GeoFT: Fine-tuning Foundation Models for Automated OSINT Geolocation

Anonymous authors
Paper under double-blind review

Abstract

Open source intelligence (OSINT) investigators face the challenge of verifying the location of media shared online. Traditional geolocation requires manual effort and cannot scale with the ever-growing volume of images and videos shared on social media. We present GeoFT, a fine-tuned version of GeoCLIP specifically optimized for geolocation in Russia and Ukraine. By focusing on street-level imagery and leveraging community-validated datasets, our model achieves significantly improved accuracy compared to existing solutions. On our test set, GeoFT reduces the average error from 3,520km to 2,150km while maintaining interpretable confidence scores. We demonstrate the model's potential for aiding OSINT investigations and discuss pathways for deployment in real-world applications.

1 Introduction

The verification of media location, or geolocation, has traditionally been a crowdsourced effort relying on human investigators using resources like Google Street View and building databases. While effective, this manual process cannot keep pace with the thousands of new images appearing daily on platforms like Decrypt. We present GeoFT, a fine-tuned model that specifically targets street-level imagery in regions of active conflict, where traditional solutions like Google Street View may be outdated or unavailable.

2 Related Work

Recent approaches to AI geolocation include GeoCLIP (Vivanco Cepeda et al., 2023), PIGEOTTO (Haas et al., 2024), and GeoDecoder (Qi et al., 2024). GeoCLIP utilizes contrastive learning between image and location embeddings but exhibits high average error (3,520km). PIGEOTTO and GeoDecoder showed promising results but lack open-source implementations. Commercial solutions like GeoSpy (GeoSpy, 2024) rely heavily on static image databases, limiting their effectiveness in rapidly changing environments. These approaches demonstrate the challenge of accurate geolocation at scale.

3 Methodology

3.1 Data Collection and Filtering

We curated a dataset combining two primary sources:

1. Eyes on Russia (EoR): 2,887 community-geotagged images from the conflict region (Eyes on Russia Project, 2024)
2. Google Street View (GSV): 16,159 street-level images gathered via API (Google LLC, 2024)

To ensure data quality, we employed GPT-4-mini as a binary classifier to filter images with the prompt: "Does this image contain street features?" This preprocessing step significantly improved training data relevance by removing indoor scenes and irrelevant imagery. The

final dataset contained approximately 19,000 images after filtering. The test set consisted of only EoR images to reflect realistic images from social media.

3.2 Model Architecture

We build upon GeoCLIP's architecture, consisting of:

1. Location encoder L(*) that transforms 2D GPS coordinates into high-dimensional features through:
   (a) Equal Earth Projection to minimize geographic distortion
   (b) Random Fourier Features (RFF) for positional encoding
   (c) Hierarchical representation using multiple RFF frequencies
   (d) MLP layers for processing
2. Image encoder V(*) based on a pre-trained CLIP ViT-L/14 model with two additional trainable linear layers (h1: 768 dimensions, h2: 512 dimensions)

During training, we keep the CLIP backbone frozen and only train the linear layers of the image encoder and the location encoder components. Training uses contrastive learning with batch and queue-based negatives to minimize the loss:

$$\mathcal{L}_i = -\sum_{j=1}^{P} \log \frac{\exp(V_{ij} \cdot L_{ij}/\tau)}{\sum_{i=0}^{|B|} \exp(V_{ij} \cdot L_{ij}/\tau) + \sum_{i=0}^{S} \exp(V_{ij} \cdot \bar{L}_i/\tau)} \tag{1}$$

where τ is the temperature parameter, $|B|$ is the batch size, S is the queue size, and $\bar{L}_i$ represents the queue of GPS embeddings used as additional negatives.

4 Results

Our evaluation reveals three key findings that demonstrate GeoFT's effectiveness and practical applicability.

First, when trained on the combined Eyes on Russia (EoR) and Google Street View (GSV) dataset, GeoFT demonstrates significant improvements over baseline models. Specifically, GeoFT achieves state-of-the-art performance on the Im2GPS3k dataset with improvements of +1.31%, +0.97%, +3.95%, +8.67%, and +3.72% at 1km, 25km, 200km, 750km, and 2500km thresholds respectively compared to prior work.

Second, the fine-tuned model exhibits substantially higher confidence in its predictions. While the baseline model shows uncertainty in its geolocation estimates (mean confidence: 0.03), GeoFT produces well-calibrated confidence scores reaching up to 0.61 for high-confidence predictions, with confidence scores strongly correlating with prediction accuracy. The standard deviation in confidence scores increased from 0% to 5%, indicating a better ability to distinguish between high and low confidence predictions.

Third, and perhaps most significantly for practical applications, GeoFT demonstrates strong performance even when trained solely on GSV data, without requiring region-specific human-labeled datasets like EoR. This finding is crucial for extending the approach to new geographic regions where specialized datasets may not be available. Table 1 compares performance between models trained on the combined dataset versus GSV data only.

As shown in Table 1, while performance does decrease without EoR data, particularly at finer granularities, the model still achieves strong results at continental (2500km) and country (750km) scales using only GSV data. This is particularly noteworthy since GSV data is widely available globally, making our approach extensible to new regions without requiring specialized datasets. The model trained only on GSV data still outperforms baseline methods, demonstrating the effectiveness of our approach even with limited training data.

The confidence metrics also reflect this pattern - the GSV-only model exhibits lower but still well-calibrated confidence scores (max 2% vs 61% with combined data), appropriately

Table 1: Performance comparison between models trained on combined EoR+GSV data versus GSV data only. While performance decreases without EoR data, the model still achieves strong results at larger scales using only widely-available GSV data.

Distance	EoR+GSV	GSV Only	Difference
2500km	100%	92%	-8%
750km	95%	54%	-41%
200km	59%	7%	-52%
25km	38%	6%	-32%
1km	6%	0%	-6%

indicating its reduced certainty in predictions. This ability to produce reliable confidence estimates, even with limited training data, is crucial for practical applications where understanding prediction reliability is important.

5 Discussion and Future Work

While GeoFT shows promising results, several extensions could enhance its utility:

1. Integration with production systems for continuous model improvement using new validated data
2. Expansion to video and aerial imagery analysis
3. Extension to other regions of interest with similar data collection methodology

The model can be deployed both as a standalone tool for OSINT investigators and integrated into existing intelligence platforms, providing automated first-pass location estimates for human verification.

6 Conclusion

GeoFT demonstrates that fine-tuning foundation models on carefully curated regional data can significantly improve geolocation accuracy. The success of our approach using only Google Street View data suggests that this methodology could be extended to other regions of interest, even without access to specialized OSINT datasets. This work represents a step toward scalable, automated support for OSINT investigations.

References

Eyes on Russia Project. Eyes on russia. `https://eyesonrussia.org`, 2024. Accessed: February 2024.

GeoSpy. Geospy: Ai-powered image geolocation. `https://geospy.ai`, 2024. Accessed: February 2024.

Google LLC. Google street view. `https://www.google.com/streetview`, 2024. Accessed: February 2024.

Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. Pigeon: Predicting image geolocations. *Preprint*, May 2024.

Feng Qi, Mian Dai, Zixian Zheng, and Chao Wang. Geodecoder: Empowering multimodal map understanding. *arXiv preprint*, Feb 2024. arXiv:2401.2401.15118.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. In *Advances in Neural Information Processing Systems*, 2023. NeurIPS 2023.