CORRECTING GEOSPATIAL DATA DISPLACEMENT WITH FOUNDATION VISION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Geospatial point annotations collected during field surveys often suffer from positional displacement due to GPS inaccuracy and environmental constraints, limiting their utility for downstream applications. Traditional alignment methods rely on multi-temporal imagery or task-specific training, restricting their practical applicability. We propose a simple approach that leverages foundation vision models to correct displaced annotations through semantic similarity matching. A semantic reference is constructed from a small set of annotated examples of the target class, and for each displaced point, we define a search region to identify the location with highest similarity to the reference set using embeddings from a feature extractor. We evaluate our method on a forestry dataset from the Amazon rainforest containing annotations for over 50 tree species. Linear probing experiments demonstrate that models trained on corrected annotations significantly outperform those trained on original displaced data, and qualitative analysis shows that corrections consistently shift points from background regions to target class. By requiring only a small set of reference examples and no retraining, our method provides a practical solution for improving geospatial annotation quality in field-based surveys.

1 INTRODUCTION

Geospatial data mapping plays a fundamental role in a wide range of environmental and ecological applications, such as forest inventory generation, biodiversity monitoring (Alvarez-Martinez et al., 2026), and natural resource management (Sharma et al., 2024). In many of these scenarios, data acquisition does not follow the traditional remote sensing pipeline in which aerial or satellite imagery is first collected and subsequently annotated. Instead, image acquisition is often performed in parallel with field surveys, where ground agents, typically domain specialists such as botanists, identify objects of interest in situ and register their locations using handheld GPS devices. These annotations are commonly represented as geospatial points corresponding to objects like individual trees and their respective species.

Despite being widely adopted, this field-based annotation process introduces a source of error. GPS devices frequently suffer from limited positional accuracy, and environmental constraints often prevent surveyors from positioning themselves exactly at the target object. As a result, the recorded point annotations are systematically displaced from their true intended locations in the imagery, we refer to this phenomenon as geospatial data displacement, and it significantly degrades the quality of the resulting datasets. In practice, even small spatial offsets can make these point-based annotations unreliable for downstream tasks (Kim et al., 2025), such as statistical analysis and the training of machine learning models (Chan et al., 2025), where precise spatial correspondence between image content and annotations is essential.

Several approaches have been proposed to correct spatial displacement in geospatial data, however, they present notable limitations. Most approaches either rely on multi-temporal image series of the same region (Duan et al., 2020) or require training task-specific networks on the target dataset (Vargas-Munoz et al., 2019), which restricts their applicability in many practical field survey settings. In this work, we propose a methodology for aligning geospatially displaced point annotations that does not require temporal data or additional training and instead leverages foundation vision models to guide the alignment process. The core idea is to guide the correction process using a semantic reference derived from a small set of examples that characterize the target class to be cor-

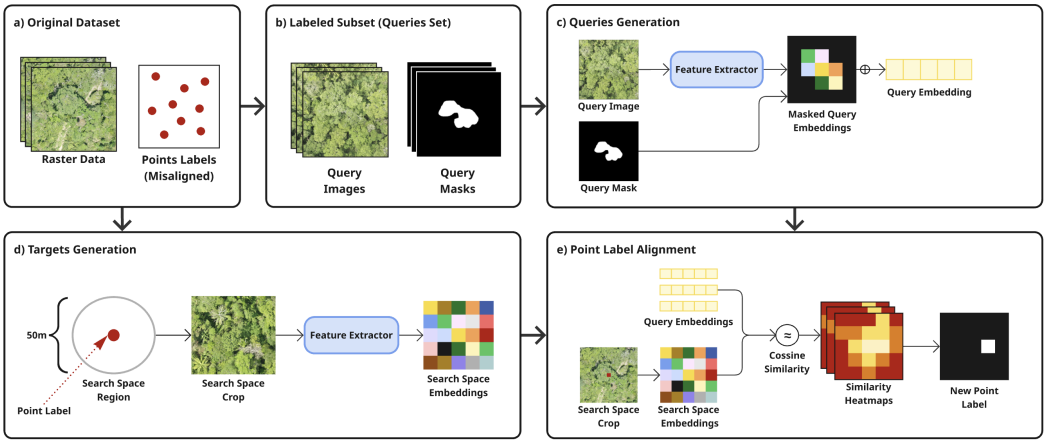

Figure 1: Overview of the proposed alignment pipeline. The figure is organized into steps (a)–(e), from reference selection and query embedding construction to search-space extraction and point correction.

rected, referred to as the query. Foundation vision models are used to extract semantic embeddings from these reference examples and from image regions around each displaced point, defined as the search space. By comparing these embeddings, we identify the location within the search space that is most semantically consistent with the reference, thereby repositioning the original misaligned point based on visual similarity rather than temporal information or task-specific patterns.

We evaluate the proposed methodology in the context of forestry monitoring using a private dataset provided by a governmental environmental agency in the Amazon region, which is employed to monitor more than 50 tree species in forest areas granted for selective logging in the Amazon rainforest. The dataset is particularly rich in terms of species diversity, spatial coverage, and annotation density, however the geospatial displacement currently limits its effective use. Correcting these misalignment not only provides a real-world demonstration of the effectiveness of the proposed method but also unlocks the potential of this dataset for a wide range of future studies.

2 METHODOLOGY

Our methodology was applied to the Amazon dataset, composed of multiple raster files covering approximately 6000 hectares of the Amazon rainforest and point annotations of individual trees and their species labels, which are misaligned with the underlying imagery. In our experiments, we focus on correcting annotations belonging to the *Shihuahuaco* class, which represents the most frequent species in the dataset. While the pipeline is described for a single class, it can be iteratively applied to other classes in a multi-class setting.

The proposed alignment strategy relies on constructing a robust semantic reference for the target class. Since no subset of annotations in the dataset can be assumed to be spatially accurate, a small set of reference samples was manually annotated with precise mask delineations. Given that, in our domain, tree species exhibit significant phenological variation over time, resulting in noticeable changes in visual appearance, the reference samples were selected to adequately capture this variability, ensuring that the semantic reference provides a complete representation of the target class.

To extract semantic representations from the reference samples, we employ DINOv3 (Siméoni et al., 2025) as the feature extractor, following recent work demonstrating its strong performance in learning rich visual representations (Malaisree et al., 2025). Although we use DINOv3 in this work, the same pipeline can be instantiated with any feature extractor capable of producing semantic representations that are discriminative for the target domain and task. In standard usage, DINO produces a single global embedding per image, however, our goal is to extract embeddings from specific regions of the image. We therefore extract the intermediate patch-level feature map before the global pooling stage, where each patch embedding encodes localized semantic information. As illustrated

in Figure 1-c, for each manually annotated reference image, embeddings are collected only from patches that fall within the annotated mask regions and are subsequently averaged to obtain a single representative embedding per reference instance. These embeddings, referred to as *query embeddings*, form the semantic reference for the target class.

For each displaced point annotation belonging to the target class, we define a spatial region where the true object is expected to lie, which we refer to as the *search space*. In our setting, the search space is defined as a circular neighborhood with a 50-meter diameter around the original point annotation, consistent with typical positional uncertainty in field surveys. As illustrated in Figure 1-d, a georeferenced image crop covering this region is extracted for each displaced point and serves as the input for the alignment procedure. Within each search space crop, we extract patch-level embeddings using the same feature extractor.

For the alignment process, as illustrated in Figure 1-e, we compute similarity maps by measuring cosine similarity between each query embedding and the patch embeddings in the search space, allowing each reference example to independently vote for visually consistent regions and generating N similarity maps per search space image, where N equals the number of reference examples in the query set. To combine evidence from all query embeddings, the similarity maps are aggregated by summation, emphasizing locations that are consistently supported across references. The aggregated heatmap is then smoothed using a Gaussian filter to suppress isolated responses and consolidate spatially coherent regions. Finally, the location with the highest similarity score is selected as the new corrected position for the displaced point annotation.

3 Results and Discussion

Evaluating the effectiveness of the proposed alignment method is inherently challenging due to the lack of a large, reliably aligned ground-truth set. As a result, we assess the proposed method using indirect evaluation strategies based on downstream learning performance and qualitative visual inspection.

3.1 Quantitative Evaluation

To quantitatively assess the impact of the proposed correction, we employ a linear probing experiment in which two classifiers are trained under identical conditions using different training datasets. In both cases, the task is to distinguish the *Shihuahuaco* class from all other species using image patches extracted around point annotations. The two training datasets differ only in how the positive samples are generated. In the first case, patches corresponding to *Shihuahuaco* are extracted from the original, displaced point annotations, while in the second case they are extracted from the corrected point annotations produced by our method. Negative samples corresponding to other species are identical in both training sets. Both classifiers are evaluated on the same validation set, which consists of a newly annotated small subset of data created specifically for validation and not used during either the correction process or training.

The model trained using corrected annotations consistently outperforms the model trained on the original displaced data, as shown in Table 1. Showing a big improvement in all metrics, with the F1-Score improving from 74% to 85%. This result indicates that the corrected positive samples are more visually and semantically consistent with the target class after the alignment process, leading to more discriminative features for classification.

3.2 Qualitative Analysis

Qualitative results shown in Figure 2 further support the previous findings, illustrating representative cases before and after correction. In several examples (Figure 2-a and b), corrected annotations successfully shift from background regions to the target tree crown. Moreover, points initially located near crown boundaries tend to move toward the center of the canopy (Figure 2-c), suggesting improved spatial positioning and removal of points from semantically irrelevant areas.

However, the independent evaluation of each point reveals an inherent limitation of the proposed approach. In cases where two or more displaced points have overlapping search spaces and multiple target objects of the desired class exist within this shared region, the method may assign both points

Table 1: Linear probing results for *Shihuahuaco* vs. other species when training with original (displaced) annotations versus our corrected annotations. Both models use a ViT-B backbone.

Training data	Backbone	Accuracy	Precision	Recall	F1
Original points (displaced)	ViT-B	0.7529	0.7986	0.7526	0.7406
Corrected points (ours)	ViT-B	**0.8588**	**0.8624**	**0.8529**	**0.8587**

(a) (b) (c) (d)

Figure 2: Qualitative alignment results. Original (displaced) points are shown in blue, and corrected points are shown in red.

to the same individual tree, as illustrated in Figure 2-d. This occurs because each point is corrected independently based solely on semantic similarity, not considering the results of the other aligments. While this represents a failure situation of the current approach, we note that such cases are very infrequent in our dataset and could potentially be addressed in future work through post-processing steps.

4 Conclusion

In this work, we introduce a simple and effective approach for correcting geospatially displaced point annotations based on semantic similarity using foundation models for feature extraction. Although we evaluate the method on a forestry dataset, it can be directly applied to other point-based annotation problems and domains, as long as a feature extractor capable of producing semantically relevant representations for the domain is available.

References

Jose M. Alvarez-Martinez et al. Four decades of remote sensing for monitoring terrestrial ecosystems: a global review and future challenges. *Science of Remote Sensing*, 13:100341, 2026.

Sixian Chan et al. Sparse point annotations for remote sensing image segmentation. *Scientific Reports*, 15:27347, 2025.

Weiwei Duan, Yao-Yi Chiang, Stefan Leyk, Johannes H. Uhl, and Craig A. Knoblock. Automatic alignment of contemporary vector data and georeferenced historical maps using reinforcement learning. *International Journal of Geographical Information Science*, 34(4):824–849, 2020.

Jooho Kim et al. Geosight: Enhancing object geo-localization with visual similarity and coordinate referencing. *International Journal of Disaster Risk Reduction*, 128:105730, 2025.

P. Malaisree et al. Dino-yolo: Enhancing object detection for remote sensing images through vision foundation models. *CoRR*, abs/2510.25140, 2025.

S. Sharma et al. Remote sensing and gis in natural resource management: A review of recent advancements and future perspectives. *Remote Sensing*, 16(22):4161, 2024.

Oriane Siméoni et al. Dinov3. *CoRR*, abs/2508.10104, 2025.

John E. Vargas-Munoz et al. Correcting rural building annotations in openstreetmap using convolutional neural networks. *CoRR*, abs/1901.08190, 2019.