CVGlobal

Leonardo Ruso
leonardo.russo@inria.fr
Diego Marcos
diego.marcos@inria.fr

INRIA
Evergreen Team
Montpellier, France

Abstract

Cross-view geo-localization and orientation estimation require models that generalize across diverse geographies. Existing evaluations are often geographically narrow, limiting conclusions about robustness and fairness. We address this by introducing CVGlobal, a geographically balanced test benchmark spanning five continents with paired aerial and street-view imagery across urban and rural contexts. CVGlobal enables geographically fair assessment and facilitates analysis of region-specific behavior.

Alongside the benchmark, we present a training-free, zero-shot baseline for cross-view orientation estimation built on vision foundation models. The method aggregates tokens along vertical (ground) and radial (aerial) directions and incorporates lightweight geometric priors (sky masking and depth cues) to encourage alignment without task-specific training. We evaluate on CVUSA and analyze results on CVGlobal. While the zero-shot baseline underperforms supervised methods in absolute accuracy, it provides meaningful signals, transparent failure modes, and a portable, low-cost reference for ablations and diagnostics.

Our study surfaces generalization gaps across regions and highlights opportunities to combine strong priors with light supervision. We release CVGlobal to catalyze geographically fair evaluation and provide a simple zero-shot baseline to serve as a reproducible yardstick for future supervised and semi-supervised approaches.

1 Introduction

Cross-view geo-localization and orientation estimation underpin applications in robotics, mapping, and outdoor AR, where systems must align ground-view observations with overhead imagery across diverse environments. Despite rapid progress, most evaluations remain geographically narrow (e.g., US-centric), raising concerns about generalization and fairness. We argue that advancing the field requires both: (i) geographically balanced benchmarks, and (ii) strong, training-light baselines that are easy to transfer and diagnose.

This work addresses both needs. First, we introduce CVGlobal, a globally balanced test benchmark spanning five continents with urban and rural regions. CVGlobal enables geographically fair assessment by design, reducing the risk that methods overfit to a single region's visual biases. Second, we present a simple zero-shot baseline for orientation estimation that exploits vision foundation models. Our method aggregates tokens along vertical (ground) and radial (aerial) directions, and uses lightweight geometric priors (sky masking and depth cues) to encourage alignment—all without task-specific training. While zero-shot

performance lags supervised methods in absolute accuracy, it provides a transparent, portable reference that helps stress-test assumptions and diagnose failure modes across regions.

We evaluate on CVUSA and report qualitative and quantitative analyses, then use CVGlobal to highlight geographic effects and characterize strengths and limitations of zero-shot orientation estimation. We find that the baseline produces meaningful signals in many scenes and offers a cost-effective yardstick for ablations and future improvements.

Our contributions are:

- CVGlobal: a geographically balanced cross-view test benchmark covering five continents with paired aerial and street-view imagery across urban and rural contexts, enabling fairer evaluation and analysis.
- A training-free, zero-shot baseline for cross-view orientation estimation built on ViT features, with interpretable vertical/radial token aggregation and simple geometric priors (sky and depth).
- An empirical study across CVUSA and CVGlobal that surfaces generalization gaps, clarifies when zero-shot signals are reliable, and establishes a portable baseline for future supervised or semi-supervised methods.

2 Related Works

Cross-view geo-localization addresses the fundamental challenge of matching images captured from drastically different viewpoints of the same geographic location. This task has evolved from traditional handcrafted approaches to sophisticated deep learning methods, driven by applications in augmented reality, robotics, among many other fields.

2.1 Cross-View Geo-localization and Orientation Estimation

Early pioneering works by Workman and Jacobs [8**?**] established the foundation for cross-view matching by demonstrating the potential of CNNs for learning feature representations across viewpoint variations. The introduction of benchmark datasets like CVUSA [8] and later VIGOR [9] catalyzed systematic research in this domain, with the latter providing more challenging same-area and cross-area evaluation protocols which better reflect real-world deployment scenarios.

Subsequent developments focused on addressing the inherent domain gap between aerial and ground imagery. Notable approaches include CVM-Net [5], which introduced dual-stream CNN architectures with polar transformations to align spatial layouts, and SAFA [**?**], which employed attention mechanisms for spatial-aware feature aggregation. Methods like CVFT [**?**] tackled the domain gap through feature transport modules, while others explored generative approaches to synthesize cross-view correspondences [7].

The recognition that orientation estimation is crucial for disambiguation and practical applications led to joint location and orientation frameworks. Recent works have emphasized the importance of spatial awareness in feature representations [**?**], particularly for applications requiring precise alignment such as outdoor augmented reality. However, many existing methods either treat orientation as a byproduct of location retrieval or require extensive supervised training with orientation labels.

The advent of Vision Transformers (ViTs) [2] has revolutionized the whole computer vision field by enabling models to capture global spatial relationships through self-attention mechanisms. Transformer-based approaches like L2LTR [**?**] and TransGeo [**?**] have demonstrated superior performance over CNN-based methods, leveraging learnable position encodings and global context modeling.

Concurrently, self-supervised learning has emerged as a powerful paradigm for learning visual representations without manual annotations. Methods such as MoCo [4], SimCLR [1], and BYOL [3] have shown that self-supervised features can match or exceed supervised counterparts on various downstream tasks. DINOv2 [6] represents the current state-of-the-art for visual transformers, providing features particularly well-suited for dense prediction tasks and fine-grained visual understanding.

Despite these advances, most transformer-based cross-view methods still rely heavily on supervised learning with orientation labels, requiring extensive annotation efforts. Recent works have begun exploring self-supervised features for geo-localization tasks, but typically treat cross-view matching as standard retrieval without considering the specific geometric constraints and orientation relationships inherent in cross-view scenarios.

2.2 Cross-Modality Datasets

Current benchmark datasets, while valuable, present limitations for comprehensive evaluation. CVUSA focuses primarily on the United States, while VIGOR, though more diverse, still covers a limited geographical scope. These datasets primarily support geo-localization and retrieval tasks, with growing interest in orientation estimation as a distinct but related problem. The lack of large-scale, geographically diverse datasets with comprehensive global coverage has hindered the development of truly robust cross-view methods that generalize across different environmental conditions and cultural contexts.

Our work addresses these limitations by introducing CVGlobal, a large-scale dataset with balanced global representation, and proposing a novel cross-view method that combines unsupervised features with multi-modal cues for orientation estimation.

3 Dataset Generation Method

In this section, we present our methodology for constructing CVGlobal, a multi-modal dataset that pairs satellite and street-view imagery across diverse global regions. Our approach systematically samples locations from five continents while ensuring geographical diversity and balanced representation between urban and rural environments.

3.1 Dataset Design and Sampling Strategy

Our dataset construction methodology is guided by two key principles: *geographical diversity* and *balanced representation*. We define sampling regions across five major continents (North America, Europe, Asia, South America, and Africa), with each continent contributing equally to the final dataset to prevent geographical bias.

For each continent, we establish two distinct sampling regions:

- **Urban regions**: Areas with high population density and significant urban infrastructure

Table 1: Geographical sampling regions defined for each continent and environment type.

Continent	Type	Location	Lat Range	Lon Range
North America	Urban	New York City	40.71°–40.81°N	74.01°–73.91°W
	Rural	California Farmland	36.78°–36.88°N	119.42°–119.32°W
Europe	Urban	Paris	48.86°–48.96°N	2.35°–2.45°E
	Rural	French Countryside	46.23°–46.33°N	2.21°–2.31°E
Asia	Urban	Tokyo	35.69°–35.79°N	139.69°–139.79°E
	Rural	Rural India (Agra)	27.18°–27.28°N	78.04°–78.14°E
South America	Urban	São Paulo	23.55°–23.45°S	46.63°–46.53°W
	Rural	Brazilian Rainforest	14.24°–14.13°S	51.93°–51.83°W
Africa	Urban	Nairobi	1.29°–1.19°S	36.82°–36.92°E
	Rural	Kenyan Savanna	2.15°–2.05°S	37.31°–37.41°E

- **Rural regions**: Areas with low population density and predominantly natural or agricultural landscapes

The sampling regions are carefully selected to represent diverse climatic, cultural, and developmental contexts within each continent. Table 1 details the specific geographical boundaries for each region.

3.2 Data Acquisition Details

The CVGlobal dataset has been generated ensuring high-quality multi-modal data collection through random coordinate sampling. For each region, we generate random coordinates and validate them for Street View availability and outdoor environments using Google Places API filtering to exclude indoor locations such as shopping malls and restaurants.

For each validated coordinate, we acquire satellite imagery (640×640 pixels at zoom level 18) and street-view images from four cardinal directions (0°, 90°, 180°, 270°). The directional images are concatenated to create panoramic representations.

4 Crossview Method

We propose PROJECT_NAME, a novel approach for cross-view orientation estimation that leverages a ViT backbne for feature extraction with orientation-aware token aggregation strategies. Our method addresses the fundamental challenge of aligning ground-level images with aerial satellite views by exploiting both spatial structure and depth information through a flexible architecture that supports multiple vision foundation models.

4.1 Problem Formulation

Given a ground-level panoramic image I_g and an aerial satellite image I_a of the same geographic location, our goal is to estimate the relative orientation θ between the two views. The ground image is extracted from a 360° panorama using a field-of-view (FOV) window defined by parameters (f_x, f_y, ψ, ϕ), where f_x and f_y represent the horizontal and vertical

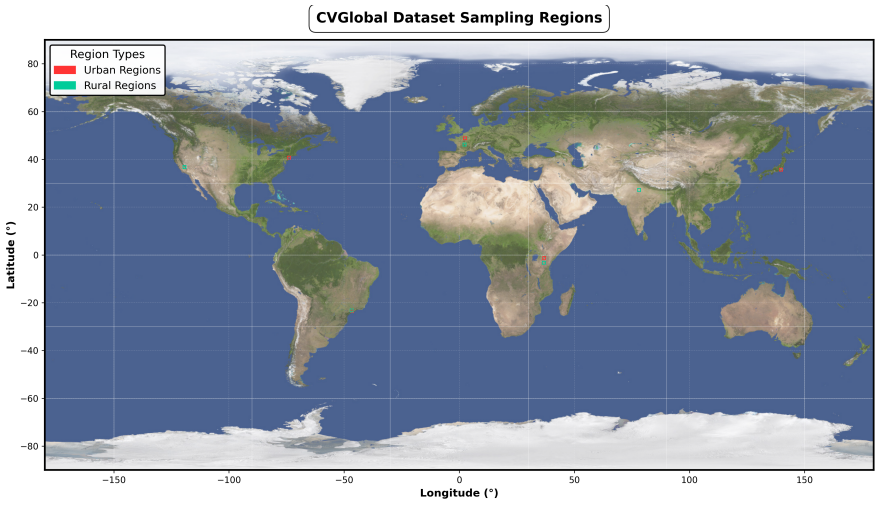

Figure 1: Global distribution of CVGlobal dataset sampling regions across five continents, showing urban (red) and rural (teal) areas with numbered locations corresponding to Table 1.

FOV angles, ψ is the yaw (rotation around the vertical axis), and ϕ is the pitch (elevation angle).

4.2 Feature Extraction

PROJECT_NAME employs a flexible architecture that can utilize different pre-trained ViT models as feature extractors. Our implementation supports multiple backbone architectures including DINOv2 and CLIP.

The feature extraction process is consistent for any backbone model used, and it will be treated as a black box for the scope of this work.

$$\mathbf{T}_g = \text{Backbone}(I_g) \tag{1}$$
$$\mathbf{T}_a = \text{Backbone}(I_a) \tag{2}$$

where $\mathbf{T}_g$, $\mathbf{T}_g \in \mathbb{R}^{(G,G,C)}$ are the ground and aerial tokens respectively, where G represents the grid size, and C represents the number of channels for each token. Their size changes on the backbone model used, depending on its patch size:

- **DINOv2** uses the pre-trained DINOv2-ViT-B/14 model with 14×14 pixel patches, producing a 16×16 token grid with 768-dimensional features.
- **CLIP** employs CLIP-ViT-Base-Patch16 with 16×16 pixel patches, generating a 14×14 token grid with 768-dimensional features.

4.3 Sky Filtering and Depth Estimation

To improve orientation estimation, we incorporate the following semantic and geometric priors to the ground and aerial images:

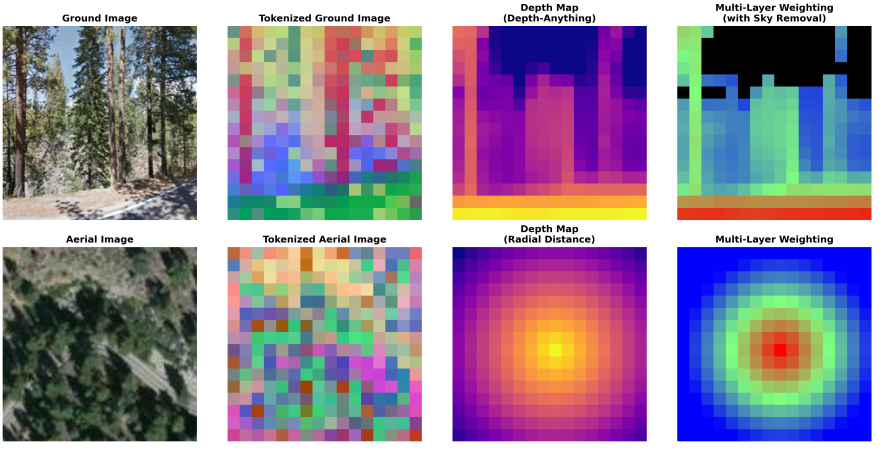

Figure 2: Visualization of tokenized depth information for aerial images, highlighting the spatial distribution of depth cues across different regions.

Sky Segmentation: We employ a lightweight CNN-based sky filter to identify and mask sky regions in ground images. The sky mask M_{sky} is computed at the patch level using majority voting within each grid cell, producing a binary mask $M_{grid} \in \{0,1\}^{G\times G}$ where 1 indicates ground and 0 indicates sky.

Depth Estimation: For ground imagery, we utilize the Depth-Anything model to generate depth maps $D^{(G)}_{grid}$ for each image. The depth information is downsampled to match the token grid, providing normalized depth values $d_{i,j} \in [0,1]$ for each spatial location (i,j).

For aerial imagery instead, we create a comparative depth map by using the radial distance from the center of the image to each pixel. Subsequently, the pixel values are averaged to compute the depth map $D^{(A)}_{grid}$ at the patch level.

4.4 Multi-Layer Depth-Weighted Token Aggregation

We introduce a novel aggregation strategy that separates tokens into three depth layers: *foreground, middleground, and background*. This approach captures the multi-scale nature of visual features in cross-view matching.

Vertical Column Analysis: For each vertical column j in the ground image feature grid, we compute depth-weighted averages over valid (non-sky) tokens:

$$\mathbf{t}_j^{fore} = \frac{\sum_{i:M_{grid}(i,j)=1} w_i^{fore} \cdot \mathbf{f}^g_{i,j}}{\sum_{i:M_{grid}(i,j)=1} w_i^{fore}} \tag{3}$$

$$\mathbf{t}_j^{mid} = \frac{\sum_{i:M_{grid}(i,j)=1} w_i^{mid} \cdot \mathbf{f}^g_{i,j}}{\sum_{i:M_{grid}(i,j)=1} w_i^{mid}} \tag{4}$$

$$\mathbf{t}_j^{back} = \frac{\sum_{i:M_{grid}(i,j)=1} w_i^{back} \cdot \mathbf{f}^g_{i,j}}{\sum_{i:M_{grid}(i,j)=1} w_i^{back}} \tag{5}$$

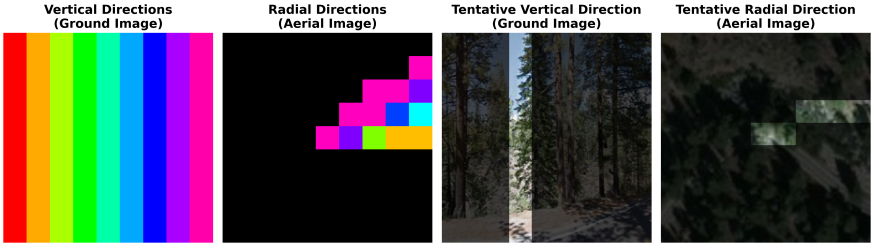

Figure 3: Radial directions used for aerial image token aggregation, visualized as a rainbow pattern. Each color represents a distinct radial direction corresponding to the angular offset from the center of the aerial image.

where the depth-dependent weights are defined as:

$$w_i^{fore} = d_{i,j} \tag{6}$$

$$w_i^{mid} = \begin{cases} \frac{d_{i,j}}{\tau} & \text{if } d_{i,j} \leq 0.5 \\ \frac{1-d_{i,j}}{d_{i,j}} & \text{otherwise} \end{cases} \tag{7}$$

$$w_i^{back} = 1 - d_{i,j} \tag{8}$$

with threshold $\tau = 0.5$. This weighting scheme emphasizes close objects for foreground, balanced weights for middleground, and distant objects for background layers.

Radial Direction Analysis: For aerial images, we extract features along radial directions from the center, using linear weight progressions:

$$\mathbf{r}_\beta^{fore} = \frac{\sum_{r=0}^{R} w_r^{fore} \cdot \mathbf{f}_{\beta,r}^a}{\sum_{r=0}^{R} w_r^{fore}} \tag{9}$$

$$\mathbf{r}_\beta^{mid} = \frac{\sum_{r=0}^{R} w_r^{mid} \cdot \mathbf{f}_{\beta,r}^a}{\sum_{r=0}^{R} w_r^{mid}} \tag{10}$$

$$\mathbf{r}_\beta^{back} = \frac{\sum_{r=0}^{R} w_r^{back} \cdot \mathbf{f}_{\beta,r}^a}{\sum_{r=0}^{R} w_r^{back}} \tag{11}$$

where β represents the angular direction, r is the radial distance from center, and the weights follow: $w_r^{fore} = 1 - r/R$ (decreasing), $w_r^{back} = r/R$ (increasing), and w_r^{mid} follows a triangular pattern peaking at the center.

4.5 Cross-Modal Orientation Estimation

We estimate orientation by finding the angular offset that minimizes the cosine distance between corresponding vertical and radial feature aggregations. For each candidate orientation θ, we compute the alignment cost:

$$\mathcal{L}(\theta) = \frac{1}{G}\sum_{i=0}^{G}\left\| 1 - \begin{bmatrix} \mathbf{t}_{G-1-i}^{fore} \\ \mathbf{t}_{G-1-i}^{mid} \\ \mathbf{t}_{G-1-i}^{back} \end{bmatrix}^{T} \begin{bmatrix} \mathbf{r}_{\phi(\theta,i)}^{fore} \\ \mathbf{r}_{\phi(\theta,i)}^{mid} \\ \mathbf{r}_{\phi(\theta,i)}^{back} \end{bmatrix} \right\| \tag{12}$$

where $\phi(\theta,i) = (\lfloor \theta/\Delta\theta \rfloor + i - G/2) \bmod |\mathcal{R}|$ maps vertical columns to radial directions, $\Delta\theta = \mathrm{FOV}_x/G$ is the angular step size.

The optimal orientation is then found as:

$$\theta^* = \arg\min_{\theta\in[0,2\pi)} \mathcal{L}(\theta) \tag{13}$$

To assess the reliability of orientation estimates, we compute a confidence score based on the Z-score of the minimum distance:

$$\text{confidence} = \frac{\mu(\mathcal{L}) - \min(\mathcal{L})}{\sigma(\mathcal{L})} \tag{14}$$

where $\mu(\mathcal{L})$ and $\sigma(\mathcal{L})$ are the mean and standard deviation of the loss values across all candidate orientations. Higher confidence scores indicate more reliable orientation estimates.

5 Results

To evaluate the performance of the proposed method, we analyze its performance in the CVUSA dataset and we compare our method against state-of-the-art approaches in cross-view orientation estimation.

The CVUSA dataset includes panoramic images, from which we extract random FOV windows with:

FOV_x	$90°$
FOV_y	$180°$
Yaw	$\psi \sim (0°, 360°)$

Table 2: FOV and yaw settings for the CVUSA dataset.

Aerial images undergo center cropping and resizing to match the standard 224×224 ViT input size.

Then, we apply our method to all the images belonging to the CVPR subset of CVUSA, storing the errors in orientation estimates for evaluation.

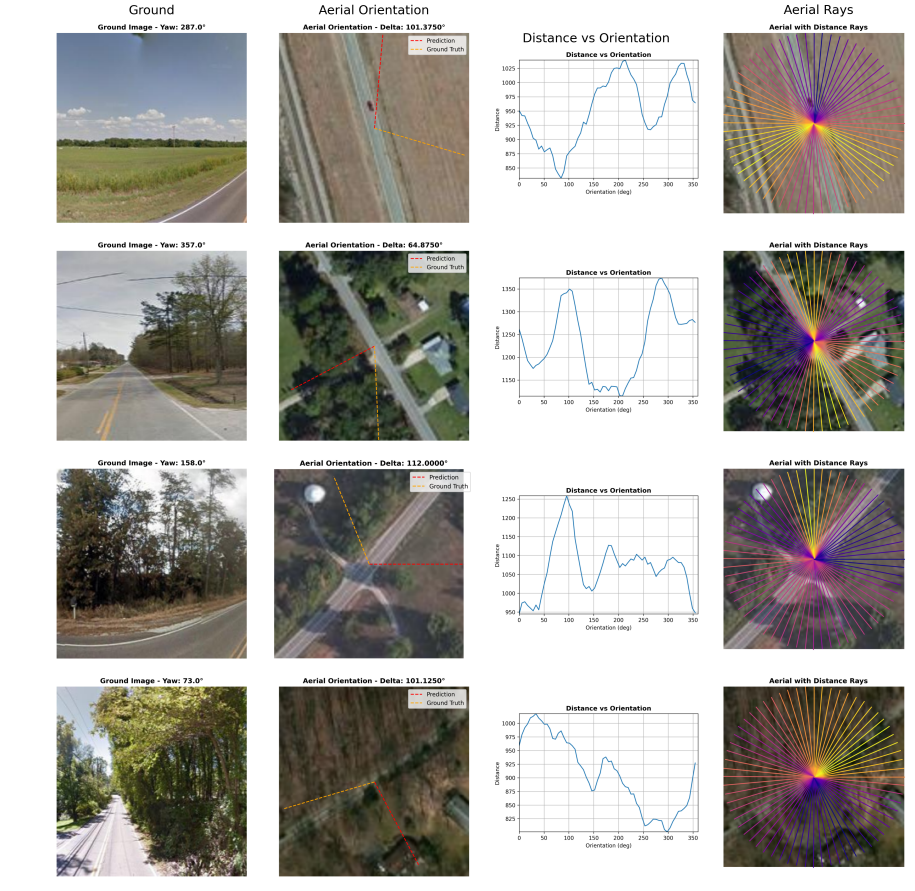

Figure 4: Orientation estimation errors for the CVPR subset of CVUSA dataset. Each bar represents the average error for a specific image, highlighting the effectiveness of our method in various scenarios.

6 Conclusions

We presented a zero-shot approach to cross-view orientation estimation that leverages strong vision foundation models with simple, interpretable token-aggregation and lightweight geometric priors (sky filtering and depth cues). While our experiments show that this zero-shot pipeline underperforms supervised counterparts in absolute accuracy, it offers several positives: no task-specific training, easy portability across backbones, transparent failure modes, and a low-cost baseline for rapid analysis and ablations. In practice, we find it provides meaningful signals in many scenes and can serve as a diagnostic tool or complementary component within stronger systems.

Beyond the method, this work contributes a new test benchmark with balanced global coverage. The CVGlobal evaluation set spans diverse continents and urban/rural contexts, enabling geographically fair assessment and reducing region-specific bias. We hope this resource will facilitate more rigorous comparisons and catalyze progress toward globally

robust cross-view localization and orientation estimation.

Future work includes integrating light supervision to bridge the zero-shot gap, exploring improved geometric priors and uncertainty calibration, and scaling evaluation across additional modalities.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[3] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[4] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[5] Sixing Hu, Mengdan Feng, Rang MH 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.

[6] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[7] Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3501–3510, 2018.

[8] Scott Workman, Richard Souvenir, and Nathan Jacobs. On the location dependence of convolutional neural network features. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 70–78, 2015.

[9] 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.