VimGeo: Efficient Cross-View Geo-Localization with Vision Mamba Architecture

Jinglin Huang[1,2,3], **Maoqiang Wu**[4], **Peichun Li**[5], **Wen Wu**[3], **Rong Yu**[1]
[1]School of Automation, Guangdong University of Technology
[2]Gauss Riemann Technologies Co., Ltd., Guangzhou, GD
[3]Frontier Research Center, Peng Cheng Laboratory
[4]School of Electronic Science and Engineering, South China Normal University
[5]Department of Computer and Information Science, University of Macau
{2112304047, yurong}@gdut.edu.cn, maoqiang.wu@vip.163.com, peichun.li@connect.um.edu.mo, wuw02@pcl.ac.cn

Abstract

Cross-view geo-localization is a crucial task with diverse applications, yet it remains challenging due to the significant variations in viewpoints and visual appearances between images from different perspectives. While recent advancements have been made, existing methods often suffer from high model complexity, excessive resource consumption, and the impact of sample learning difficulty on optimization. To overcome these limitations, we optimize the Vision Mamba (Vim) model, built on a State Space Model (SSM) architecture, by replacing the traditional classification head with *Channel Group Pooling (CGP)* for efficient feature integration. This optimization reduces model parameters by 1.5% and computational complexity by 0.4%. Additionally, we propose a novel *Dynamic Weighted Batch-tuple Loss (DWBL)* to dynamically adjust the weighting of negative samples, improving model performance. By combining CGP and DWBL, we develop an efficient end-to-end network, VimGeo, which achieves state-of-the-art performance with enhanced computational efficiency. Specifically, VimGeo achieves a Recall@1 of **81.67%** on the CVACT_test dataset, outperforming prior approaches. Extensive experiments on CVUSA, CVACT, and VIGOR datasets validate VimGeo's effectiveness and competitiveness in cross-view geo-localization tasks, achieving the leading results among sequence modeling-based methods. The implementation is available at: https://github.com/VimGeoTeam/VimGeo.

1 Introduction

Cross-view geo-localization determines the geographic location of a ground-level image (query image) by retrieving the most visually similar geo-tagged aerial image (reference image) and utilizing its location tag. This GPS-free task is crucial in GPS-denied environments like urban canyons and finds applications in autonomous driving [Fervers *et al.*, 2023], robot navigation [Fervers *et al.*, 2023], and UAV [Zheng *et al.*, 2020]. Challenges such as perspective and appearance differences between ground and aerial views necessitate advanced methods to bridge these gaps effectively.

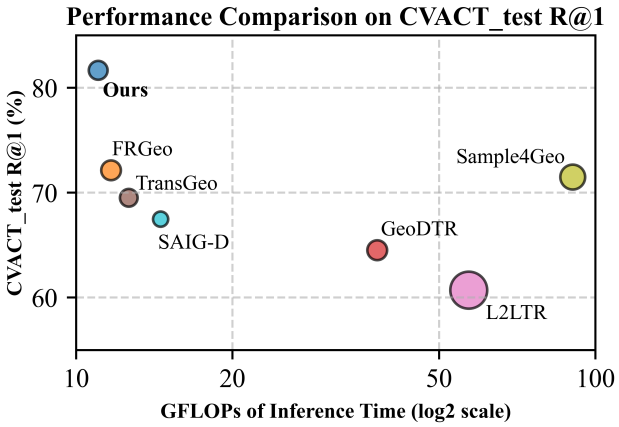

Figure 1: Comparison of R@1 accuracy on CVACT_test. The bubble size represents the number of trainable parameters. Our proposed *VimGeo* method, based on the ViM architecture, integrates the dynamic weight mechanism and CGP. This approach achieves competitive R@1 accuracy while significantly reducing trainable parameters and GFLOPs.

Recently, convolution-based methods have shown impressive performance in cross-view image-matching tasks due to their lightweight nature [Zhang and Zhu, 2024; Deuser *et al.*, 2023]. Meanwhile, Transformer-based architectures, with their powerful global modeling capabilities and self-attention mechanisms, have gained wide adoption in cross-view geo-localization tasks [Yang *et al.*, 2021; Zhu *et al.*, 2022; Zhang *et al.*, 2023]. Despite their success, the recently proposed Vision Mamba (Vim) [Zhu *et al.*, 2024] offers distinct advantages for cross-view geo-localization. By avoiding the high computational cost of self-attention mechanisms, Vim achieves sub-quadratic time complexity and linear memory requirements. This results in faster inference and significantly lower GPU memory usage, particularly for high-resolution images. Compared to methods such as Sample4Geo and L2LTR (see Figure 1), VimGeo achieves a superior trade-off

between computational efficiency and retrieval accuracy.

Furthermore, Vim employs a pure SSM architecture to model input images as sequential data without relying on 2D convolution priors. This design provides robust visual representations, making it suitable for diverse cross-view scenarios. Notably, VimGeo surpasses sequence modeling-based methods, highlighting its capability to generalize across complex geometric layouts (Figure 1).

Based on the unique advantages of Vim, we have developed a cross-view geo-localization system architecture, *VimGeo*, which leverages its strengths to achieve state-of-the-art performance.

Aside from the choice of feature extraction architecture for cross-view geo-localization, one of the primary challenges is the significant viewpoint disparity between ground-view and aerial-view images. The distinct imaging modalities of these two views result in considerable differences in appearance, resolution, and perspective, making geometric alignment between images highly challenging. Traditional methods have employed various feature extraction strategies to address this issue.

[Zhu *et al.*, 2022] proposed *TransGeo*, a Transformer-based framework for cross-view image geo-localization. This framework extracts features using a classification head combined with fully connected layers. However, when extracting features of higher complexity, the fully connected layers in this approach significantly increase the number of parameters, resulting in higher model complexity and longer inference times. [Zhang and Zhu, 2024] introduced a *Feature Recombination Module (FRM)* to align geometric spatial layouts between views. Their approach involves dividing feature maps into regions and applying spatial average pooling. While this method is effective for CNN-based architectures with downsampling, it divides each channel into four parts for pooling, which can lead to excessive feature loss when applied to Transformer-like modules.

In contrast to traditional strategies, our method leverages global feature compression rather than relying solely on a classification head. Specifically, we remove the classification head and the final fully connected layer. Directly pooling across all global feature channels can result in significant feature loss. To address this, we propose the *Channel Group Pooling (CGP)* module.

CGP partitions feature maps into distinct channel groups as shown in Figure 2(b). Within each group, average pooling is applied, and the resulting features are merged into a compact representation. By adjusting the size of each channel group, CGP effectively controls the output feature dimensions, similar to a fully connected layer. However, unlike fully connected layers, CGP achieves this without increasing the model's parameter count or inference time, ensuring greater efficiency and scalability.

We further investigated the loss function, which is one of the critical components in cross-view geo-localization tasks. [Zhu *et al.*, 2022] normalized the embedded output features using L_2 normalization and employed the soft-margin triplet loss as the training objective. [Zhang and Zhu, 2024] introduced a novel weighted $(B+1)$-tuple loss (WBL) as the optimization objective, which allows joint comparison of multiple negative samples by incorporating a weighting factor α. The proposed WBL significantly improved convergence speed and final performance.

However, their designed loss function overlooks the impact of sample difficulty on the optimization results. To address this issue, we propose a novel dynamic weight-based triplet loss, *Dynamic Weighted Batch-tuple Loss (DWBL)*, as the optimization objective. This loss function enables the model to pay more attention to the difficulty of learning samples. The improved loss function effectively enhances the final performance and demonstrates superior optimization capabilities for cross-view geo-localization tasks.

Our work focuses on the following key innovations:

- **Optimized Vim Architecture for Cross-View Feature Matching.** We adapted the Vim framework by removing its classification head and incorporating *Channel Group Pooling (CGP)* for global feature fusion. This design enhances the model's ability to capture global image information, improving localization accuracy while reducing parameters for greater efficiency.
- **Dynamic Weight Mechanism for Contrastive Learning.** We proposed a novel loss function, *Dynamic Weighted Batch-tuple Loss (DWBL)*, which dynamically adjusts negative sample weights to better handle hard and easy negatives. By balancing the influence of negatives with varying difficulty, DWBL enhances the robustness of optimization and improves the model's ability to achieve consistent performance across tasks.
- **Superior Computational Efficiency and Performance.** VimGeo achieves competitive or state-of-the-art results on CVUSA, CVACT, and VIGOR datasets while maintaining lower computational complexity, highlighting its advantages in both efficiency and accuracy.

2 Related Work

We preliminary investigated the existing cross-view geo-localization methods, focusing on feature extraction architectures and feature optimization functions.

2.1 Feature Extraction Architectures for Cross-View Geo-Localization

The choice of feature extraction architectures plays a pivotal role in cross-view geo-localization. [Workman *et al.*, 2015] first introduced CNNs into cross-view matching tasks, inspired by the remarkable success of CNNs in computer vision tasks [Krizhevsky *et al.*, 2012]. Subsequently, [Hu *et al.*, 2018] combined NetVlad [Arandjelovic *et al.*, 2016] and a dual-branch VGG architecture [Simonyan, 2014] to achieve viewpoint-invariant image representations. [Shi *et al.*, 2019] proposed the SAFA method, which aligns aerial images geometrically through polar coordinate transformations based on the VGG architecture [Simonyan, 2014]. Building on this, [Shi *et al.*, 2020] introduced DSM, which incorporates a sliding window approach for geolocating ground images with limited fields of view, employing a two-stream convolutional network architecture [Simonyan and Zisserman, 2014].

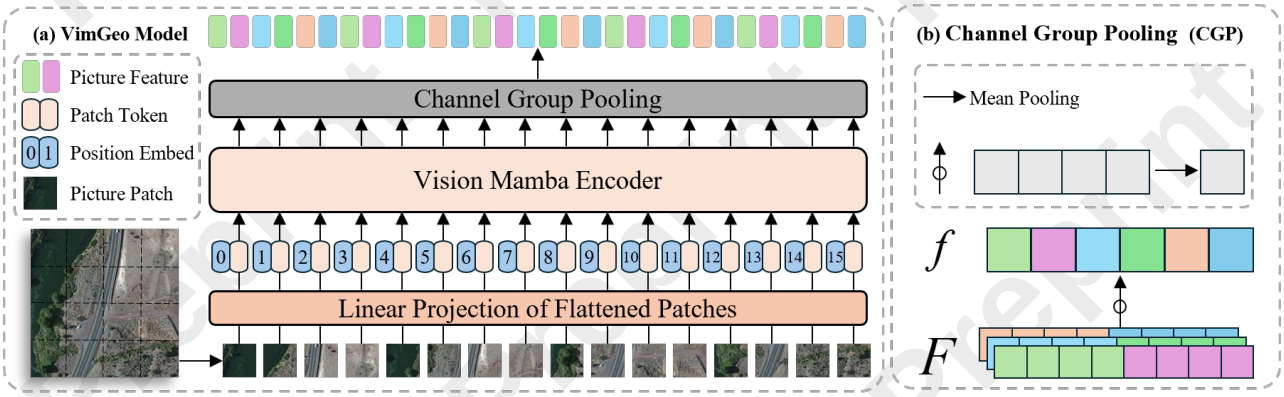

Figure 2: (a) Architecture of the proposed VimGeo model. (b) Visualization of the *Channel Group Pooling (CGP)* module.

The CDE method, proposed by [Toker *et al.*, 2021], integrates GANs [Goodfellow *et al.*, 2014] with SAFA for both geo-localization and ground image synthesis. Meanwhile, L2LTR [Yang *et al.*, 2021], TransGeo [Zhu *et al.*, 2022], and GeoDTR [Zhang *et al.*, 2023] all build upon the Vision Transformer (ViT) architecture [Dosovitskiy, 2020], with L2LTR leveraging a hybrid design, TransGeo adopting a pure ViT-based framework, and GeoDTR further introducing geometric layout descriptors and semantic data augmentation to enhance cross-view matching performance, especially on cross-area benchmarks.

Furthermore, FRGeo [Zhang and Zhu, 2024] and Sample4Geo [Deuser *et al.*, 2023] explicitly rely on the convolutional techniques proposed by [Liu *et al.*, 2022]. These methods mitigate geometric misalignment issues through explicit geometric spatial layout alignment and data augmentation, highlighting the significance of convolutional methods in cross-view geo-localization tasks.

2.2 Feature Optimization Functions for Cross-View Geo-Localization

Feature optimization functions are crucial in cross-view geo-localization tasks, focusing on improving the discriminative power of extracted features and facilitating effective learning. [Workman *et al.*, 2015] pioneered the use of Euclidean Loss for cross-view matching, directly measuring the similarity between image features. Building on this, [Hu *et al.*, 2018] introduced the Weighted Soft-Margin Triplet Loss [Arandjelovic *et al.*, 2016], incorporating a weighting factor α to enhance convergence speed. They further extended this to the Weighted Quadruplet Loss, improving the distinction between positive and negative samples while avoiding the reliance on predefined margins.

[Toker *et al.*, 2021] proposed the CDE method, introducing a Retrieval Loss that minimizes the Euclidean distance of projected features for improved matching accuracy. In addition, [Yang *et al.*, 2021] and [Zhu *et al.*, 2022] employed the Soft-Margin Triplet Loss in their L2LTR and TransGeo frameworks, respectively. TransGeo further leveraged bidirectional information flow to enhance cross-view matching. [Zhang *et al.*, 2023] introduced GeoDTR, which combines Counterfactual Loss with Weighted Soft-Margin Triplet Loss, offering improved robustness by contrasting hypothetical descriptors and optimizing geometric layouts.

To address the challenge of utilizing diverse negative samples, [Zhang and Zhu, 2024] proposed a Weighted (B+1)-Tuple Loss in the FRGeo method, which effectively leverages all negative samples within a mini-batch to accelerate convergence and improve final performance. Furthermore, [Deuser *et al.*, 2023] introduced Sample4Geo, utilizing Symmetric InfoNCE Loss to maximize the information derived from mini-batch negative samples while enhancing feature alignment through bidirectional optimization.

3 Methodology

3.1 Problem Formulation

We consider a dataset consisting of ground-aerial image pairs, denoted as $\{(I_i^g, I_i^a)\}_N$, where I_i^g and I_i^a are ground and aerial images, respectively, and N is the total number of pairs. Each pair corresponds to a unique geo-location. Geo-tags are available only for the aerial images $\{I_i^a\}_N$.

The objective of cross-view geo-localization is to find the aerial image I_r^a that corresponds to a given query ground image I_q^g, where $q, r \in \{1, 2, \ldots, N\}$. This match determines the geo-location of the query image I_q^g.

For this dataset, we extract feature representations $\{(f_i^g, f_i^a)\}_N$, designed such that the distance between matched pairs is smaller than the distance to any unmatched pair. This condition is mathematically represented as:

$$d(f_q^g, f_q^a) < \{d(f_q^g, f_i^a) \mid \forall i \in \{1, \ldots, N\}, i \neq q\}, \quad (1)$$

where $d(\cdot, \cdot)$ represents the L_2 distance.

Accordingly, the geo-localization task can be defined as finding:

$$r = \arg \min_{i \in \{1, \ldots, N\}} d(f_q^g, f_i^a). \quad (2)$$

A retrieval is correct if $r = q$.

3.2 Model Overview

The proposed model, *VimGeo*, addresses the challenges of cross-view geo-localization by leveraging the Vision Mamba (ViM) [Zhu *et al.*, 2024] architecture and incorporating innovative modules for feature extraction and optimization. As

illustrated in Figure 2(a), VimGeo adopts a dual-branch network structure consisting of two branches: the ground-view branch and the aerial-view branch. Each branch independently processes a given ground-aerial image pair (I_g, I_a).

At the first stage, input images are divided into patches and flattened, followed by a linear projection step that transforms the patches into sequences of feature vectors. These projected patches are then passed into the Vision Mamba Encoder, which captures global contextual information within the images. Unlike traditional methods that rely heavily on classification heads and fully connected layers, VimGeo eliminates these computationally intensive components, opting instead for a more efficient global feature integration mechanism.

To mitigate the issue of excessive feature loss during pooling, the *Channel Group Pooling (CGP)* module, as depicted in Figure 2(b), is introduced. CGP divides the feature map into adjustable channel groups, applies average pooling within each group, and subsequently flattens and combines the pooled features. This design not only preserves critical global information but also provides flexibility in adjusting the output feature dimensions, similar to fully connected layers, without introducing additional parameters or computational overhead.

To further enhance cross-view matching, the model incorporates the *Dynamic Weighted Batch-tuple Loss (DWBL)* as its optimization objective, shown in Figure 3. The DWBL process computes a similarity matrix between query and reference features, dynamically assigning weights to negative samples based on their proximity to the anchor. This mechanism ensures effective utilization of all negative samples, significantly improving optimization and robustness in cross-view geo-localization tasks.

The final feature representations $f^g \in \mathbb{R}^{L \cdot \frac{C}{G}}$ and $f^a \in \mathbb{R}^{L \cdot \frac{C}{G}}$ are optimized through the DWBL mechanism, which dynamically balances the difficulty of learning from negative samples, achieving enhanced performance on benchmark datasets.

3.3 Channel Group Pooling (CGP)

The *Channel Group Pooling (CGP)* mechanism takes the raw features F^g and F^a, produced by the backbone, to compute the final representations f^g and f^a. The input features have dimensions $[L, C]$, where L denotes the sequence length (i.e., the number of spatial patches or tokens), and C indicates the number of channels.

As part of the preprocessing, the satellite image features are first mapped to the same spatial regions as the ground image features, following the mapping procedure used in the *Feature Recombination Module (FRM)* from FRGeo [Zhang and Zhu, 2024]. Specifically, the satellite image features F^a are divided into 4 regions according to the spatial division method, similar to the ground image. Each of these regions is processed independently to align the feature representations between the two views.

Next, the channel dimension C is divided into groups of size G. The feature tensor F is then reshaped to group the channels, and mean pooling is applied along the last dimension (group size):

$$F \in \mathbb{R}^{L \times \frac{C}{G} \times G} \xrightarrow{\text{mean}(x, \text{dim}=-1)} F' \in \mathbb{R}^{L \times \frac{C}{G}}, \tag{3}$$

which reduces the tensor's size to $[L, \frac{C}{G}]$.

Finally, the tensor is flattened, combining the sequence and channel dimensions:

$$F' \in \mathbb{R}^{L \times (\frac{C}{G})} \rightarrow f \in \mathbb{R}^{L \cdot \frac{C}{G}}. \tag{4}$$

This process produces the final feature representations f^g and f^a for the ground and aerial images, respectively:

$$f^g = \text{CGP}(F^g), \quad f^a = \text{CGP}(F^a). \tag{5}$$

For a visual illustration of how the CGP module works, refer to part (b) of Figure 2, which demonstrates the effective application of CGP in processing the feature maps.

3.4 Optimization Objective

In previous works [Hu *et al.*, 2018; Shi *et al.*, 2019; Yang *et al.*, 2021; Zhu *et al.*, 2022], the weighted soft-margin ranking loss has been widely used. This loss is computed by constructing triplets within each mini-batch, focusing on only one negative sample during each update. While effective to some extent, this approach severely limits the utilization of information from other negative samples in the mini-batch, resulting in prolonged training times and suboptimal accuracy.

Building on the weighted (B+1)-tuple loss (WBL) proposed in [Zhang and Zhu, 2024], which constructs (B+1)-tuples by pushing the anchor sample away from all other negative samples in the mini-batch, we extend this concept further to design a more efficient loss function. The WBL framework leverages the relationship among multiple negative samples in a batch, providing a stronger optimization foundation. However, its static weighting of negative samples may limit its potential in handling harder negatives effectively.

To address this, we propose the *Dynamic Weighted Batch-tuple Loss (DWBL)* (Figure 3), which not only constructs batch-tuples but also introduces a dynamic weighting mechanism. Unlike traditional WBL, DWBL dynamically adjusts the importance of each negative sample based on its similarity to the anchor sample. By assigning higher weights to harder negatives, DWBL ensures that these challenging samples are emphasized and effectively learned. As iterations progress, DWBL ensures that negatives of varying difficulty gradually move farther from the positive sample, with their distances converging to a similar range. This mechanism guarantees the separation of all negatives from the positive sample, leading to improved accuracy and robust optimization.

To formalize this approach, we present the mathematical formulation of DWBL. Given a set of mini-batch image pairs $\{(I_i^g, I_i^a)\}_B$, their corresponding feature representations are denoted as $\{(f_i^g, f_i^a)\}_B$, where B is the number of image pairs in the mini-batch. Since each pair constitutes one training instance, we denote the number of instances in the mini-batch as N, and in our case, $N = B$. When f_i^g is selected as the anchor sample, f_i^a serves as the positive sample, while the set of negative samples is $\{f_j^a\}_{j \neq i}$. Within each mini-batch, the DWBL is defined as:

$$L = \frac{1}{2}\Big(L_{\text{s}}(f^g, f^a) + L_{\text{s}}(f^a, f^g)\Big), \tag{6}$$

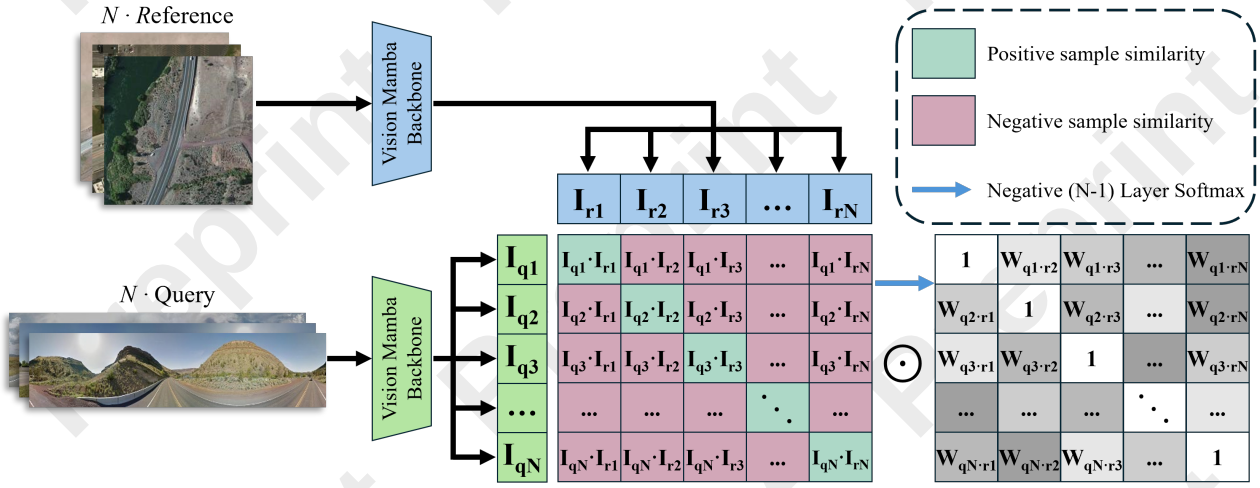

Figure 3: Illustration of the proposed DWBL process. The similarity matrix between query and reference features is computed, and dynamic weights are assigned to negative samples based on their proximity to the anchor. This mechanism ensures the effective utilization of all negative samples, leading to improved optimization.

where the single-direction loss L_{s} is computed as:

$$L_{\mathrm{s}}(f^q, f^k) = \frac{1}{N}\sum_{i=1}^{N}\log\left(1+\sum_{j\neq i}\exp\left(\alpha(s_{ij}^{\mathrm{neg}} - s_i^{\mathrm{pos}})\right)\right). \quad (7)$$

Here, s_i^{pos} represents the similarity between the anchor f_i^q and its corresponding positive f_i^k, while s_{ij}^{neg} denotes the similarity between the anchor f_i^q and a negative sample f_j^k. The weights for negative samples are dynamically adjusted using a softmax function over the similarities. To ensure that the total contribution of negative samples aligns with their original scale, the weights are scaled by the number of negative samples, $N-1$:

$$w_{ij}^{\mathrm{neg}} = (N-1)\frac{\exp(s_{ij}^{\mathrm{neg}})}{\sum_{j\neq i}\exp(s_{ij}^{\mathrm{neg}})}. \quad (8)$$

The dynamic weighting mechanism adaptively prioritizes hard negatives (those with subtle visual differences from the anchor) by amplifying their gradient contributions. The parameter α precisely controls this reweighting process to capture fine-grained similarity variations.

DWBL specifically optimizes for fine-grained retrieval by reshaping the loss landscape: it strengthens gradients for challenging cases with marginal visual differences while reducing emphasis on obviously dissimilar samples. This approach is particularly effective for discriminating between highly similar candidates where conventional losses fail to capture subtle discriminative features. Our experiments demonstrate superior performance in fine-grained cross-view matching tasks.

4 Experiment

4.1 Datasets and Experimental Settings

Datasets. Our method is evaluated on three widely-used cross-view geo-localization datasets: CVUSA [Zhai *et al.*, 2017], CVACT [Liu and Li, 2019], and VIGOR [Zhu *et al.*, 2021]. CVUSA consists of 35,532 training pairs and 8,884 testing pairs, primarily featuring suburban landscapes. CVACT includes 35,532 training pairs, 8,884 validation pairs (CVACT_val), and 92,802 testing pairs (CVACT_test), focusing on urban regions in Canberra for city-scale geo-localization. In contrast, VIGOR contains 105,214 ground images and 90,618 aerial images, allowing query ground images to originate from arbitrary locations within the target area. It employs the Same-area and Cross-area protocols for evaluation under the standard setup.

Evaluation Metrics. Following prior works [Hu *et al.*, 2018; Liu and Li, 2019; Shi *et al.*, 2019; Shi *et al.*, 2020], we adopt R@K (K = {1, 5, 10, 1%}) as the primary metric to evaluate performance. This metric measures the likelihood of correct matches within the top-K retrieved results. Additionally, for the VIGOR dataset, we report the hit rate, which quantifies the probability that the top-1 aerial image contains the query ground image's location. Together, these metrics provide a comprehensive evaluation of both standard and fine-grained geo-localization tasks.

Implementation Details. Our model is built upon the Vision Mamba (Vim) architecture [Zhu *et al.*, 2024], a state-of-the-art bidirectional state space model designed for efficient visual representation learning. Vim replaces the conventional attention-based mechanism with a pure-SSM-based backbone, offering subquadratic-time computation and linear memory complexity while retaining the capability for global context modeling and positional awareness. The model is initialized with pretrained parameters on ImageNet-1K [Deng *et al.*, 2009].

The model is trained using 4 NVIDIA A6000 GPUs. We utilize the AdamW optimizer [Loshchilov, 2017], with the hyperparameter α in the single-direction loss function (Equation 7) set to 10, ensuring optimal convergence. Training experiments were conducted on high-resolution image datasets,

leveraging the computational efficiency and memory savings of the Vim architecture, which outperforms Transformer-based alternatives in terms of speed and memory consumption.

4.2 Comparison with State-of-the-art Methods

Our proposed VimGeo model is evaluated against leading methods on three benchmark datasets: CVUSA, CVACT, and VIGOR. The experiments assess its performance across various cross-view geo-localization scenarios, including standard, fine-grained, and beyond one-to-one tasks.

Table 1 and Table 2 compare our VimGeo model with state-of-the-art methods on the CVUSA and CVACT datasets under various settings. The results demonstrate that VimGeo achieves competitive performance across all metrics. On the CVUSA dataset, VimGeo achieves an R@1 of **96.19%**, showcasing its effectiveness in cross-view geo-localization tasks. While Sample4Geo achieves higher metrics across all ranks (R@1, R@5, R@10, and R@1%), VimGeo maintains strong performance as one of the top sequence modeling-based methods, as shown in Table 1.

On the CVACT_val dataset, VimGeo achieves an R@1 of 87.62%, which is competitive among sequence modeling-based methods, although slightly lower than Sample4Geo (<u>90.81%</u>). These results highlight VimGeo's balanced performance across various metrics. Furthermore, on the CVACT_test dataset, VimGeo significantly outperforms all other methods, achieving an R@1 of **81.69%**, which is **9.54%** higher than the next best-performing method, FRGeo.

Table 3 summarizes the results on the VIGOR dataset under both Same-area and Cross-area protocols. Initial experiments showed that the *Dynamic Weighted Batch-tuple Loss (DWBL)* was less effective for VIGOR, likely due to its unique challenges, prompting the use of *Weighted (B+1)-tuple Loss (WBL)* instead. In the Same-area setting, Ours (WBL) achieves competitive performance among sequence modeling-based methods (R@1: 55.24%), while in the Cross-area setting, it maintains adaptability with an R@1 of **19.31%**, comparable to TransGeo.

Method	R@1	R@5	R@10	R@1%
SAFA†	89.84%	96.93%	98.14%	99.64%
CDE†	92.56%	97.55%	98.33%	99.57%
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†	95.43%	**98.86%**	**99.34%**	**99.86%**
FRGeo	97.06%	99.25%	99.47%	99.85%
Sample4Geo	<u>98.68%</u>	<u>99.68%</u>	<u>99.78%</u>	<u>99.87%</u>
Ours	**96.19%**	98.62%	99.00%	99.52%

Table 1: Comparisons between VimGeo (Ours) and state-of-the-art methods on the CVUSA dataset. † indicates applying polar transform to aerial images. For spatial modeling-based methods, the highest values are underlined. For sequence modeling-based methods, the highest values are highlighted in bold. This notation is consistently applied to the subsequent tables as well.

Method	R@1	R@5	R@10	R@1%
		CVACT_val		
SAFA†	81.03%	92.80%	94.84%	98.17%
DSM†	82.49%	92.44%	93.99%	97.32%
CDE†	83.28%	93.57%	95.42%	98.22%
L2LTR†	84.89%	94.59%	95.96%	98.37%
TransGeo	84.95%	94.14%	95.78%	98.37%
SEH†	84.75%	93.97%	95.46%	98.11%
GeoDTR†	86.21%	**95.44%**	**96.72%**	**98.77%**
FRGeo	90.35%	96.45%	97.25%	98.74%
Sample4Geo	<u>90.81%</u>	<u>96.74%</u>	<u>97.48%</u>	<u>98.77%</u>
Ours	**87.62%**	94.88%	96.06%	98.06%
		CVACT_test		
SAFA†	55.50%	79.94%	85.08%	94.49%
DSM†	35.63%	60.07%	69.10%	84.75%
CDE†	61.29%	85.13%	89.14%	98.32%
L2LTR†	60.72%	85.85%	89.88%	96.12%
GeoDTR†	64.52%	88.59%	91.96%	**98.74%**
FRGeo	<u>72.15%</u>	91.93%	94.05%	98.66%
Sample4Geo	71.51%	<u>92.42%</u>	<u>94.45%</u>	<u>98.70%</u>
Ours	**81.69%**	**92.42%**	**94.32%**	97.19%

Table 2: Comparison between VimGeo (Ours) and state-of-the-art methods on the CVACT dataset under CVACT_val and CVACT_test settings.

These results collectively demonstrate VimGeo's capability to balance performance across various datasets and protocols, establishing it as a robust and adaptable model for cross-view geo-localization tasks.

4.3 Computational Costs

Figure 1 compares our proposed VimGeo model with six state-of-the-art models on the CVACT_test dataset in terms of computational complexity (GFLOPs), trainable parameters, and retrieval accuracy. VimGeo achieves the best computational complexity (11.023 GFLOPs) and requires only 50.868M trainable parameters, while achieving a retrieval accuracy of R@1: **81.69%**, significantly outperforming all competing methods. In contrast, models like GeoDTR and L2LTR incur much higher computational costs but fail to match VimGeo's accuracy, and Sample4Geo, despite achieving moderate accuracy, does so at the expense of significantly higher computational complexity (90.414 GFLOPs). Meanwhile, methods such as FRGeo and TransGeo exhibit relatively lower computational demands but cannot achieve the accuracy levels of VimGeo. These results highlight VimGeo's ability to balance computational efficiency and performance, making it a scalable and effective solution for cross-view geo-localization, particularly in resource-constrained scenarios.

4.4 Ablation Study

To evaluate the effectiveness of the proposed components, including *Channel Group Pooling (CGP)* and *Dynamic Weighted Batch-tuple Loss (DWBL)*, we conducted a series of ablation experiments by sequentially integrating these components into the Baseline model. The Baseline model adopts

Method	R@1	R@5	R@10	R@1%	Hit
Same-area Protocol					
Siamese-VGG	18.69%	43.64%	55.36%	97.55%	21.90%
SAFA	33.93%	58.42%	68.12%	98.24%	36.87%
SAFA+Mining	38.02%	62.87%	71.12%	97.63%	41.81%
VIGOR	41.07%	65.81%	74.05%	98.37%	44.71%
TransGeo	**61.48%**	**87.54%**	**91.88%**	**99.56%**	**73.09%**
FRGeo	71.26%	91.38%	94.32%	99.52%	82.41%
Sample4Geo	77.86%	95.66%	97.21%	99.61%	89.82%
Ours	37.84%	66.63%	75.13%	96.81%	40.81%
Ours (WBL)	55.24%	80.75%	76.12%	97.30%	57.43%
Cross-area Protocol					
Siamese-VGG	2.77%	8.61%	12.94%	62.64%	3.16%
SAFA	8.20%	19.59%	26.36%	77.61%	8.85%
SAFA+Mining	9.23%	21.12%	28.02%	77.84%	9.92%
VIGOR	11.00%	23.56%	30.76%	80.22%	11.64%
TransGeo	18.99%	**38.24%**	**46.91%**	**88.94%**	**21.21%**
FRGeo	37.54%	59.58%	67.34%	94.28%	40.66%
Sample4Geo	61.70%	83.50%	88.00%	98.17%	69.87%
Ours	8.17%	19.46%	26.47%	77.33%	8.64%
Ours (WBL)	**19.31%**	37.50%	46.03%	86.96%	20.72%

Table 3: Comparison between VimGeo (Ours) and state-of-the-art methods on the VIGOR dataset under Same-area and Cross-area protocols. Testing revealed that DWBL is unsuitable for this dataset's task, and thus the loss function was replaced.

Method	R@1	R@5	R@10	R@1%
Base	94.10%	98.39%	98.93%	99.68%
Base + CGP	95.36%	98.56%	**99.11%**	**99.71%**
Base + DWBL	94.17%	98.28%	98.86%	99.67%
Base + CGP + DWBL	**96.19%**	**98.62%**	99.00%	99.52%

Table 4: Performance comparison on the CVUSA dataset. The proposed model with CGP and DWBL achieves the best performance across all metrics.

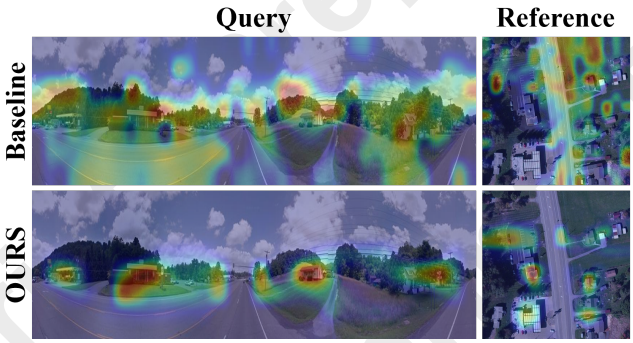

Figure 4: Comparison of heatmaps generated by the VimGeo model on street and satellite view images.

a dual-branch architecture with Vim [Zhu *et al.*, 2024] as the backbone. To ensure a fair comparison, all configurations were trained under consistent hyperparameters and strategies. The results, summarized in Table 1, show the contribution of each component to model performance.

The experiments started with the Baseline model, which includes the Vim backbone without additional components. Adding the CGP module enhances the model's ability to capture global image representations, improving feature integration and localization accuracy. Integrating the DWBL loss further optimizes the model by effectively utilizing negative samples during training, resulting in improved optimization. Finally, combining CGP and DWBL into the Baseline model forms our complete VimGeo architecture, which achieves the best performance in terms of R@1 and R@5, demonstrating the complementary nature and effectiveness of these components.

As shown in Table 4, both CGP and DWBL independently contribute to notable performance improvements. The integration of both components (Baseline + CGP + DWBL) achieves the best performance, demonstrating their complementary nature and highlighting the effectiveness of the proposed VimGeo architecture.

4.5 Visualization Analysis

To better understand VimGeo's learning process and the differences in attention regions across models, we visualize heatmaps, as shown in Figure 4. The Baseline model exhibits scattered and inconsistent attention, focusing on non-sky structures in a disorganized manner, while VimGeo demonstrates a more stable and focused attention distribution by emphasizing prominent buildings alongside road information. These prominent buildings provide stable reference features for cross-view geo-localization, as they remain consistent in appearance across different ground-aerial image pairs. VimGeo's enhanced focus on these critical regions is attributed to the effectiveness of the *Channel Group Pooling (CGP)* mechanism and the *Dynamic Weighted Batch-tuple Loss (DWBL)* loss, which together enable the model to align spatial layouts across views and prioritize regions essential for accurate localization.

5 Conclusion

In this paper, we propose VimGeo, an innovative and efficient cross-view geo-localization method designed to address geometric spatial misalignments between cross-view images. Through the introduction of the *Channel Group Pooling (CGP)* mechanism, our approach effectively mitigates ambiguities and improves the extraction of discriminative localization features. Additionally, the *Dynamic Weighted Batch-tuple Loss (DWBL)* enhances optimization by strategically focusing on hard negatives to refine discriminative power, prioritizing top-1 matching accuracy.

Comprehensive experiments demonstrate that VimGeo achieves superior performance across the CVUSA, CVACT, and VIGOR datasets. Furthermore, it offers significant advantages in computational efficiency and model scalability, highlighting its practicality and suitability for cross-view geo-localization tasks.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants U22A2054,

62201311, and 62401213, in part by the Peng Cheng Laboratory Major Key Project under Grants PCL2023AS1-5 and PCL2024A01, in part by the Young Elite Scientists Sponsorship Program by CAST under Grant 2023QNRC001, and in part by the Young Innovative Talents Program for Universities of Guangdong Province under Grant 2024KQNCX062.

References

[Arandjelovic *et al.*, 2016] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5307, 2016.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[Deuser *et al.*, 2023] Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4Geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16847–16856, 2023.

[Dosovitskiy, 2020] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Fervers *et al.*, 2023] Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, and Rainer Stiefelhagen. Uncertainty-aware vision-based metric cross-view geolocalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21621–21631, 2023.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

[Hu *et al.*, 2018] Sixing Hu, Mengdan Feng, Rang M.H. Nguyen, and Gim Hee Lee. 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*, pages 7258–7267, 2018.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

[Liu and Li, 2019] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5624–5633, 2019.

[Liu *et al.*, 2022] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[Loshchilov, 2017] Ilya Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Shi *et al.*, 2019] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32, 2019.

[Shi *et al.*, 2020] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. 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*, pages 4064–4072, 2020.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems*, 27, 2014.

[Simonyan, 2014] Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Toker *et al.*, 2021] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. 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*, pages 6488–6497, 2021.

[Workman *et al.*, 2015] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, December 2015.

[Yang *et al.*, 2021] Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. *Advances in Neural Information Processing Systems*, 34:29009–29020, 2021.

[Zhai *et al.*, 2017] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 867–875, 2017.

[Zhang and Zhu, 2024] Qingwang Zhang and Yingying Zhu. Aligning geometric spatial layout in cross-view geo-localization via feature recombination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7251–7259, 2024.

[Zhang *et al.*, 2023] Xiaohan Zhang, Xingyu Li, Waqas Sultani, Yi Zhou, and Safwan Wshah. Cross-view geo-localization via learning disentangled geometric layout correspondence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3480–3488, 2023.

[Zheng *et al.*, 2020] Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1395–1403, 2020.

[Zhu *et al.*, 2021] Sijie Zhu, Taojiannan Yang, and Chen Chen. VIGOR: Cross-view image geo-localization beyond

one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2021.

[Zhu *et al.*, 2022] Sijie Zhu, Mubarak Shah, and Chen Chen. 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*, pages 1162–1171, 2022.

[Zhu *et al.*, 2024] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.