CVGlobal

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

INRIA
Evergreen Team
Montpellier, France

Abstract

This document demonstrates the format requirements for papers submitted to the British Machine Vision Conference. The format is designed for easy on-screen reading, and to print well at one or two pages per sheet. Additional features include: pop-up annotations for citations [0, 0]; a margin ruler for reviewing; and a greatly simplified way of entering multiple authors and institutions.

All authors are encouraged to read this document, even if you have written many papers before. As well as a description of the format, the document contains many instructions relating to formatting problems and errors that are common even in the work of authors who *have* written many papers before.

1 Introduction

Image alignment is a fundamental task in computer vision with wide-ranging applications including image stitching, 3D reconstruction, and augmented reality. Traditional approaches for image alignment often rely on handcrafted feature detectors and descriptors, followed by robust matching techniques. While effective in many scenarios, these methods can struggle when faced with significant viewpoint changes or complex scenes.

The recent advancements in deep learning, particularly with Vision Transformer (ViT) models, have opened new avenues for addressing challenges in image alignment. ViTs have demonstrated exceptional capabilities in learning rich and discriminative visual representations, enabling them to capture semantic information effectively. This has led to significant progress in various computer vision tasks, including image classification, object detection, and semantic segmentation.

In this work, we explore the potential of pre-trained ViT models for aligning ground-view and aerial-view images, a challenging task due to the drastic difference in perspectives. We leverage the powerful representations learned by these models to establish correspondences between salient elements in both views. Our approach builds upon the hypothesis that similar objects or regions, such as roads, buildings, or vegetation, should exhibit similar embeddings or tokens in the model's output space, even when observed from vastly different viewpoints.

To enhance the alignment accuracy, we incorporate sky filtering and depth estimation techniques to eliminate irrelevant sky regions and prioritize ground features. We also introduce a novel alignment strategy based on averaging tokens along vertical lines in the ground image and radial lines in the aerial image, enabling robust comparison of representations across the two views.

Our experimental results on the CVUSA dataset demonstrate the effectiveness of our proposed approach. We achieve promising alignment accuracy, even in the presence of challenging scenarios with significant viewpoint changes and complex scenes.

The main contributions of this paper can be summarized as follows:

- We propose a novel approach for aligning ground-view and aerial-view images using pre-trained ViT models.
- We introduce a robust alignment strategy based on averaging tokens along vertical and radial lines, enabling effective comparison of representations across the two views.
- We incorporate sky filtering and depth estimation techniques to improve alignment accuracy by focusing on relevant ground features.
- Our experimental results on the CVUSA dataset demonstrate the effectiveness of our proposed approach.

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 feature-based approaches to sophisticated deep learning methods, driven by applications in robotics, augmented reality, and autonomous navigation.

2.1 Cross-View Geo-localization and Orientation Estimation

Early pioneering works by Workman and Jacobs [10**?**] established the foundation for cross-view matching by demonstrating the potential of CNNs for learning robust feature representations across viewpoint variations. The introduction of benchmark datasets like CVUSA [10] and later VIGOR [12] catalyzed systematic research in this domain, with the latter providing more challenging same-area and cross-area evaluation protocols that 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 [6], 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 [9].

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.

2.2 Vision Transformers and Self-Supervised Learning

The advent of Vision Transformers (ViTs) [3] has revolutionized cross-view geo-localization by enabling models to capture long-range dependencies and 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 robust visual representations without manual annotations. Methods such as MoCo [5], SimCLR [2], and BYOL [4] have shown that self-supervised features can match or exceed supervised counterparts on various downstream tasks. DINOv2 [8] represents the current state-of-the-art, 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.

2.3 Multi-Modal Integration and Dataset Limitations

Recent research has recognized the value of incorporating complementary modalities to improve cross-view matching performance. Sky segmentation techniques help eliminate uninformative regions in ground-level images, focusing attention on building structures and terrain features visible in aerial views [10]. Advanced monocular depth estimation models like Depth-Anything [11] provide robust geometric cues that enable multi-scale feature aggregation and informed spatial reasoning.

Attention mechanisms have proven particularly effective for cross-view alignment, with cross-attention layers learning to focus on corresponding regions between aerial and ground views. However, most attention-based methods require supervised training with orientation labels and struggle with the limited field-of-view constraints common in real-world applications.

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 novel orientation-aware methods that combine self-supervised features with multi-modal cues for robust cross-view orientation estimation without requiring extensive supervision.

3 Dataset Generation Method

In this section, we present our methodology for constructing CVGlobal, a large-scale 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.

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

3.1 Dataset Design and Sampling Strategy

Our dataset construction methodology is guided by three key principles: *geographical diversity*, *balanced representation*, and *multi-modal consistency*. 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
- **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 Pipeline

Our data acquisition pipeline ensures high-quality multi-modal data collection through systematic coordinate generation, validation, and image retrieval. For each sampling 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. Our pipeline implements robust error handling with exponential backoff retry strategies and supports resumption capabilities for interrupted data collection sessions.

The resulting CVGlobal dataset provides balanced, geographically diverse paired satellite and street-view imagery suitable for training and evaluating cross-view methods across varied global contexts.

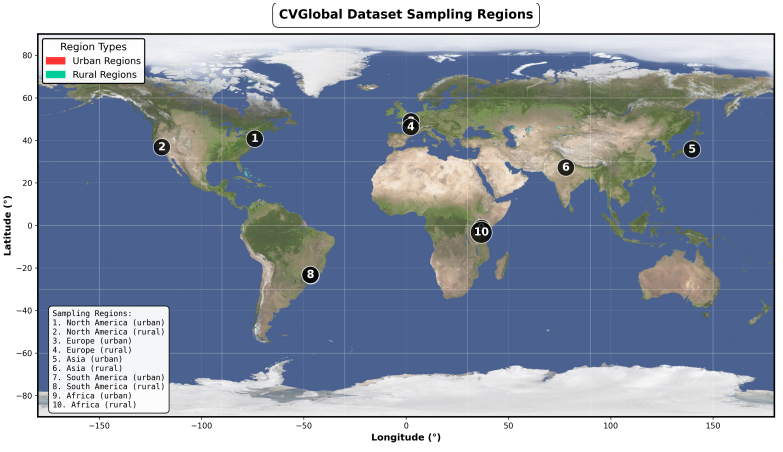

Figure 1: Global distribution of CVGlobal dataset sampling regions across five continents, showing urban (red) and rural (teal) areas with numbered locations corresponding to Table 1.

4 Crossview Method

We propose CroDINO, a novel approach for cross-view orientation estimation that leverages backbone-agnostic feature extraction with orientation-aware token aggregation strategies. Our method addresses the fundamental challenge of aligning ground-level panoramic 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 FOV angles, ψ is the yaw (rotation around the vertical axis), and ϕ is the pitch (elevation angle).

4.2 Architecture Overview

4.2.1 Backbone-Agnostic Feature Extraction

CroDINO employs a flexible architecture that can utilize different pre-trained vision models as feature extractors. Our implementation supports multiple backbone architectures including Vision Transformers (DINOv2, CLIP) and Convolutional Neural Networks (ResNet50), allowing for comparative analysis and optimal performance selection.

The backbone-agnostic design consists of:

- **Flexible Feature Extractor**: Support for DINOv2-ViT-B/14, CLIP-ViT-Base-Patch16, or ResNet50 as frozen feature extractors.

- **Unified Token Interface**: Standardized token representation regardless of backbone architecture.
- **Dynamic Grid Adaptation**: Automatic grid size calculation based on token dimensions from different architectures.

4.2.2 Multi-Backbone Token Processing

The feature extraction process varies by backbone but produces consistent token representations:

Vision Transformer Backbones: For DINOv2 and CLIP models, patch embeddings are extracted and normalized:

$$\mathbf{F}_g^{raw} = \text{Backbone}(I_g) \tag{1}$$
$$\mathbf{F}_a^{raw} = \text{Backbone}(I_a) \tag{2}$$
$$\mathbf{F}_g = \text{L2Normalize}(\mathbf{F}_g^{raw}[:, 1:, :]) \tag{3}$$
$$\mathbf{F}_a = \text{L2Normalize}(\mathbf{F}_a^{raw}[:, 1:, :]) \tag{4}$$

CNN Backbones: For ResNet50, convolutional features are flattened into token-like representations:

$$\mathbf{F}_g^{conv} = \text{ResNet50}(I_g) \tag{5}$$
$$\mathbf{F}_a^{conv} = \text{ResNet50}(I_a) \tag{6}$$
$$\mathbf{F}_g = \text{Reshape}(\mathbf{F}_g^{conv}, (-1, D)) \tag{7}$$
$$\mathbf{F}_a = \text{Reshape}(\mathbf{F}_a^{conv}, (-1, D)) \tag{8}$$

where D represents the feature dimension (768 for ViTs, 2048 for ResNet50). The grid size G is dynamically calculated as $G = \sqrt{N}$ where N is the number of tokens.

4.3 Orientation-Aware Token Aggregation

4.3.1 Sky Filtering and Depth Estimation

To improve orientation estimation, we incorporate semantic and geometric priors:

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: We utilize the Depth-Anything model to generate depth maps D for ground images. 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).

4.3.2 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:

$$\mathbf{t}_j^{fore} = \frac{\sum_{i:M_{grid}(i,j)=1} w_i^{fore} \cdot \mathbf{f}_{i,j}^g}{\sum_{i:M_{grid}(i,j)=1} w_i^{fore}} \tag{9}$$

$$\mathbf{t}_j^{mid} = \frac{\sum_{i:M_{grid}(i,j)=1} w_i^{mid} \cdot \mathbf{f}_{i,j}^g}{\sum_{i:M_{grid}(i,j)=1} w_i^{mid}} \tag{10}$$

$$\mathbf{t}_j^{back} = \frac{\sum_{i:M_{grid}(i,j)=1} w_i^{back} \cdot \mathbf{f}_{i,j}^g}{\sum_{i:M_{grid}(i,j)=1} w_i^{back}} \tag{11}$$

where the depth-dependent weights are defined as:

$$w_i^{fore} = d_{i,j} \tag{12}$$

$$w_i^{mid} = \begin{cases} \frac{d_{i,j}}{\tau} & \text{if } d_{i,j} \leq 0.5 \\ \frac{1-d_{i,j}}{d_{i,j}} & \text{otherwise} \end{cases} \tag{13}$$

$$w_i^{back} = 1 - d_{i,j} \tag{14}$$

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:

$$\mathbf{r}_\beta^{fore} = \frac{\sum_{r=0}^{R} w_r^{fore} \cdot \mathbf{f}_{\beta,r}^a}{\sum_{r=0}^{R} w_r^{fore}} \tag{15}$$

$$\mathbf{r}_\beta^{mid} = \frac{\sum_{r=0}^{R} w_r^{mid} \cdot \mathbf{f}_{\beta,r}^a}{\sum_{r=0}^{R} w_r^{mid}} \tag{16}$$

$$\mathbf{r}_\beta^{back} = \frac{\sum_{r=0}^{R} w_r^{back} \cdot \mathbf{f}_{\beta,r}^a}{\sum_{r=0}^{R} w_r^{back}} \tag{17}$$

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.

4.4 Orientation Estimation

4.4.1 Cross-Modal Alignment via Cosine Distance Minimization

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:

$$\mathcal{L}(\theta) = \frac{1}{G}\sum_{i=0}^{G}\left\| 1 - \begin{bmatrix} \mathbf{t}^{fore}_{G-1-i} \\ \mathbf{t}^{mid}_{G-1-i} \\ \mathbf{t}^{back}_{G-1-i} \end{bmatrix}^{T} \begin{bmatrix} \mathbf{r}^{fore}_{\phi(\theta,i)} \\ \mathbf{r}^{mid}_{\phi(\theta,i)} \\ \mathbf{r}^{back}_{\phi(\theta,i)} \end{bmatrix} \right\| \tag{18}$$

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

The optimal orientation is found through exhaustive search:

$$\theta^* = \arg\min_{\theta \in [0,360)} \mathcal{L}(\theta) \tag{19}$$

4.4.2 Confidence Estimation

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

$$\text{confidence} = \frac{\mu(\mathcal{L}) - \min(\mathcal{L})}{\sigma(\mathcal{L})} \tag{20}$$

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.

4.5 Implementation Details

4.5.1 Backbone-Specific Configurations

Our implementation supports three backbone architectures:

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. Positional embeddings are interpolated for different input sizes.

CLIP: Employs CLIP-ViT-Base-Patch16 with 16×16 pixel patches, generating a 14×14 token grid with 768-dimensional features. L2 normalization is applied to token representations for improved stability.

ResNet50: Uses convolutional features from the penultimate layer, which are reshaped into 2048-dimensional tokens. The spatial resolution depends on the input size and stride configuration.

4.5.2 Training Strategy and Preprocessing

Our approach operates in a largely unsupervised manner, leveraging pre-trained features without requiring orientation labels during training. The orientation estimation is performed through geometric alignment of feature aggregations.

Data Preprocessing: We extract random FOV windows from panoramic images with:

- Horizontal FOV: 90° (configurable)
- Vertical FOV: 180°

- Random yaw: $\psi \sim \text{Uniform}(0^\circ, 360^\circ)$
- Fixed pitch: $\phi = 90^\circ$

Aerial images undergo center cropping and resizing to match the ground image dimensions.

4.5.3 Pipeline Architecture

The complete pipeline processes image pairs through the following stages:

1. **Feature Extraction**: Backbone-specific token generation with dynamic grid size calculation
2. **Sky Segmentation**: CNN-based sky filtering with guided filter refinement
3. **Depth Estimation**: Depth-Anything model for geometric understanding
4. **Token Aggregation**: Multi-layer depth-weighted aggregation for both vertical and radial directions
5. **Orientation Search**: Exhaustive search over discretized orientation space with cosine similarity

All models are implemented in PyTorch and support both CPU and GPU execution. The orientation search space is discretized with angular steps of $\Delta\theta = \text{FOV}_x/G$, where the grid size G is dynamically determined from the backbone's token dimensions. This flexible approach ensures optimal resolution regardless of the chosen backbone architecture.

References

[1] Authors. The frobnicatable foo filter, 2006. ECCV06 submission ID 324. Supplied as additional material `eccv06.pdf`.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] N. David Mermin. What's wrong with these equations? *Physics Today*, October 1989. http://www.cvpr.org/doc/mermin.pdf.

[8] 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.

[9] 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.

[10] 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.

[11] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[12] 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.