Aligning Geometric Spatial Layout in Cross-View Geo-Localization via Feature Recombination

Qingwang Zhang, Yingying Zhu[*]

College of Computer Science and Software Engineering, Shenzhen University, China
zhangqingwang2022@email.szu.edu.cn, zhuyy@szu.edu.cn

Abstract

Cross-view geo-localization holds significant potential for various applications, but drastic differences in viewpoints and visual appearances between cross-view images make this task extremely challenging. Recent works have made notable progress in cross-view geo-localization. However, existing methods either ignore the correspondence between geometric spatial layout in cross-view images or require high costs or strict constraints to achieve such alignment. In response to these challenges, we propose a Feature Recombination Module (FRM) that explicitly establishes the geometric spatial layout correspondences between two views. Unlike existing methods, FRM aligns geometric spatial layout by directly recombining features, avoiding image preprocessing, and introducing no additional computational and parameter costs. This effectively reduces ambiguities caused by geometric misalignments between ground-level and aerial-level images. Furthermore, it is not sensitive to frameworks and applies to both CNN-based and Transformer-based architectures. Additionally, as part of the training procedure, we also introduce a novel weighted $(B+1)$-tuple loss (WBL) as optimization objective. Compared to the widely used weighted soft margin ranking loss, this innovative loss enhances convergence speed and final performance. Based on the two core components (FRM and WBL), we develop an end-to-end network architecture (FRGeo) to address these limitations from a different perspective. Extensive experiments show that our proposed FRGeo not only achieves state-of-the-art performance on cross-view geo-localization benchmarks, including CVUSA, CVACT, and VIGOR, but also is significantly superior or competitive in terms of computational complexity and trainable parameters. Our project homepage is at https://zqwlearning.github.io/FRGeo.

Introduction

The goal of cross-view geo-localization is to determine the geographical location of a ground image (known as a query image) from geo-tagged aerial images (known as reference images) without relying on GPS or other positioning devices. Existing methods to cross-view geo-localization commonly frame the problem as a retrieval task. In practical deployment, the task involves retrieving the reference image that is most similar to the query image and utilizing its location label as predictive result. This task offers an alternative means for geo-localization in real scenarios, particularly crucial in environments where GPS signals are obstructed or perturbed by noise. The potential applications of this task are extensive, encompassing areas such as autonomous driving (Häne et al. 2017; Kim and Walter 2017), robotic navigation (McManus et al. 2014), and 3D reconstruction (Middelberg et al. 2014).

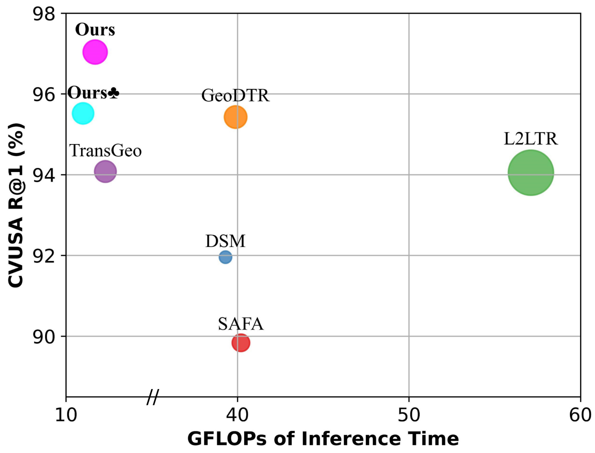

Figure 1: Performance comparison on CVUSA R@1. Bubble size indicates the number of trainable parameters. Ours♣ indicates the integration of FRM and WBL into the TransGeo 1-stage, which is a pure Transformer-based method. Ours (FRGeo) achieves the highest R@1 while enjoying significantly less number of trainable parameters and GFLOPs.

Despite the enticing potential for application, the task of cross-view matching presents substantial challenges due to the dramatic changes in viewpoints and visual appearances between ground and aerial images. Consequently, it is paramount to understand and correspond both image content (appearances and semantics) and geometric spatial layout across views. Considering that the correspondence of geometric spatial layout can be implicitly modeled by models autonomously or guided by external priors, existing methods (Shi et al. 2019, 2020) for aligning the geometric spatial layout between different views achieve alignment by warping aerial images to match ground images. This help

[*] Corresponding author.

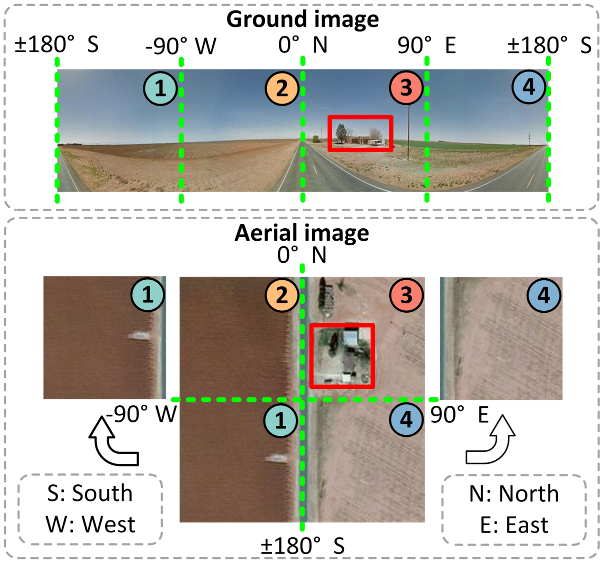

Figure 2: Geometric spatial layout correspondence between two views. The context information contained in the same indexed regions is closely related. For example, a building in region ③ of the ground image is also located in region ③ of the aerial view (the red boxes indicate the building).

reduce ambiguities caused by geometric misalignments between cross-view images. However, such a method results in obvious distortions in appearance, introduces additional image preprocessing steps, and requires assumptions of spatial alignment to the center and orientation. Other endeavors (Liu and Li 2019) improve performance by introducing orientation information for each pixel through the addition of orientation maps, yet this also introduces strict constraints and augmented computational costs. While these methods hold considerable promise, they either entail intricate designs, incurring substantial preprocessing time and computational costs, or entail strict dataset requirements. The presence of the aforementioned issues limits the applicability of these methods, prompting us to seek a low-cost method to minimize cross-view image misalignment, accomplish alignment of geometric spatial layout, and relax strict dataset requirements, thereby enabling wider applications.

Contrary to traditional strategies, our method is centered around establishing a clear geometric spatial layout correspondence between two views at the region level, which derives discriminative matching cues from these approximately aligned regions. Specifically, we observe that both ground and aerial images cover the same field of view (FoV, observed visible region) range, and they can be naturally divided into 4 regions according to North (0°), East (90°), South (±180°) and West ($-90°$). We use 4 indexes ①, ②, ③, and ④ to represent these regions, as shown in Figure 2. Since the regions with the same index are different representations of the same FoV under different views, the context information they contain are closely related. Inspired by the above observation, we propose the Feature Recombination Module (FRM). FRM uses the division in Figure 2 to divide the feature maps into different regions and performs spatial average pooling within each region, then recombines to obtain final representations.

Significantly, unlike polar transform or the addition of orientation maps to the network, our method does not need to strictly align the geometric spatial layout between two views at the **pixel** level. Instead, we adopt a simpler and more flexible alignment method to approximately align at the **region** level, which has no additional transformations or precise alignment, and is therefore more realistic, more tractable and more widely applicable (even to datasets with non-central alignment, *e.g.*, VIGOR). In addition, our method operates directly on feature maps, thereby avoiding any appearance distortion and image preprocessing. Remarkably, it introduces no additional computational or parameter costs, and thanks to its simple design, FRM can be plugged into any CNN or Transformer (Vaswani et al. 2017) architecture.

We also delve into the loss, which is one of the crucial part of the cross-view geo-localization task. Previous works have widely used weighted soft margin ranking loss (Hu et al. 2018), which has the limitation of considering only one negative sample during the construction of a triplet while not interacting with the other negative samples in each update. To address this issue, we propose a novel weighted $(B+1)$-tuple loss (WBL) as our optimization objective that allows joint comparison multiple negative samples and introduces a weighted coefficient α. This proposed loss enhances convergence speed and final performance. Extensive experiments demonstrate that our method (named FRGeo with FRM and WBL as key components) not only achieves state-of-the-art performance but also exhibits significant advantages or competitiveness in terms of computational complexity and trainable parameters, as illustrated in Figure 1.

Our main contributions can be summarized as follows:

- We propose a novel Feature Recombination Module (FRM), which explicitly establishes the correspondence of geometric spatial layouts between two views at the region level, to reduce ambiguities caused by geometric misalignments. It has the advantages of no image preprocessing, being lightweight, and plug-and-play.
- We design a weighted $(B+1)$-tuple loss (WBL) as part of the training procedure, enabling simultaneously pushing away of multiple negative samples, which effectively speeds convergence and improves performance compared to traditional weighted soft margin ranking loss.
- The Feature Recombination Geo-localization network (FRGeo) outperforms previously developed deep networks for the cross-view geo-localization task on CVUSA, CVACT, and VIGOR datasets. Furthermore, FRGeo exhibits a noteworthy advantage or competitiveness in terms of computational complexity and trainable parameters.

Related Work

We roughly categorize existing cross-view geo-localization methods into feature-based and geometry-based methods.

Feature-based Methods

Feature-based methods focus on learning discriminative image representations to differentiate similar images. Workman, Souvenir, and Jacobs (2015) first introduce CNNs to cross-view matching, drawing inspiration from the success of CNNs in the computer vision (Krizhevsky, Sutskever, and Hinton 2012). Subsequently, Hu et al. (2018) integrate the NetVlad (Arandjelovic et al. 2016) with a dual-branch VGG (Simonyan and Zisserman 2015) backbone network to obtain viewpoint-invariant representations. They also propose a weighted soft margin ranking loss to expedite network training, an optimization objective that has found widespread application in subsequent research. Despite promising, most of the above feature-based methods rely on models implicitly modeling spatial information and rarely pay enough attention to the importance of explicitly aligning geometric spatial layouts. In this study, we explicitly establish the geometric spatial layout correspondence between views via a Feature Recombination Module, to reduce ambiguities caused by geometric misalignments.

Geometry-based Methods

Geometry-based methods aim to reduce ambiguities caused by geometric misalignments between ground and aerial images. Liu and Li (2019) introduce orientation maps to inject the orientation information of each pixel into the network. Shi et al. (2019) uses polar transform to warp aerial images, aligning the geometric spatial layout of ground-aerial image pairs. Subsequently, the same authors introduced DSM (Shi et al. 2020) using a sliding window for geo-localization of limited field of view ground images. CDE (Toker et al. 2021) combines GAN (Goodfellow et al. 2014) and SAFA (Shi et al. 2019) for geo-localization and ground image synthesis. GeoDTR (Zhang et al. 2023) extracts geometric layout descriptors from raw features, also proposing layout simulation and semantic data augmentations. While the above methods have improved performance, many still rely on polar transform for fine-grained geometric spatial layout alignment, leading to appearance distortions and additional preprocessing. Moreover, these methods exhibit strict dataset requirements, rendering them unsuitable for non-centrally aligned datasets, *e.g.*, VIGOR (Zhu, Yang, and Chen 2021b). Our method, however, aligns at the more macro level, avoiding pixel-level micro geometry alignment. Consequently, it does not demand data to possess strict center alignment properties, accommodating non-centrally aligned datasets, *e.g.*, VIGOR. Remarkably, benefitted from the model design, our method does not rely on polar transform while introducing no additional computational or parameter costs.

Recent researches, several methods employing the Transformer as backbone have emerged. L2LTR (Yang, Lu, and Zhu 2021) explores a hybrid ViT-based model, whereas TransGeo (Zhu, Shah, and Chen 2022) introduces a pure Transformer-based model. These methods exclusively rely on the Transformer to implicitly model spatial information. Nevertheless, our method explicitly aligns the geometric spatial layout across different views, thereby reducing ambiguities caused by geometry misalignments and leading to enhanced performance. Furthermore, in comparison to L2LTR, FRGeo exhibits conspicuous advantages in terms of computational complexity and trainable parameters, all without necessitating the 2-stage training paradigm proposed by TransGeo.

Methodology

Problem Formulation

A set of ground-aerial image pairs is denoted as $\{(\mathbf{I}_i^g, \mathbf{I}_i^a)\}^N$, where the superscripts g and a denote ground and aerial images, respectively; N denotes the number of pairs. Each ground-aerial image pair corresponds to distinct geo-location, where the geo-tags is unknown for ground images $\{\mathbf{I}_i^g\}^N$ and known for aerial images $\{\mathbf{I}_i^a\}^N$. In cross-view geo-localization task, given a query ground image $\mathbf{I}_q^g$ with index q, $q \in \{1, 2, ..., N\}$, the objective is to retrieve the optimal matching reference aerial image $\mathbf{I}_r^a, r \in \{1, 2, ..., N\}$, to determine the specific geo-location of $\mathbf{I}_q^g$.

For a given set $\{(\mathbf{I}_i^g, \mathbf{I}_i^a)\}^N$, we infer the corresponding image representations as $\{(\mathbf{f}_i^g, \mathbf{f}_i^a)\}^N$. These representations are expected to possess the following properties: the distance between matched image pairs is smaller than the distance between unmatched image pairs, expressed as $d(\mathbf{f}_q^g, \mathbf{f}_q^a) < \{d(\mathbf{f}_q^g, \mathbf{f}_i^a) | \forall i \in \{1, ..., N\}, i \neq q\}$, $d(\cdot, \cdot)$ denotes the L_2 distance. Consequently, the cross-view geo-localization task can be made explicit as:

$$r = \underset{i \in \{1,...,N\}}{\arg\min}\ d(\mathbf{f}_q^g, \mathbf{f}_i^a) \tag{1}$$

If the retrieval is correct, r equals q. For the sake of notation simplicity, we will omit the subscript i in the subsequent sections, except when discussing the loss function.

FRGeo Model

Model Overview. The propose model (FRGeo) introduces a Siamese neural network composed of two branches: the ground and aerial view branches, as depicted in Figure 3 (a). For a given ground-aerial image pair $(\mathbf{I}^g, \mathbf{I}^a)$, a preliminary stage involves the extraction of raw features utilizing either CNN-based or Transformer-based backbone. This extraction obtains $\mathbf{F}^g \in \mathbb{R}^{H^g \times W^g \times C}$ and $\mathbf{F}^a \in \mathbb{R}^{H^a \times W^a \times C}$, where H^g, W^g, H^a, and W^a correspond to the height and width of raw features from the ground and aerial images, and C denotes channel of raw features. Subsequently, the raw features $\mathbf{F}^g$ and $\mathbf{F}^a$ undergo processing through the Feature Recombination Module (FRM) to obtain the final image feature representations, $\mathbf{f}^g \in \mathbb{R}^{4C}$ and $\mathbf{f}^a \in \mathbb{R}^{4C}$. The optimization of model parameters is achieved by employing our proposed weighted $(B+1)$-tuple loss (WBL). In the following, we will describe the core components of FRGeo in detail.

Feature Recombination Module. The FRM utilizes raw features $\mathbf{F}^g$ and $\mathbf{F}^a$ extracted by the backbone as its inputs, obtaining the final image feature representations $\mathbf{f}^g$ and $\mathbf{f}^a$ as outputs. On the spatial dimension, the raw feature of each view is divided into 4 regions according to the division method shown in Figure 2, as shown in Figure 3 (b). Considering $\mathbf{F}_{\text{SW}}^g$, $\mathbf{F}_{\text{WN}}^g$, $\mathbf{F}_{\text{NE}}^g$ and $\mathbf{F}_{\text{ES}}^g$ denote ①, ②, ③, and ④ regions of the ground raw feature $\mathbf{F}^g$; $\mathbf{F}_{\text{SW}}^a$, $\mathbf{F}_{\text{WN}}^a$, $\mathbf{F}_{\text{NE}}^a$

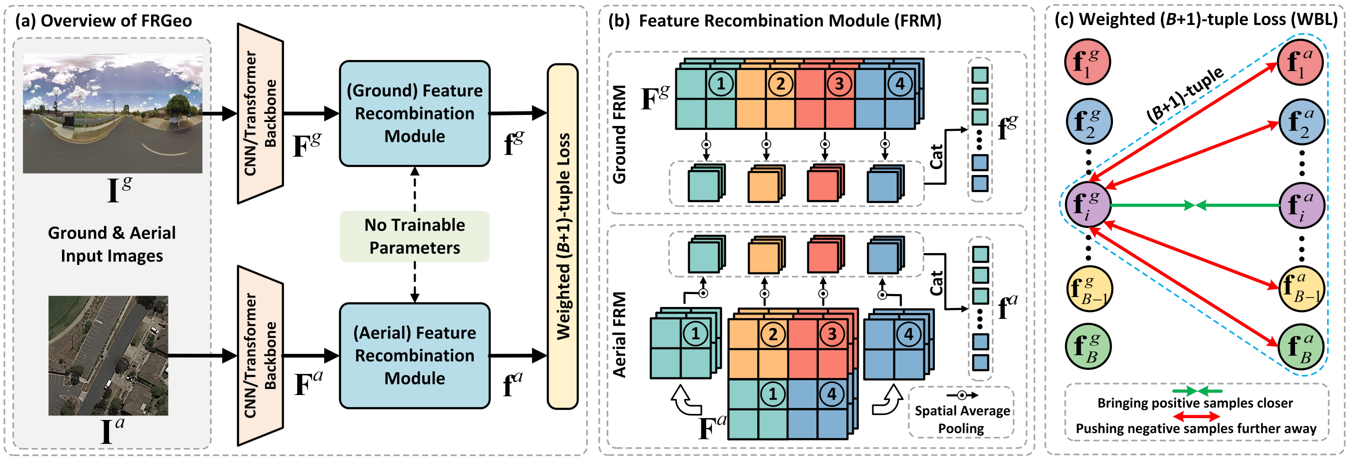

Figure 3: (a) Overview of our proposed FRGeo model. (b) Illustration of the proposed Feature Recombination Module (FRM). (c) Illustration of the proposed Weighted $(B+1)$-tuple Loss (WBL).

and $\mathbf{F}^a_{\text{ES}}$ denote ①, ②, ③, and ④ regions of the aerial raw feature $\mathbf{F}^a$, respectively. Their formal representations:

$$\mathbf{F}^g_{\text{SW}} = \mathbf{F}^g(:, 0 : W^g/4, :) \tag{2}$$

$$\mathbf{F}^g_{\text{WN}} = \mathbf{F}^g(:, W^g/4 : W^g/2, :) \tag{3}$$

$$\mathbf{F}^g_{\text{NE}} = \mathbf{F}^g(:, W^g/2 : 3W^g/4, :) \tag{4}$$

$$\mathbf{F}^g_{\text{ES}} = \mathbf{F}^g(:, 3W^g/4 : W^g, :) \tag{5}$$

$$\mathbf{F}^a_{\text{SW}} = \mathbf{F}^a(H^a/2 : H^a,\ 0 : W^a/2, :) \tag{6}$$

$$\mathbf{F}^a_{\text{WN}} = \mathbf{F}^a(0 : H^a/2, 0 : W^a/2, :) \tag{7}$$

$$\mathbf{F}^a_{\text{NE}} = \mathbf{F}^a(0 : H^a/2, W^a/2 : W^a, :) \tag{8}$$

$$\mathbf{F}^a_{\text{ES}} = \mathbf{F}^a(H^a/2 : H^a, W^a/2 : W^a, :) \tag{9}$$

where $/$ denotes integer division. The final image feature representations $\mathbf{f}^g$ and $\mathbf{f}^a$ are calculated by:

$$\mathbf{f}^g = Cat(P_{avg}(\mathbf{F}^g_{\text{SW}}), P_{avg}(\mathbf{F}^g_{\text{WN}}), P_{avg}(\mathbf{F}^g_{\text{NE}}), P_{avg}(\mathbf{F}^g_{\text{ES}})) \tag{10}$$

$$\mathbf{f}^a = Cat(P_{avg}(\mathbf{F}^a_{\text{SW}}), P_{avg}(\mathbf{F}^a_{\text{WN}}), P_{avg}(\mathbf{F}^a_{\text{NE}}), P_{avg}(\mathbf{F}^a_{\text{ES}})) \tag{11}$$

where $Cat(\cdot,\cdot)$ denotes the concatenation, and $P_{avg}(\cdot)$ denotes spatial average pooling.

Optimization Objective. In previous works (Hu et al. 2018; Shi et al. 2019; Yang, Lu, and Zhu 2021; Zhu, Shah, and Chen 2022), the most widely employed loss is weighted soft margin ranking loss (Hu et al. 2018), which is computed by constructing triplets within each mini-batch. The problem lies in the fact that this loss uses only one negative sample in each update, consequently limiting the effective utilization of information from other negative samples. It ends up with slow convergence and suboptimal performance. Drawing inspiration from the work by Sohn (2016), we propose the weighted $(B + 1)$-tuple loss (WBL). WBL employs the construction of $(B + 1)$-tuple thus pushing away the distance between the anchor sample and all other $B - 1$ negative samples within mini-batch during each update, as depicted in Figure 3 (c). The formulation of WBL is provided below.

For a set of mini-batch samples $\{(\mathbf{I}^g_i, \mathbf{I}^a_i)\}^B$, corresponding to this are collections of image feature representations $\{(\mathbf{f}^g_i, \mathbf{f}^a_i)\}^B$, where B denotes the number of pairs in the mini-batch. When $\mathbf{f}^g_i$ is chosen as the anchor sample, the corresponding positive sample is denoted as $\mathbf{f}^a_i$, while the set of negative samples is denoted as $\{\mathbf{f}^a_j\}^B_{j\neq i}$. Within each mini-batch, it is feasible to construct $2B$ $(B + 1)$-tuples. To improve the convergence rate, we introduce a weighted coefficient α to $\left(d(\mathbf{f}^g_i, \mathbf{f}^a_i) - d(\mathbf{f}^g_i, \mathbf{f}^a_j)\right)$, resulting in WBL, which serves as our optimization objective:

$$\mathcal{L}_{WBL}(\mathbf{f}^g_i, \mathbf{f}^a_i, \{\mathbf{f}^a_j\}^B_{j\neq i}) =$$
$$log\left(1 + \sum_{j=1, j\neq i}^{B} \exp\left[\alpha\left(d(\mathbf{f}^g_i, \mathbf{f}^a_i) - d(\mathbf{f}^g_i, \mathbf{f}^a_j)\right)\right]\right) \tag{12}$$

where $d(\cdot,\cdot)$ denotes the L_2 distance. Loss in a mini-batch can be calculated by the following equation:

$$\mathcal{L}(\{(\mathbf{f}^g_i, \mathbf{f}^a_i)\}^B) = \frac{1}{2B}\sum_{c\in\mathbb{C}}\sum_{i=1}^{B}\mathcal{L}_{WBL}(\mathbf{f}^c_i, \mathbf{f}^c_i, \{\mathbf{f}^{\mathbb{C}-c}_j\}^B_{j\neq i}) \tag{13}$$

where $\mathbb{C}$ denotes the set of superscripts on the view $\{g, a\}$.

Experiment

Datasets and Experimental Settings

Datasets. We evaluate our method on three public cross-view geo-localization datasets: CVUSA (Zhai et al. 2017), CVACT (Liu and Li 2019) and VIGOR (Zhu, Yang, and Chen 2021b). CVUSA and CVACT support standard and fine-grained cross-view geo-localization, both of which are one-to-one retrievals; VIGOR supports beyond one-to-one retrieval, *i.e.*, one-to-many retrieval.

- **CVUSA** contains 35,532 image pairs for training and 8,884 image pairs for testing. This dataset consists of images mainly collected at suburban areas.
- **CVACT** provides 35,532 image pairs for training and 8,884 image pairs for validation (CVACT_val). It also provides 92,802 image pairs to support fine-grained city-scale geo-localization (CVACT_test). These images cover the urban area (Canberra) densely.

- **VIGOR** comprises 105,214 ground images and 90,618 aerial images, which assuming that the query ground images can belong to arbitrary locations in the target area without center-aligned settings. We follow the setting of VIGOR with both Same-area and Cross-area protocols.

Evaluation Metrics. Following previous works (Hu et al. 2018; Liu and Li 2019; Shi et al. 2019, 2020), we utilize the R@$K, K = \{1, 5, 10, 1\%\}$ metrics to evaluate our model, which represents the probability of correct matches among the top-K retrieved results. Additionally, for VIGOR, we report the hit rate, which denotes the probability that the top-1 retrieved reference image covers the query image.

Implementation Details. We employ a ConvNeXt-T (Liu et al. 2022) as the backbone with off-the-shelf pre-trained parameters on ImageNet-1K (Deng et al. 2009). α is set 10 in Equation (12). We train the model on a NVIDIA V100 Server with AdamW (Loshchilov and Hutter 2017) optimizer.

Comparison with State-of-the-art Methods

We compare our method with 8 state-of-the-art methods on the CVUSA and CVACT datasets, including SAFA (Shi et al. 2019), DSM (Shi et al. 2020), CDE (Toker et al. 2021), L2LTR (Yang, Lu, and Zhu 2021), TransGeo (Zhu, Shah, and Chen 2022), SEH (Guo et al. 2022), and GeoDTR (Zhang et al. 2023). On the VIGOR dataset, we compare our method with 5 state-of-the-art methods, including Siamese-VGG (Zhu, Yang, and Chen 2021a), SAFA, SAFA+Mining (Zhu, Yang, and Chen 2021b), VIGOR (Zhu, Yang, and Chen 2021b), and TransGeo. In the main paper, we evaluate the performance of our model for three tasks, including standard cross-view geo-localization, fine-grained cross-view geo-localization and beyond one-to-one retrieval.

Standard Cross-view Geo-localization. We first evaluate our model on the standard cross-view geo-localization task. The results on the CVUSA and CVACT_val datasets are shown in Table 1 and 2, respectively. The findings lead to the results that, in comparison to previous works, FRGeo achieves state-of-the-art or competitive performance. Remarkably, even without resorting to polar transform, FRGeo outperforms methods employing it. This highlights the benefit of our method in aligning geometric spatial layouts. Furthermore, we propose FRM and WBL can be seamlessly integrated into the TransGeo 1-stage, surpassing the raw TransGeo on metrics such as R@1, R@5, and R@10, and obtaining comparable performance in R@1%. This demonstrates that FRM and WBL are pluggable and not only applicable to CNN-based models, but also can significantly improve the performance of Transformer-based models.

Fine-grained Cross-view Geo-localization. In order to thoroughly evaluate the representational capacity of the model, we conducte a comprehensive evaluation of our method in the fine-grained cross-view geo-localization task. Specifically, we compare FRGeo with state-of-the-art methods on the challenging large-scale CVACT_test dataset - viz. 10× bigger than CVACT validation set, which is fully GPS-tagged for accurate localization. The results are shown in Table 2. Furthermore, we also report the experimental results of integrating FRM and WBL with the TransGeo 1-stage, arriving at conclusions consistent with the standard cross-view geo-localization. In comparison with all previous works, FRGeo achieves state-of-the-art or competitive performances. The results also demonstrate the superiority of our method.

Method	R@1	R@5	R@10	R@1%
SAFA	81.15%	94.23%	96.85%	99.49%
SAFA†	89.84%	96.93%	98.14%	99.64%
DSM†	91.93%	97.50%	98.54%	99.67%
CDE†	92.56%	97.55%	98.33%	99.57%
L2LTR	91.99%	97.68%	98.65%	99.75%
L2LTR†	94.05%	98.27%	98.99%	99.67%
TransGeo	94.08%	98.36%	99.04%	99.77%
SEH†	95.11%	98.45%	99.00%	99.78%
GeoDTR	93.76%	98.47%	99.22%	99.85%
GeoDTR†	95.43%	98.86%	99.34%	**99.86%**
Ours♣	95.52%	98.66%	99.13%	99.74%
Ours	**97.06%**	**99.25%**	**99.47%**	99.85%

Table 1: Comparisons between FRGeo (Ours) and state-of-the-art methods on the CVUSA dataset. † indicates applying polar transform to aerial images. Ours♣ indicates FRM and WBL integrating into the TransGeo 1-stage. Best and second best results shown in bold and underline, respectively.

Beyond One-to-one Retrieval. The beyond one-to-one retrieval task is performed on the recently introduced VIGOR. This dataset assumes that query images can belong to arbitrary locations in the target area, thus is not spatially aligned to the center. Consequently, it is the more complex and realistic benchmark. Many existing one-to-one retrieval methods falter in this context, however, our method remains well performing. In Table 3, our proposed method outperforms the competing methods by a substantial amount. For both Same-area and Cross-area evaluation protocols, the R@1 of our method achieve 71.26% (**+9.78%**) and 37.54% (**+18.55%**), respectively, with relative improvements of 15.91% and 97.68%. The above results demonstrate the powerful learning ability and widespread applicability of our method, as well as its robustness to cross-distribution shifts and the advantage of handling the datasets of spatially unaligned to the center.

Computational Costs

In Figure 1, the proposed method is compared with 5 state-of-the-art methods in terms of computational complexity and trainable parameters. It is intuitively clear from the figure that our method uses the least GFLOPs, which are just less than one-third of those of SAFA, DSM, L2LTR and GeoDTR. This observation implies that our method holds the potential for achieving faster processing speed and higher efficiency in practical applications. In terms of trainable parameters, our method is also competitive, especially when compared with L2LTR. It is important to emphasize that our method not only achieves the state-of-the-art performance, but also maintains the least computational complexity and competitive trainable parameters. These experimental results reflect the powerful

Method	CVACT_val				CVACT_test			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
SAFA	78.28%	91.60%	93.79%	98.15%	-	-	-	-
SAFA†	81.03%	92.80%	94.84%	98.17%	55.50%	79.94%	85.08%	94.49%
DSM†	82.49%	92.44%	93.99%	97.32%	35.63%	60.07%	69.10%	84.75%
CDE†	83.28%	93.57%	95.42%	98.22%	61.29%	85.13%	89.14%	98.32%
L2LTR	83.14%	93.84%	95.51%	98.40%	58.33%	84.23%	88.60%	95.83%
L2LTR†	84.89%	94.59%	95.96%	98.37%	60.72%	85.85%	89.88%	96.12%
TransGeo	84.95%	94.14%	95.78%	98.37%	-	-	-	-
SEH†	84.75%	93.97%	95.46%	98.11%	-	-	-	-
GeoDTR	85.43%	94.81%	96.11%	98.26%	62.96%	87.35%	90.70%	98.61%
GeoDTR†	86.21%	95.44%	96.72%	**98.77%**	64.52%	88.59%	91.96%	**98.74%**
Ours♣	88.60%	95.35%	96.26%	98.13%	69.52%	89.79%	92.36%	98.20%
Ours	**90.35%**	**96.45%**	**97.25%**	98.74%	**72.15%**	**91.93%**	**94.05%**	98.66%

Table 2: Comparison between FRGeo (Ours) and state-of-the-art methods on CVACT dataset. Notations are the same as Table 1.

Method	Same-area					Cross-area				
	R@1	R@5	R@10	R@1%	Hit	R@1	R@5	R@10	R@1%	Hit
Siamese-VGG	18.69%	43.64%	55.36%	97.55%	21.90%	2.77%	8.61%	12.94%	62.64%	3.16%
SAFA	33.93%	58.42%	68.12%	98.24%	36.87%	8.20%	19.59%	26.36%	77.61%	8.85%
SAFA+Mining	38.02%	62.87%	71.12%	97.63%	41.81%	9.23%	21.12%	28.02%	77.84%	9.92%
VIGOR	41.07%	65.81%	74.05%	98.37%	44.71%	11.00%	23.56%	30.76%	80.22%	11.64%
TransGeo	61.48%	87.54%	91.88%	**99.56%**	73.09%	18.99%	38.24%	46.91%	88.94%	21.21%
Ours	**71.26%**	**91.38%**	**94.32%**	99.52%	**82.41%**	**37.54%**	**59.58%**	**67.34%**	**94.28%**	**40.66%**

Table 3: Comparison between FRGeo (Ours) and state-of-the-art methods on VIGOR dataset, including Same-area and Cross-area protocols. Hit means hit rate (Zhu, Yang, and Chen 2021b). Notations are the same as Table 1.

practical value (lighter and more accurate) of our method.

Ablation Study

Effectiveness of Components. To demonstrate the effectiveness of our proposed components (FRM and WBL), we conducte a series of experiments by sequentially integrating these components into the Baseline model (*i.e.*, Baseline, Baseline + FRM, Baseline + WBL, Baseline + FRM + WBL). Specifically, the Baseline model adopts a Siamese architecture with ConvNeXt-T (Liu et al. 2022) serving as the backbone. To enable a fair comparison, the choice of hyperparameters and training strategy for all subsequent models remain entirely consistent with those of the Baseline. The results on the CVUSA, CVACT, and VIGOR are shown in Table 4. It is evident that upon the introduction of either FRM or WBL, there is a remarkable improvement in model performance. The optimal performance is achieved when both components are simultaneously applied, as observed in our FRGeo. These experimental results successfully validate the effectiveness of our proposed FRM and WBL.

Additionally, we monitored the evolution of the R@1 metric during the initial 1 to 20 epochs of training on the CVUSA and CVACT datasets, as depicted in Figure 4 (Left and Middle). It is apparent that the utilization of FRM and WBL not only improves performance but also speeds up convergence. Remarkably, after a mere 10 epochs of training, our method achieves performance comparable to, if not superior to, the other state-of-the-art methods that typically require at least 100 epochs to reach similar results. This excellent performance is attributed to the rationalization of the geometric spatial layouts of the FRM explicitly aligned cross-view images and the effectiveness of the WBL in pushing away multiple negative samples simultaneously.

Few-shot Training. To further verify the effectiveness of FRM and WBL, we conduct a series of few-shot training aimed at training a model capable of generalizing from a limited set of training samples (He et al. 2020). To support this task, we randomly select a certain percentage (100%/80%/60%/40%/20%) of samples from the CVUSA training set for training, while keeping the test set unchanged. These training subsets are employed to train the models, and subsequent testing is performed on the raw test set to observe the impact of varying training subset sizes on both the Baseline model and FRGeo, as shown in Figure 4 (Right). The results consistently demonstrate that the simultaneous omission of FRM and WBL continues to impair performance, particularly evident when the training subset size is smaller. For instance, with only 20% of samples participating in training, the Baseline drops by as much as 18.84% compared to the R@1 performance of FRGeo. This shows that the combi-

Method	CVUSA				VIGOR Same-area				
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	Hit
Baseline	94.10%	98.57%	99.22%	99.83%	56.25%	83.98%	88.96%	99.00%	68.93%
Baseline + FRM	96.70%	99.12%	99.39%	99.84%	67.38%	88.95%	92.38%	99.37%	77.86%
Baseline + WBL	95.27%	99.03%	99.38%	99.82%	66.58%	90.83%	93.99%	99.51%	80.53%
Baseline + FRM + WBL (**Ours**)	**97.06%**	**99.25%**	**99.47%**	**99.85%**	**71.26%**	**91.38%**	**94.32%**	**99.52%**	**82.41%**
	CVACT_val				CVACT_test				
Baseline	84.77%	95.24%	96.66%	**98.77%**	59.19%	86.59%	90.74%	98.71%	-
Baseline + FRM	88.79%	96.20%	97.02%	98.71%	68.35%	90.34%	93.08%	98.69%	-
Baseline + WBL	87.42%	96.06%	97.04%	98.75%	65.28%	89.28%	92.49%	**98.76%**	-
Baseline + FRM + WBL (**Ours**)	**90.35%**	**96.45%**	**97.25%**	98.74%	**72.15%**	**91.93%**	**94.05%**	98.66%	-

Table 4: Effectiveness of the proposed components. Sequentially integrate FRM and WBL into the baseline model, and their performance is reported on the CVUSA, CVACT and VIGOR datasets. Best results shown in bold.

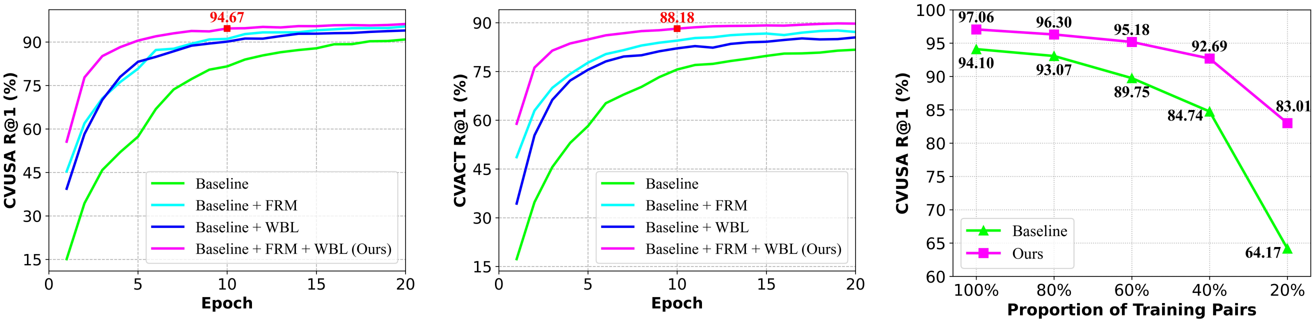

Figure 4: Left and Middle: The training curve (R@1) on CVUSA (Left) and CVACT (Middle) datasets. The red dot data indicates the performance of FRGeo (Ours) at the 10th training epoch. Right: Few-shot training on CVUSA dataset. 100% indicates using all training samples. Best viewed on screen with zoom-in.

nation of FRM and WBL not only improves performance but also enhances generalization ability.

Visualization Analysis

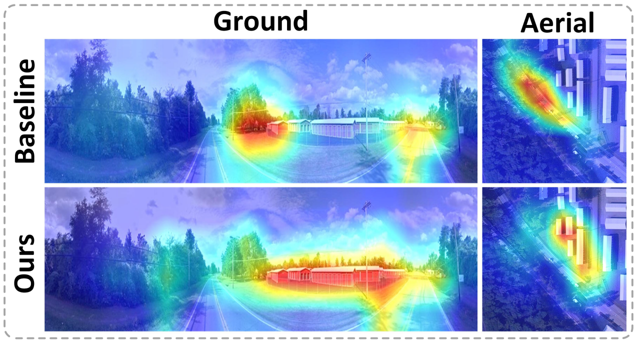

Figure 5: Heatmap visualization of the Baseline and FRGeo model. Best viewed on screen with zoom-in.

To comprehend what FRGeo has learned, and to compare the differences in the regions that different models focus on, we visualize heatmaps. Figure 5 shows the heatmaps of both the Baseline and FRGeo model. It is discernible that the Baseline mainly focuses on road information, while FRGeo pays more attention to some salient buildings in addition to road information. We argue that these buildings are more discriminative localization cues for cross-view geo-localization task, since buildings in different ground-aerial images often have substantial differences in appearance and layout. We attribute the heightened focus of FRGeo on discriminative regions (*e.g.*, salient buildings) to the efficacy of the FRM, which aligns geometric spatial layout alignment cross-view images, thereby making it easier for the model to learn them.

Conclusion

In this paper, we propose a novel and efficient cross-view geo-localization method for aligning the geometric spatial layout between cross-view images by feature recombination, reducing ambiguities caused by geometry misalignments and making discriminative localization cues easier to be learned. Moreover, we introduce the weighted $(B + 1)$-tuple loss, and show that it notably accelerates training speed and improves the performance of our method. Extensive experiments demonstrate that our method achieves state-of-the-art performance on the CVUSA, CVACT, and VIGOR datasets with significant advantages or competitiveness in terms of computational complexity and trainable parameters.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62072318, in part by the Key Project of Department of Education of Guangdong Province under Grant 2023ZDZX1016, and in part by Shenzhen Science and Technology Program under Grant 20220810142553001.

References

Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; and Sivic, J. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5297–5307.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Guo, Y.; Choi, M.; Li, K.; Boussaid, F.; and Bennamoun, M. 2022. Soft exemplar highlighting for cross-view image-based geo-localization. *IEEE transactions on image processing*, 31: 2094–2105.

Häne, C.; Heng, L.; Lee, G. H.; Fraundorfer, F.; Furgale, P.; Sattler, T.; and Pollefeys, M. 2017. 3D visual perception for self-driving cars using a multi-camera system: Calibration, mapping, localization, and obstacle detection. *Image and Vision Computing*, 68: 14–27.

He, J.; Hong, R.; Liu, X.; Xu, M.; Zha, Z.-J.; and Wang, M. 2020. Memory-augmented relation network for few-shot learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1236–1244.

Hu, S.; Feng, M.; Nguyen, R. M.; and Lee, G. H. 2018. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7258–7267.

Kim, D.-K.; and Walter, M. R. 2017. Satellite image-based localization via learned embeddings. In *2017 IEEE international conference on robotics and automation (ICRA)*, 2073–2080. IEEE.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Liu, L.; and Li, H. 2019. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5624–5633.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11976–11986.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

McManus, C.; Churchill, W.; Maddern, W.; Stewart, A. D.; and Newman, P. 2014. Shady dealings: Robust, long-term visual localisation using illumination invariance. In *2014 IEEE international conference on robotics and automation (ICRA)*, 901–906. IEEE.

Middelberg, S.; Sattler, T.; Untzelmann, O.; and Kobbelt, L. 2014. Scalable 6-dof localization on mobile devices. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, 268–283. Springer.

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

Shi, Y.; Yu, X.; Campbell, D.; and Li, H. 2020. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4064–4072.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition.

Sohn, K. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.

Toker, A.; Zhou, Q.; Maximov, M.; and Leal-Taixé, L. 2021. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6488–6497.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. *Neural Information Processing Systems,Neural Information Processing Systems*.

Workman, S.; Souvenir, R.; and Jacobs, N. 2015. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, 3961–3969.

Yang, H.; Lu, X.; and Zhu, Y. 2021. Cross-view geo-localization with layer-to-layer transformer. *Advances in Neural Information Processing Systems*, 34: 29009–29020.

Zhai, M.; Bessinger, Z.; Workman, S.; and Jacobs, N. 2017. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 867–875.

Zhang, X.; Li, X.; Sultani, W.; Zhou, Y.; and Wshah, S. 2023. Cross-view geo-localization via learning disentangled geometric layout correspondence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3480–3488.

Zhu, S.; Shah, M.; and Chen, C. 2022. Transgeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1162–1171.

Zhu, S.; Yang, T.; and Chen, C. 2021a. Revisiting street-to-aerial view image geo-localization and orientation estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 756–765.

Zhu, S.; Yang, T.; and Chen, C. 2021b. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3640–3649.