Campus using the pre-trained DINOv2 [41] as backbone. All methods are implemented within the standard symmetric cross-view framework. Finally, the overall framework illustrated in Fig. 4, which integrates DyCL, HAPPIER, and clustering loss as defined in Eq. (6), is denoted as DyCL+HAPPIER. We denote the overall framework, which integrates all loss components in Eq. (6), as DyCL in Table I.

It can be observed that the proposed DyCL framework achieves superior results across all evaluation metrics, outperforming other methods by a clear margin. Compared to other leading methods, DyCL demonstrates a fundamental enhancement in maintaining the overall hierarchical ranking quality (e.g., up to 2.73% improvement in H-AP for the Satellite $\rightarrow$ Drone task), even when using a more compact ResNet-50 backbone. These results preliminarily indicate that

TABLE I
PERFORMANCE OF DIFFERENT METHODS ON THREE HIERARCHICAL EVALUATION METRICS. FOR BOTH SINGLE-SCALE AND MULTI-SCALE MODELS, THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN BOLD AND UNDERLINED, RESPECTIVELY. HERE, $\sum$ TRIPLET REFERS TO THE MULTI-SCALE METRIC LEARNING DESCRIBED IN SECTION V-B, WHERE TRIPLET LOSS IS DIRECTLY APPLIED AT ALL SCALES.

	Method	Satellite→Drone			Drone→Satellite		
		H-AP	ASI	NDCG	H-AP	ASI	NDCG
Single-Scale	Contrastive [1], [26]	28.35	34.89	71.69	37.56	35.23	59.68
	NSM [33]	30.88	37.56	72.90	39.23	37.07	60.16
	Soft Margin Triplet [2]	30.72	41.32	73.36	41.82	38.96	62.03
	Triplet [24], [25]	31.14	38.02	73.21	40.82	38.73	61.32
	Instance [29]	38.23	47.01	76.71	45.86	46.18	64.30
	LPN [37]	42.78	48.86	76.83	49.62	48.82	64.08
	InfoNCE [38], [39]	43.37	51.68	**78.75**	50.38	51.54	67.38
	Sample4Geo [7]	38.01	45.16	78.62	46.13	40.51	65.34
	DAC [21]	37.62	42.82	77.54	45.07	37.92	63.07
	CAMP [22]	38.39	44.41	78.45	46.07	39.33	64.70
	MEAN [23]	35.06	38.38	74.97	41.97	33.74	59.20
	CV-Cities [6]	**45.00**	**54.78**	78.25	**53.88**	**55.45**	**70.28**
Multi-Scale	CSL [15]	13.12	19.01	60.24	16.56	14.44	38.23
	Multi-Similarity [40]	36.32	44.52	71.2	46.79	40.91	63.64
	$\sum$ Triplet	44.82	49.43	77.45	50.97	44.98	66.28
	DyCL (ours)	**47.73**	**56.27**	**81.06**	**54.35**	**55.95**	**70.63**
	Sample4Geo+MSRerank	43.55	51.31	80.49	52.07	42.89	68.26
	CV-Cities+MSRerank	49.77	58.90	80.23	58.17	59.98	74.10
	DyCL+MSRerank	**54.63**	**60.11**	**82.72**	**59.14**	**60.82**	**74.11**

utilizing spatial relationships among targets to construct the latent feature space during training yields direct gains in the final hierarchical retrieval metrics. In Table II, we further report the mAP and R@1 results of different methods across all relevance levels. This comprehensive comparison enables a more insightful analysis of underlying mechanisms leading to the ultimate advantages.

At the small scale, recent methods such as Sample4Geo [7], DAC [21], CAMP [22] and MEAN [23] achieve superior mAP and R@1 performance. This success can be largely attributed to their targeted architecture designs and advanced backbones. However, as the spatial scale increases, a significant degradation of performance, especially in mAP, is observed for these methods. Theoretically, this is primarily because, in the absence of explicit spatial constraints, these models rely heavily on distinguishing hard negative samples during training to enhance local discrimination. These hard negatives often correspond to targets that are geographically adjacent to the anchor. Consequently, while achieving high mAP and R@1 scores at the instance level, models tend to push spatially adjacent samples as far as possible in the ranking list. It leads to suboptimal mAPs at larger scales, where adjacent samples are actually considered as positives. In practical applications, this poses a potential risk that once accurate matching fails, the top-ranked results may fail to provide any reliable contextual clues. Comparatively, DyCL mitigates this deficiency through hierarchical loss designs. It consistently outperforms single-scale baselines at large scales and achieves higher overall accuracy, effectively aligning with the objective of pertinent geo-localization.

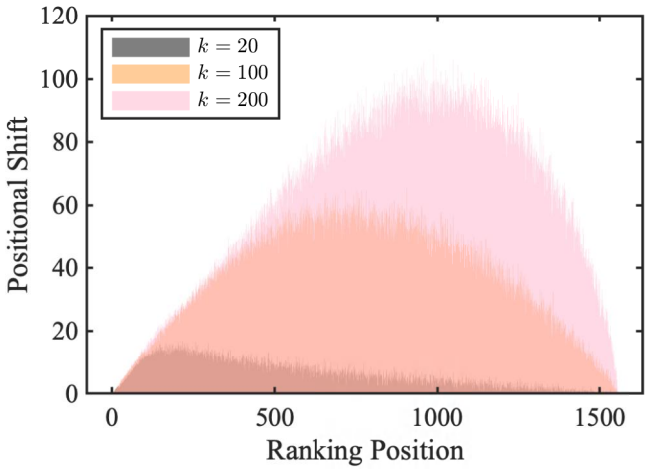


Fig. 7. Distribution of position shifts at different ranking positions after re-ranking with various k values. On the training set, each sample is treated as a query, ranking remaining samples as the retrieval set. Then the curves are obtained by averaging the position changes at each position.

D. Multi-scale Re-ranking

To gain a systematic analysis of the re-ranking algorithm in the context of HR, we started from performing re-ranking with different k values on the training set of DA-Campus. Using the Kendall Tau distance [43], we measured position shifts throughout the ranking list, as illustrated in Fig. 7. The plots intuitively show that varying k explicitly affects different sections of the ranking list. Based on this observation, we empirically selected k^l values for MSRerank as described in Eq. (7) via prior knowledge from the training set.

In Table I and Table II, we listed the performance of MSRerank when applied to different CVGL methods including Sample4Geo, CV-Cities, and our proposed DyCL. It can be observed that although k^l is calibrated based on training statistics, the post-processing strategy generalizes well to the test set. It delivers significant and consistent improvements regardless of the underlying baseline, yielding stable gains in both hierarchical metrics and multi-scale mAP/R@1. Moreover, from the comparisons in Table II, one can see that MSRerank differs from standard re-ranking by preserving R@1 accuracy and boosting mAP performance at larger scales. These comparisons empirically validate that MSRerank can be seamlessly integrated into various CVGL frameworks as a robust post-processing module.

To further verify the efficacy and flexibility of MSRerank, we conducted more experiments on standard HR benchmark datasets, DyML-Vehicle, DyML-Animal and DyML-Product [15], as summarized in Table III. It can be seen that MSRerank consistently improves the mAP over standard re-ranking results. Due to the segmented accumulative strategy, here the performance of R@1 remains unchanged. These comparisons demonstrate that MSRerank can further serve as a generic post-processing technique to reliably enhance various hierarchical retrieval models.

TABLE II
PERFORMANCE OF DIFFERENT METHODS IN TERMS OF MAP AND R@1 ACROSS ALL SPATIAL SCALES. OVERALL ACCURACIES ARE CALCULATED AS THE MEAN VALUES OF THE RESULTS AT THE SMALL, MIDDLE, AND LARGE SCALES.

	Method	**Satellite→Drone**								**Drone→Satellite**							
		Small		Middle		Large		Overall		Small		Middle		Large		Overall	
		mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1
Single-Scale	Contrastive [1], [26]	21.55	55.12	30.54	65.21	32.21	70.89	28.10	63.74	37.89	38.74	38.10	56.87	35.64	64.65	37.21	53.42
	NSM [33]	31.45	56.67	30.70	66.67	28.70	71.33	30.28	64.89	42.38	47.56	38.22	57.68	33.71	64.41	38.10	56.55
	Soft Margin Triplet [2]	26.42	52.21	33.69	66.55	32.45	80.26	30.85	66.34	38.89	44.91	45.74	55.42	41.26	68.37	41.96	56.23
	Triplet [24], [25]	28.94	54.45	34.24	67.89	32.34	82.21	31.84	68.18	39.69	42.56	44.96	55.41	39.92	67.36	41.54	55.11
	Instance [29]	28.18	51.67	36.63	68.00	52.37	81.00	39.06	66.89	38.02	38.26	44.27	54.34	55.92	71.57	46.07	54.72
	LPN [37]	33.06	55.88	40.53	72.56	54.39	82.62	42.66	70.35	44.58	43.20	47.36	58.22	56.24	72.21	49.39	57.88
	InfoNCE [38], [39]	30.91	54.33	41.18	70.00	59.51	84.0	43.87	69.44	40.53	41.23	47.93	56.49	63.28	76.96	50.58	58.23
	Sample4Geo [7]	41.00	67.67	39.75	77.33	37.69	80.00	39.48	75.00	49.98	53.33	41.68	65.50	35.67	**77.09**	42.44	65.31
	DAC [21]	41.88	69.33	36.13	77.07	32.93	83.02	36.98	76.47	**51.99**	**54.48**	39.18	65.33	28.85	72.93	40.01	64.25
	CAMP [22]	**42.23**	**69.67**	38.83	76.33	36.04	80.67	39.03	75.56	51.48	53.38	41.36	66.98	29.56	74.98	40.80	65.11
	MEAN [23]	38.61	66.01	31.34	74.43	23.71	77.02	31.22	72.49	46.79	50.88	34.41	63.54	25.80	65.99	35.67	60.14
	CV-Cities [6]	34.71	67.81	**43.18**	**77.71**	**59.80**	**84.70**	**45.89**	**76.74**	45.53	53.53	**51.65**	**67.65**	**65.19**	75.19	**54.12**	**65.46**
Multi-Scale	CSL [15]	13.66	39.33	12.74	45.00	12.34	49.67	13.04	44.67	17.77	17.77	15.87	21.57	14.36	25.41	16.00	21.58
	Multi-Similarity [40]	25.53	48.32	34.36	70.54	48.33	82.39	36.07	67.08	34.60	35.22	45.31	52.72	60.87	63.28	46.93	50.41
	$\sum$ Triplet	28.63	62.69	41.10	72.73	62.79	84.21	44.17	73.21	38.36	46.00	49.65	62.31	65.63	80.21	51.21	62.84
	DyCL (ours)	**36.08**	**64.67**	**45.77**	**78.33**	**63.78**	**89.00**	**48.54**	**77.33**	**44.87**	**49.21**	**52.05**	**66.64**	**66.89**	**82.03**	**54.45**	**66.29**
	Sample4Geo+MSRerank	**48.89**	67.88	47.76	77.96	45.55	82.08	47.40	75.97	**53.76**	62.35	45.99	73.80	40.23	79.66	46.66	71.94
	CV-Cities+MSRerank	43.05	**67.92**	50.66	77.75	65.01	85.19	52.91	76.95	51.27	**62.80**	55.02	75.07	70.33	78.12	58.87	72.00
	DyCL+Rerank	44.01	64.87	50.53	**78.64**	65.20	**89.92**	53.18	**77.81**	50.93	59.49	55.45	**75.20**	67.55	**84.97**	57.98	**73.22**
	DyCL+MSRerank	43.55	64.87	**52.29**	**78.64**	**70.58**	**89.92**	**55.47**	**77.81**	49.88	59.49	**56.66**	**75.20**	**71.29**	**84.97**	**59.24**	**73.22**

TABLE III
PERFORMANCE COMPARISON BETWEEN STANDARD RE-RANKING AND THE PROPOSED MSRERANK ALGORITHM ON CONVENTIONAL HR DATASETS. THE TRAINED HAPPIER MODEL IS ADOPTED AS THE BASELINE.

Method	**DyML-Vehicle**		**DyML-Animal**		**DyML-Product**	
	mAP	R@1	mAP	R@1	mAP	R@1
Triplet [24], [25]	10.0	13.8	11.0	18.2	9.3	11.2
Multi-Similarity [40]	10.4	17.4	11.6	16.7	10.0	12.7
TL_{SH} [42]	26.1	84.0	37.5	66.3	36.32	69.6
NSM [33]	27.7	88.7	38.8	69.6	35.6	57.4
$\sum TL_{SH}$ [42]	25.5	81.0	38.9	65.9	36.9	58.5
$\sum$ NSM [33]	32.0	89.4	42.6	**70.0**	36.8	60.8
CSL [15]	30.0	87.1	40 .8	60.9	31.1	52.7
HAPPIER [14]	37.0	89.1	43.8	68.9	38.0	63.7
HAPPIER+Rerank [34]	38.55	**90.52**	44.53	69.51	41.21	**64.5**
HAPPIER+MSRerank	**40.7**	**90.52**	**46.44**	69.51	**46.54**	**64.5**

TABLE IV
CROSS-DATASET GENERALIZATION PERFORMANCE EVALUATED ON THE UNIVERSITY-1652. HERE, MAP AND R@1 ARE THE OVERALL ACCURACIES AVERAGED ACROSS MULTIPLE SCALES. FOR SIMPLICITY, WE ONLY LIST THE RESULTS OF THE DRONE NAVIGATION TASK (SATELLITE → DRONE). SIMILAR TRENDS ARE OBSERVED FOR THE SYMMETRIC DRONE → SATELLITE TASK.

Method	H-AP	ASI	NDCG	mAP	R@1
Sample4Geo [7]	39.66	46.35	81.22	74.48	95.66
DAC [21]	40.79	47.15	83.38	76.33	96.48
CAMP [22]	39.96	45.89	83.56	75.18	**96.51**
MEAN [23]	37.21	41.56	81.80	73.37	96.40
DyCL	**41.93**	**50.40**	**83.88**	**81.92**	87.87

E. Cross-Dataset Generalization Analysis

To further assess the generalization capability of the proposed method, we conducted a cross-dataset evaluation on the University-1652 [12]. Leveraging its recently released building-level GPS coordinates, we benchmarked various CVGL models from the perspective of the DACVGL task. As shown in Table IV, we compare our proposed DyCL framework against four state-of-the-art CVGL methods. For the existing standard methods, we evaluate their official models trained on University-1652 with HR metrics. For our proposed method, we directly employ the model trained on DA-Campus and test its generalization capability on the University-1652.

It is worth noting that University-1652 was not originally designed for the DACVGL scenario. However, DyCL still exhibits superior performance in distance-aware metrics. Although high-performance CVGL models achieve almost saturated matching accuracy at the instance level, they suffer from a similar deficiency to that observed on DA-Campus. As the retrieval scope expands, there is a significant performance degradation at larger spatial scales. In contrast, although the domain gap limits the R@1 accuracy of our method, it still achieves higher overall mAP and hierarchical precisions. This advantage primarily stems from that DyCL systematically utilizes coordinate information at both the data and algo-

TABLE V
COMPARISON OF MODEL PARAMETERS AND INFERENCE TIME ON DA-CAMPUS.

Method	Backbone	Params (M)	Inference Time (ms)
Sample4Geo [7]	ConvNeXt-B	88.5	30.97
Sample4Geo+MSRerank	ConvNeXt-B	88.5	46.66
CV-Cities [6]	ViTb14-mix	90.5	46.68
CV-Cities+MSRerank	ViTb14-mix	90.5	62.31
DyCL	ResNet-50	23.6	15.16
DyCL+Rerank	ResNet-50	23.6	20.94
DyCL+MSRerank	ResNet-50	23.6	32.02

rithmic levels. Specifically, the elaborately designed dynamic contrastive loss explicitly exploits the comprehensive distance information provided by the dense and continuous sampling strategy of DA-Campus via hierarchical spatial constraints. Consequently, these results demonstrate that our proposed DyCL can serve as a strong baseline for DACVGL, and the learned distance-aware representations generalize well to unseen environments.

F. Algorithmic Complexity Analysis

We compare the computational costs and inference times of different methods in Table V. All experiments are conducted on a workstation running Ubuntu 20.04 with an NVIDIA Tesla P40 GPU. The reported inference time corresponds to the average per-query latency, obtained by executing the inference pipeline over the entire test set. For the re-ranking algorithms, after completing feature extraction and similarity computation, we also execute the post-processing over the entire test set to measure the runtime. This setting aligns with practical deployment, since both reference features and reciprocal neighborhoods can be cached offline to accelerate the re-ranking process. From the comparison, one can observe that the inference time of DyCL primarily depends on the computational complexity of the backbone network. Our proposed loss function only affects the training stage, without introducing any additional learnable parameters or auxiliary modules. In addition, the extra latency introduced by the standard re-ranking module remains stable across different backbones, typically within 6 ms per query. Our proposed MSRerank module incurs an overhead roughly three times of the standard re-ranking, which is consistent with the description in Algorithm 1. The main complexity of MSRerank arises from running the standard re-ranking algorithm multiple times at different scales.

G. Ablation Studies

In this section, we conduct comprehensive ablation studies on DA-Campus to validate the proposed DyCL framework by analyzing the impact of key hyper-parameters and the contributions of individual loss components.

Hyper-parameters. We analyze the sensitivity of the hyper-parameters in Fig. 8, specifically the weighting coefficients $\lambda_1, \lambda_2, \lambda_3$ in the combined loss defined in Eq. (6), and the scaling factor τ in Eq. (4). First, we investigate the balance between $\mathcal{L}_{\text{clust}}$ and $\mathcal{L}_{\text{HAPPIER}}$ in Fig. 8 (b). With fixed $\lambda_3 = 1.0$, we find that the performance is relatively insensitive to variations of λ_2. This observation is consistent with the results in [14]. Then with fixed λ_2 and λ_3, we search for the optimal λ_1 in Fig. 8 (a), identifying a peak at $\lambda_1 = 0.2$. Subsequently, using the optimal λ_1 and λ_2, we re-evaluate λ_3 over a broader range in Fig. 8 (c). The results indicate that model performance is primarily affected by the ratio between $\mathcal{L}_{\text{DyCL}}$ and $\mathcal{L}_{\text{HAPPIER}}$. An improper balance (e.g., when λ_3 is too low relative to λ_1) may even yield performance degradation. Consequently, we empirically select $\lambda_1 = 0.2$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.9$ as the optimal configuration for all other experiments. Finally, in Fig. 8 (d), we further validate the scaling factor τ. This factor determines the effective magnitude of the hierarchical margins (m^l), thereby controlling the strictness of the spatial constraints. We evaluate τ at values of 16, 32, and 64, and observe that the best overall performance is achieved when $\tau = 32$.

Loss components. To further validate the efficacy of each loss component and investigate their relationships, we compare the performance of different configurations in Table VI. Results demonstrate that $\mathcal{L}_{\text{clust}}$ as a foundation, brings stable improvements to other components. On top of that, a linear combination of $\mathcal{L}_{\text{HAPPIER}}$ and $\mathcal{L}_{\text{DyCL}}$ exhibits strong complementarity as they optimize the feature space from distinct perspectives. The ranking-based $\mathcal{L}_{\text{HAPPIER}}$ excels at the small scale (37.13% mAP) by refining fine-grained retrieval orders. In contrast, while the structure-based $\mathcal{L}_{\text{DyCL}}$ imposes strict constraints that slightly compromise discrimination at the small scale, it ensures superior robustness at larger scales (64.79% mAP) via explicit margin control that maintains cross-scale geometric consistency.

Ultimately, the combination of all three components yields the best overall performance with only a negligible trade-off in instance-level matching.

H. Visualization of Qualitative Results

To intuitively evaluate the practical deployment potential of DyCL, we visualize the top-ranked retrieval results on the DA-Campus dataset. Specifically, we randomly selected three scenes for each task and displayed the top six retrieval results in Fig. 9. As illustrated in the figure, DyCL successfully retrieves the corresponding target image at the top-1 position in some scenarios. Conversely, in certain challenging scenarios, the model fails to produce the correct ranking order. However, note that even in these scenarios, it consistently prioritizes images that are geographically closer to the target (marked with blue frame), effectively mitigating the cost of localization failure. For instance, in the last row of Fig. 9(a), the correct building appears at the edge of the retrieved image. In the second row of Fig. 9(b), the retrieved images depict surrounding buildings with a distinctively similar architectural style to the query. These instances demonstrate that DyCL effectively leverages the correlation between spatial proximity and visual relevance, offering valuable guidance for target localization when exact matching fails.

Failure cases indicate that DA-Campus remains a challenging benchmark. The primary difficulty arises from the

TABLE VI
ABLATION STUDY OF LOSS COMPONENTS ON DRONE NAVIGATION (SATELLITE → DRONE) TASK.

Loss Components			H-Metrics			Small Scale		Middle Scale		Large Scale		Overall	
$\mathcal{L}_{\text{clust}}$	$\mathcal{L}_{\text{HAPPIER}}$	$\mathcal{L}_{\text{DyCL}}$	H-AP	ASI	NDCG	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1
✓			30.88	37.56	72.90	31.45	56.67	30.70	66.67	28.70	71.33	30.28	64.89
	✓		43.29	51.98	78.20	35.18	63.60	42.11	75.26	54.55	82.78	43.95	73.88
		✓	45.33	52.06	78.35	30.70	58.68	43.96	75.39	63.30	86.77	45.99	73.61
✓	✓		44.95	53.75	80.36	**37.13**	**66.67**	43.71	76.33	54.66	85.67	45.17	76.22
✓		✓	46.02	55.63	80.22	32.31	60.33	44.56	76.00	**64.79**	88.33	47.22	74.89
✓	✓	✓	**47.73**	**56.27**	**81.06**	36.08	64.67	**45.77**	**78.33**	63.78	**89.00**	**48.54**	**77.33**

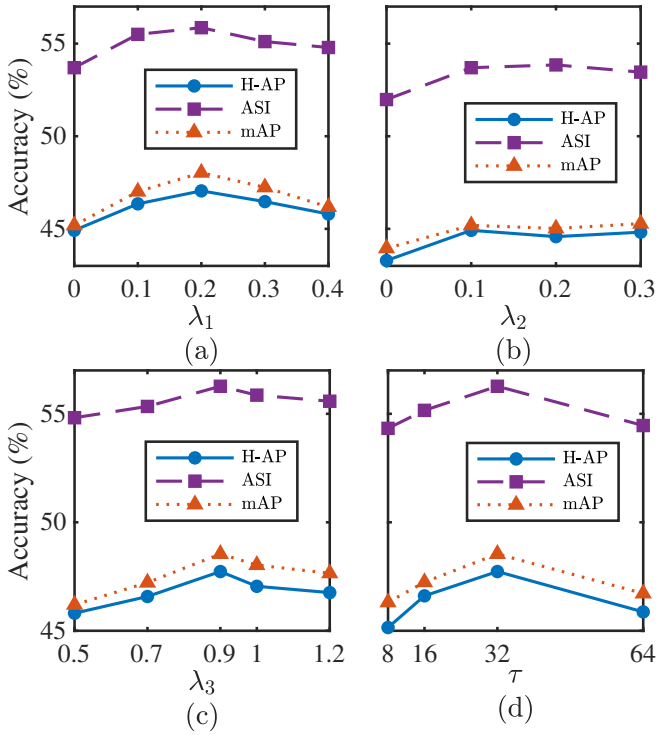


Fig. 8. Analysis of hyper-parameters. For simplicity, only results for the drone navigation (Satellite → Drone) task are shown. (a)-(c) Impact of the weighting coefficients λ_1, λ_2, and λ_3 in the combined loss function defined in Eq. (6), respectively. (d) Impact of the scaling factor τ in Eq. (4).

(a) Drone navigation on DA-Campus (Satellite → Drone)

(b) Drone-view target localization on DA-Campus (Drone → Satellite)

Fig. 9. Qualitative results of DyCL on the DA-Campus dataset. (a) Top-6 retrieval results for drone navigation. (b) Top-6 retrieval results for drone-view target localization. Consistent with the color scheme in Fig. 5, the border colors indicate the geographic relevance between the retrieved images and the query: Green represents the exact match at the small scale ($\mathcal{S}_c^0$). Blue and Yellow indicate geographically proximal neighbors at the middle ($\mathcal{S}_c^1$) and large ($\mathcal{S}_c^2$) scales, respectively. Red denotes distant negative samples ($\mathcal{S}_c^{>2}$). Therefore, an ideal distance-aware ranking should follow the sequence: Green → Blue → Yellow → Red.

dense distribution of target buildings, where adjacent structures often exhibit high morphological homogeneity. This visual ambiguity not only complicates feature discrimination between buildings but also results in images containing multiple buildings in a complex perspective relationship, rather than a single isolated target. Consequently, achieving precise localization requires the model to comprehensively understand the complex spatial relationships among multiple buildings. We believe addressing this challenge of DACVGL holds significant value for practical navigation and localization tasks.

VI. CONCLUSION

This paper revisits CVGL from a distance-aware perspective and introduces the DACVGL task. As a typical hierarchical contrastive learning problem, DACVGL exhibits two notable properties. First, supervision at multiple scales is inherently interrelated. Jointly learning embeddings within a unified feature space benefits overall model performance. Second, supervision contains complex structural information, which motivates the adoption of a contrastive learning approach rather than conventional metric learning. To address these challenges, we construct the DA-Campus benchmark and propose a novel DyCL framework. Experimental results demonstrate that DyCL can serve as a strong baseline for DACVGL. Beyond quantitative improvements, it offers practical advantages by mitigating severe localization errors and ensuring that top-ranked results remain spatially proximal to the target.

Future work will focus on a more comprehensive investigation of DACVGL, especially in more heterogeneous real-world scenarios. Currently, the data of DA-Campus is primarily collected from university campuses via synthetic UAV flights, which limits the diversity of architectural styles and environmental conditions. To further narrow the domain gap between

synthetic views and practical CVGL applications, we plan to extend our construction protocol to broader urban landscapes and introduce realistic weather and dynamic illumination variations into imagery. In such scenarios, both the diversity of building appearances and the high visual similarity among adjacent targets pose more complex challenges to retrieval tasks. To tackle these challenges, more specialized network architectures will be further designed and validated within the framework of DACVGL.

REFERENCES

[1] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays, "Learning deep representations for ground-to-aerial geolocalization," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5007–5015.

[2] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1437–1451, 2018.

[3] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang, "Ranking with local regression and global alignment for cross media retrieval," in *Proceedings of the 17th ACM International Conference on Multimedia*, 2009, p. 175–184.

[4] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[5] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 5962–5979, 2022.

[6] G. Huang, Y. Zhou, L. Zhao, and W. Gan, "Cv-cities: Advancing cross-view geo-localization in global cities," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 1592–1606, 2025.

[7] F. Deuser, K. Habel, and N. Oswald, "Sample4geo: Hard negative sampling for cross-view geo-localisation," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 16 801–16 810.

[8] Y. Zhu, H. Yang, Y. Lu, and Q. Huang, "Simple, effective and general: A new backbone for cross-view image geo-localization," 2023. [Online]. Available: https://arxiv.org/abs/2302.01572

[9] S. Hu, M. Feng, R. M. H. Nguyen, and G. H. Lee, "Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7258–7267.

[10] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1–9.

[11] L. Liu and H. Li, "Lending orientation to neural networks for cross-view geo-localization," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5617–5626.

[12] Z. Zheng, Y. Wei, and Y. Yang, "University-1652: A multi-view multi-source benchmark for drone-based geo-localization," *ACM Multimedia*, 2020.

[13] Z. Zheng, Y. Shi, T. Wang, J. Liu, J. Fang, Y. Wei, and T.-s. Chua, "Uavm'23: 2023 workshop on uavs in multimedia: Capturing the world from a new perspective," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 9715–9717.

[14] E. Ramzi, N. Audebert, N. Thome, C. Rambour, and X. Bitot, "Hierarchical average precision training for pertinent image retrieval," in *Computer Vision – ECCV 2022*, 2022, pp. 250–266.

[15] Y. Sun, Y. Zhu, Y. Zhang, P. Zheng, X. Qiu, C. Zhang, and Y. Wei, "Dynamic metric learning: Towards a scalable metric space to accommodate multiple semantic scales," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 5389–5398.

[16] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, p. 422–446, Oct. 2002. [Online]. Available: https://doi.org/10.1145/582415.582418

[17] Y. Shi, X. Yu, L. Liu, T. Zhang, and H. Li, "Optimal feature transport for cross-view image geo-localization," 2019. [Online]. Available: https://arxiv.org/abs/1907.05021

[18] Y. Shi, L. Liu, X. Yu, and H. Li, "Spatial-aware feature aggregation for image based cross-view geo-localization," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[19] Y. Shi, X. Yu, D. Campbell, and H. Li, "Where am i looking at? joint location and orientation estimation by cross-view matching," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4063–4071.

[20] N. N. Vo and J. Hays, "Localizing and orienting street views using overhead imagery," in *Computer Vision – ECCV 2016*, 2016, pp. 494–509.

[21] P. Xia, Y. Wan, Z. Zheng, Y. Zhang, and J. Deng, "Enhancing cross-view geo-localization with domain alignment and scene consistency," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 12, pp. 13 271–13 281, 2024.

[22] Q. Wu, Y. Wan, Z. Zheng, Y. Zhang, G. Wang, and Z. Zhao, "Camp: A cross-view geo-localization method using contrastive attributes mining and position-aware partitioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[23] Z. Chen, Z.-X. Yang, and H.-J. Rong, "Multilevel embedding and alignment network with consistency and invariance learning for cross-view geo-localization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–15, 2025.

[24] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[25] P. Li, P. Pan, P. Liu, M. Xu, and Y. Yang, "Hierarchical temporal modeling with mutual distance matching for video based person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 503–511, 2021.

[26] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, 2006, pp. 1735–1742.

[27] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[28] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3774–3782.

[29] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, "Dual-path convolutional image-text embeddings with instance loss," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 16, no. 2, May 2020. [Online]. Available: https://doi.org/10.1145/3383184

[30] F. Tramer, N. Carlini, W. Brendel, and A. Madry, "On adaptive attacks to adversarial example defenses," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[31] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1096–1104.

[32] R. Fagin, R. Kumar, and D. Sivakumar, "Comparing top k lists," in *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '03. USA: Society for Industrial and Applied Mathematics, 2003, p. 28–36.

[33] A. Zhai and H.-Y. Wu, "Classification is a strong baseline for deep metric learning," 2019. [Online]. Available: https://arxiv.org/abs/1811.12649

[34] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3652–3661.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[37] T. Wang, Z. Zheng, C. Yan, J. Zhang, Y. Sun, B. Zheng, and Y. Yang, "Each part matters: Local patterns facilitate cross-view geo-localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 2, pp. 867–879, 2022.

[38] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2019. [Online]. Available: https://arxiv.org/abs/1807.03748

[39] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever,

"Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 18–24 Jul 2021, pp. 8748–8763.

[40] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5017–5025.

[41] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2024. [Online]. Available: https://arxiv.org/abs/2304.07193

[42] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krähenbühl, "Sampling matters in deep embedding learning," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2859–2867.

[43] M. G. KENDALL, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1-2, pp. 81–93, 06 1938. [Online]. Available: https://doi.org/10.1093/biomet/30.1-2.81

Suofei Zhang received the B.S. and M.S. degrees in Testing and Measuring Technology and Instrumentation from Jiangsu University, Zhenjiang, P. R. China, in 2004 and 2007, respectively, and the Ph.D. degree in Information Science and Engineering from Southeast University, Nanjing, P. R. China, in 2013. He was a Visiting Scholar with ENSTA Paris, France, from 2009 to 2010.

He is currently a Lecturer with the School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, P. R. China. His research interests include computer vision, image retrieval, drone navigation and camera localization.

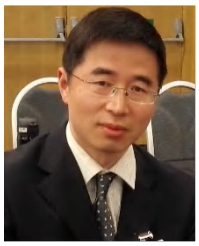
Xinxin Wang was born in 1999. He received the M.S. degree in Communication and Information Systems from Nanjing University of Posts and Telecommunications, Nanjing, P. R. China, in 2025. He is currently an Engineer with China Mobile Communications Group Co., Ltd., Lianyungang, China. His research interests include computer vision and geolocation techniques.

Xiaofu Wu received the B.S. and M.S. degrees in electrical engineering from the Nanjing Institute of Communications Engineering, Nanjing, China, in 1996 and 1999, respectively, and the Ph.D. degree in electrical engineering from the Peking University, Beijing, China, in 2005. From 2005 to 2007, he was with the Southeast University as a Post-Doctoral researcher at the National Mobile Communication Research Laboratory.

Since 2012, he has been with the Nanjing University of Posts & Telecommunications, where he is currently a full Professor. His research interests are in coding and information theory, information-theoretic security, machine learning and computer vision.

Quan Zhou (Senior Member, IEEE) received the B.S. degree in electronics and information engineering from China University of Geosciences, Wuhan, P. R. China, in 2002, and the M.S. and Ph.D. degrees in electronics and information engineering from Huazhong University of Science and Technology, Wuhan, P. R. China, in 2006 and 2013, respectively. He was a Visiting Scholar with Temple University, Philadelphia, PA, USA, from 2019 to 2020.

He is currently a Full Professor with Nanjing University of Posts and Telecommunications, Nanjing, P. R. China. He has authored or coauthored more than 100 related academic articles, including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON MEDICAL IMAGING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and *Pattern Recognition*. His research interests include deep learning, pattern recognition, and computer vision.



Haifeng Hu received the B.E. degree from Anhui University, Anhui, China, and the Ph.D. degree in signal processing from the Nanjing University of Posts and Telecommunications, in 2008. From 2012 to 2013, he visited Columbia University as a Visiting Researcher.

He is currently a Full Professor with the Department of Telecommunication and Information Engineering, Nanjing University of Posts and Telecommunications. His research interests include deep learning for communication networks and other AI-assisted applications.