Enhancing Contrastive Learning for Geolocalization by Discovering Hard Negatives on Semivariograms

Boyi Chen
boyi.chen@tum.de
Professorship Big Geospatial Data Management, Technical University of Munich, Germany

Zhangyu Wang*
zhangyu.wang@maine.edu
University of Maine, United States

Fabian Deuser
fabian.deuser@unibw.de
University of the Bundeswehr Munich, Germany

Johann Maximilian Zollner
maximilian.zollner@tum.de
Professorship Big Geospatial Data Management, Technical University of Munich, Germany

Martin Werner
martin.werner@tum.de
Professorship Big Geospatial Data Management, Technical University of Munich, Germany

Abstract

Accurate and robust image-based geo-localization at a global scale is challenging due to diverse environments, visually ambiguous scenes, and the lack of distinctive landmarks in many regions. While contrastive learning methods show promising performance by aligning features between street-view images and corresponding locations, they neglect the underlying spatial dependency in the geographic space. As a result, they fail to address the issue of false negatives - image pairs that are both visually and geographically similar but labeled as negatives, and struggle to effectively distinguish hard negatives, which are visually similar but geographically distant. To address this issue, we propose a novel spatially regularized contrastive learning strategy that integrates a semivariogram, which is a geostatistical tool for modeling how spatial correlation changes with distance. We fit the semivariogram by relating the distance of images in feature space to their geographical distance, capturing the expected visual content in a spatial correlation. With the fitted semivariogram, we define the expected visual dissimilarity at a given spatial distance as reference to identify hard negatives and false negatives. We integrate this strategy into GeoCLIP and evaluate it on the OSV5M dataset, demonstrating that explicitly modeling spatial priors improves image-based geo-localization performance, particularly at finer granularity.

CCS Concepts

• **Computing methodologies → Visual content-based indexing and retrieval**; **Image representations**; **Learning latent representations**.

Keywords

contrastive learning, geostatistics, geolocalization

[*] Corresponding author.

SIGSPATIAL '25, Minneapolis, MN, USA
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2086-4/2025/11
https://doi.org/10.1145/3748636.3763220

ACM Reference Format:
Boyi Chen, Zhangyu Wang, Fabian Deuser, Johann Maximilian Zollner, and Martin Werner. 2025. Enhancing Contrastive Learning for Geolocalization by Discovering Hard Negatives on Semivariograms. In *The 33rd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '25), November 3–6, 2025, Minneapolis, MN, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3748636.3763220

1 Introduction

Image-based geo-localization aims to determine the geographic location of a query image. This is particularly useful in scenarios where Global Positioning System (GPS) signals are unavailable or unreliable, such as dense urban areas with tall buildings or mountainous regions. It has a wide range of applications in disaster response and augmented reality [8, 10]. However, image-based geo-localization at the global scale is a critical challenge in computer vision and GeoAI due to the diversity of visual appearances across the world and the lack of distinctive landmarks in many regions.

Recently, contrastive learning has significantly advanced the field of image-based geo-localization by enabling models to capture rich semantic information that is crucial for recognizing locational cues in images. Contrastive Language-Image Pre-Training (CLIP) [13] has become a powerful learning paradigm in geo-localization tasks, especially at the global scale, where representations can be learned by leveraging the supervision of large-scale geo-tagged imagery. By encouraging semantically similar pairs to be close in the embedding space and pushing dissimilar pairs apart, contrastive learning effectively aligns visual inputs with their corresponding labels, whether textual descriptions or geographic coordinates.

Despite these advances, current approaches mainly treat geographic coordinates as pure numerical inputs and ignore their interactions with the observed images. They do not consider the fact that images geographically near to each other are often visually similar due to spatial autocorrelation, a phenomenon explained by Tobler's First Law of Geography [16], which states that "everything is related to everything else, but near things are more related than distant things". This spatial autocorrelation remains underutilized in current geo-localization models. Furthermore, although contrastive learning architectures such as CLIP benefit from hard negatives by promoting more discriminative representations, the

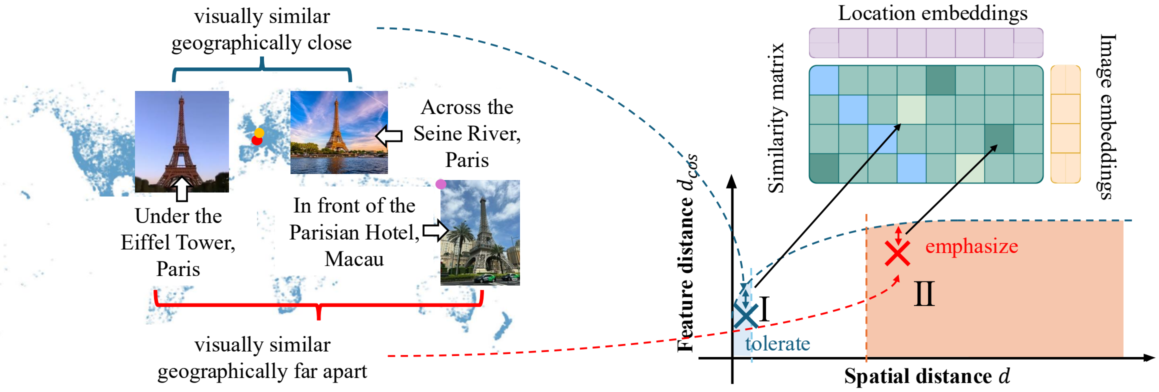

Figure 1: Spatial-aware reweighting strategy. Based on the semivariogram, we (1) emphasize visually similar but geographically distant pairs (hard negatives), and (2) tolerate visually similar and geographically close pairs (false negatives).

visually similar samples may not be hard negatives but potentially false negatives. Hence, properly identifying and handling these negatives is a key to improving the performance of contrastive learning models.

Unlike general vision tasks, geo-localization involves spatial data, where similarity depends on distance-dependent patterns. Thus, we propose to integrate the geographic information into the contrastive learning loss with a semivariogram, to make the model aware of the geographic reality. The semivariogram is a geostatistical tool that models the expected dissimilarity between samples as a function of their geographic distances. We model the expected relation in street view images across different regions, which allows us to distinguish between hard negatives and false negatives. Consequently, we penalize hard negatives during training to reduce the impact of false negatives on the performance of the model.

Contributions. In summary, our contributions are: (1) We introduce a semivariogram-based approach to model spatial autocorrelation in image similarity, demonstrating that geographic distance and visual similarity are systematically related at a global scale. (2) We design a semivariogram-guided reweighting strategy within contrastive learning to explicitly identify and handle hard negatives and false negatives, improving the model's spatial awareness.

2 Related Work

Image-based geo-localization methods can be broadly categorized into classification-based and retrieval-based approaches. Classification-based methods, such as PlaNet [20], divide the Earth's surface into a set of cells and treat the localization task as an image classification problem, the model predicts which geographic cell the input image most likely belongs to. Retrieval-based methods like Im2GPS [18] match the query image with a reference database to query the most similar one as prediction. StreetCLIP [6] adapts CLIP to geolocalization by matching images to a batch of synthetically generated location descriptions. During inference, the model predicts the location whose description yields the highest similarity to the image embedding, enabling effective zero-shot localization. Sample4Geo [3] using a symmetric InfoNCE contrastive learning loss with a novel hard negative sampling, selecting samples from both geographic proximity and visual similarity, leading to state-of-the-art performance cross-view geolocalization. Generative methods [5] propose a probabilistic formulation using diffusion and flow matching, with significant improvement at large scales.

3 Method

The proposed method introduces a negative mining strategy based on semivariogram information. We test its effectiveness with a contrastive learning architecture GeoCLIP [17]. Instead of treating all negative samples equally, we model spatial autocorrelation across the geographic space by fitting a theoretical semivariogram between the feature dissimilarities of image embeddings and their geographic distances. The fitted semivariogram defines a relationship between the spatial distance from an anchor image to its negative samples and their expected visual dissimilarity. Negative samples that significantly deviate from this relationship are considered to be potential hard negatives/false negatives. During training, we put more emphasis (i.e., higher weights) on hard negatives and less emphasis on false negatives.

3.1 GeoCLIP

GeoCLIP consists of two main components: an image encoder $\mathcal{V}(\cdot)$ and a location encoder $\mathscr{L}(\cdot)$. The image encoder is based on a frozen pre-trained CLIP image encoder ViT-L/14 [4], following a two-layer trainable MLP to fine-tune the extracted features. The location encoder first applies the Equal Earth Projection (EEP) [14] to reduce the distortion caused by unequal division of longitude and latitude. In order to capture rich details at different scales, GeoCLIP constructs a hierarchical representation using Random Fourier Features [15] with M different frequencies σ in a range from 2^0 to 2^8. Then the encoded hierarchical features are passed through a trainable MLP to learn the suitable embeddings. Let the Random Fourier Feature transformation be denoted as γ and the location MLP as f. Given a geo-tagged image (I_i, G_i) where I_i is an image and G_i is its coordinate, the image encoder $\mathcal{V}(\cdot)$ and the location encoder $\mathscr{L}(\cdot)$ for the feature extraction can then be written as:

$$V_i = \mathcal{V}(I_i) \tag{1}$$

$$L_i = \mathscr{L}(G_i) = \sum_{k=1}^{M} f_k(\gamma(EEP(G_i), \sigma_k) \tag{2}$$

During each step of training, we draw a batch of size B, $\mathcal{B} = \{(I_i, G_i)\}_{i=1}^{B}$, from the training dataset $\mathcal{D}_{train} = \{(I_n, G_n)\}_{n=1}^{N}$. For the i^{th} sample (V_i, L_i) in batch $\mathcal{B}$, the negative set $\mathcal{N}_i$ is defined as $(\mathcal{B} - \{(V_i, L_i)\}) \bigcap \mathcal{Q}$, where $\mathcal{Q}$ is a dynamic queue of geo-tagged images constructed from $\mathcal{D}_{train}$ as is described in [17]. For each batch, the training objective is to minimize the InfoNCE loss [12]:

$$\sum_{i=1}^{B} \mathcal{L}_i = \sum_{i=1}^{B} -\log \frac{\exp(V_i \cdot L_i/\tau)}{\exp(V_i \cdot L_i/\tau) + \sum_{L_j^- \in \mathcal{N}_i} \exp(V_i \cdot L_j^-/\tau)} \tag{3}$$

In practice, we use the average loss computed on $P = 2$ augmentations of each image to mitigate overfitting, as is a common trick used in contrastive learning [2].

3.2 Semivariogram

Given a finite set of locations G with associated univariate observations z, the classical semivariogram [9] models the spatial dependency by quantifying how the dissimilarity between observations increases as their spatial distance increases. The empirical semivariogram at distance h over a small tolerance ε is defined as:

$$\hat{\gamma}(h \pm \varepsilon) := \frac{1}{2|N(h \pm \varepsilon)|} \sum_{(G_i, G_j) \in N(h \pm \varepsilon)} |z_i - z_j|^2 \tag{4}$$

where $N(h \pm \varepsilon) := \{(G_i, G_j) | h - \varepsilon \le d(i, j) \le h + \varepsilon\}$ denotes the set of point pairs whose geographic distance d is within the threshold $h \pm \varepsilon$ (in our case, d is the great-circle distance from G_i to G_j). MC-GTA [19] generalized this to multivariate observations (e.g., image embeddings) by replacing the $|z_i - z_j|^2$ term with Wasserstein-2 distances. This generalization inspires us to adopt semivariograms in our geo-localization setting. We treat the image embeddings V_i as multivariate observations and measure their dissimilarity using cosine distances. Formally, given a pair of image embeddings (V_i, V_j) located at (G_i, G_j), we compute:

$$d_{\cos}(i, j) = 1 - \frac{V_i \cdot V_j}{||V_i||||V_j||} \tag{5}$$

where $\cdot$ denotes the dot product between the two feature vectors to measure the similarity. Replacing the univariate difference $|z_i - z_j|^2$ in Equation 4 with this cosine distance, we estimate the **empirical embedding semivariogram** $\hat{\gamma}_e$ that captures the expected image embedding dissimilarity as a function of geographic distance:

$$\hat{\gamma}_e(h \pm \varepsilon) := \frac{1}{2|N(d \pm \varepsilon)|} \sum_{(G_i, G_j) \in N(d \pm \varepsilon)} d_{\cos}(i, j). \tag{6}$$

Then we use the spherical semivariogram model to fit an analytical semivariogram function γ from the empirical semivariogram γ_e, which we will use in the next section.

3.3 Semivariogram Based Reweighting

In the loss computation (3), all negative pairs are treated equally regardless of their spatial context. However, this neglects the impact of false negatives and hard negatives in contrastive learning, specifically for geo-localization. To address these challenges, we propose a reweighting mechanism guided by the fitted semivariogram, allowing us to assign higher importance to hard negatives and reduce the impact of false negatives during training.

Let $d_{\cos}(i, j)$ and $d(i, j)$ denote the cosine distance and haversine distance between the negatives to the true sample. From the fitted semivariogram, we can define the expected visual dissimilarity $\gamma(d(i, j))$ at this spatial distance $d(i, j)$, and compare it to the calculated dissimilarity $d_{\cos}(i, j)$, yielding the deviation:

$$\delta(i, j) = d_{\cos}(i, j) - \gamma(d(i, j)) \tag{7}$$

This deviation quantifies whether a negative sample is more similar or less similar than expected given its spatial distance. We hypothesize that hard negatives are samples with large spatial distance and $\delta(i, j) < 0$, and that false negatives are samples with small spatial distance and $\delta(i, j) < 0$. Based on this hypothesis, we define the weight w_{ij}:

$$w_{ij} = \begin{cases} \exp\left(-\delta(i, j)/s_1\right), \text{ if } \delta(i, j) < 0 \text{ and } d(i, j) > \theta_1 & \text{(hard negative)} \\ \exp\left(\delta(i, j)/s_2\right), \text{ if } \delta(i, j) < 0 \text{ and } d(i, j) < \theta_2 & \text{(false negative)} \\ 1, \text{ otherwise} \end{cases} \tag{8}$$

where s_1, s_2 are two hyperparameters used to scale $\delta(i, j)$ for the sake of numerical stability. Finally, we reweight the loss function:

$$\mathcal{L}_i = -\log \frac{\exp(V_i \cdot L_i/\tau)}{\exp(V_i \cdot L_i/\tau) + \sum_{L_j^- \in \mathcal{N}_i} \exp(w_{ij}(V_i \cdot L_j^-/\tau))} \tag{9}$$

Intuitively, Equation 8 emphasizes (with weights > 1) the cases when far apart observations look more similar than they should – these are the negative samples that are difficult to distinguish visually, i.e. hard negatives; it also tolerates (with weights < 1) the cases when near observations look more similar than they should – both their distance and image similarity indicate that they are unlikely negative samples.

4 Experiment

Dataset: We evaluate our method on the OpenStreetView-5M (OSV5M) dataset [1], a large-scale benchmark for image-based geo-localization. OSV5M contains around 5 million geo-tagged street-view images sourced from Mapillary. We estimate the empirical semivariogram by computing pairwise cosine distances of image embeddings processed by frozen CLIP image encoder, against their corresponding haversine distances. Shown in Figure 2, illustrates a clear spatial dependency: image dissimilarity increases with increasing geographic distance up to a certain range, beyond which the spatial distance doesn't contribute to the feature distance anymore.

Evaluation: As a retrieval-based approach, GeoCLIP constructs a GPS gallery for retrieval. The goal is to retrieve the corresponding coordinates from the gallery for a given image. The model computes cosine similarities between the query image embedding and all GPS gallery embeddings, and the coordinate with the highest similarity is selected as the predicted location. Following the common metrics, which evaluate the retrieval accuracy at different levels: 25km, 200km, and 750km. A prediction is considered correct if it falls within the specified threshold radius.

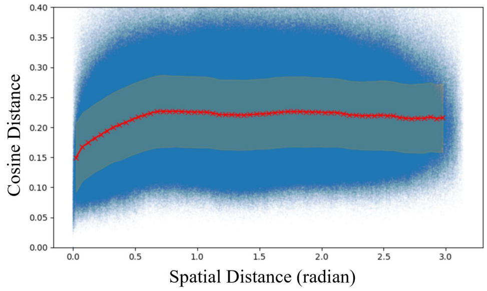

Figure 2: Embedding Semivariogram of OSV5M. Feature distance increases with the spatial distance until a certain range.

Results and Discussion: We evaluate the effectiveness of our proposed strategy by integrating it into the GeoCLIP framework. As reported in Table 1, our method consistently outperforms the original GeoCLIP across all evaluation scales. It achieves 52.1% and 21.5% accuracy at the region (200km) and city level (25km), improving upon GeoCLIP's 50.1% and 19.8%. These gains in fine-grained highlight the benefit of incorporating spatial information into contrastive learning.

Compared to other strong baselines (results from [5]), our method achieves the highest region and city level accuracy, while models like RFM perform 4% better at country level, our method maintains competitive global accuracy while achieving four times higher at city scale. This demonstrates that our semivariogram-guided penalty successfully manage a balance in both coarse and fine level performance.

Table 1: Comparison of geolocalization accuracy on OSV-5M.

Method	Accuracy ↑ (in %)		
	City(25km)	Region(200km)	Country(750km)
ISNs [11]	4.2	39.4	66.8
Hybrid [1]	5.9	39.4	68.0
SC Retrieval [6]	19.9	45.8	73.4
vMF [7]	0.6	17.2	52.7
RFM $\mathbb{S}_2$[5]	5.4	44.2	**76,2**
GeoCLIP [17]	19.8	50.1	71.4
Ours	**21.5**	**52.1**	72.1

5 Conclusion

In this work, we proposed a spatially aware strategy for contrastive learning in geo-localization. By explicitly modeling geographic information using a generalized semivariogram, we provided a perspective to discover hard negatives and false negatives and address two critical challenges in contrastive learning for localization. Integrated into the GeoCLIP framework, our method achieves improvements at fine-grained scales, notably exceeding the original GeoCLIP and other state-of-the-art baselines at the city and region levels, without losing accuracy at the coarse level. These results demonstrate the value of incorporating spatial dependency into representation learning.

Our findings suggest that geographically-aware training can significantly enhace GeoAI models. A promising direction for future research is systematically investigate how geospatial training strategies contribute to the performance and generalization of geospatial artificial intelligence.

References

[1] Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, Constantin Aronssohn, Nacim Bouia, Stephanie Fu, Romain Loiseau, Van Nguyen Nguyen, Charles Raude, Elliot Vincent, et al. 2024. Openstreetview-5m: The many roads to global visual geolocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21967–21977.

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

[3] Fabian Deuser, Konrad Habel, and Norbert Oswald. 2023. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16847–16856.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[5] Nicolas Dufour, Vicky Kalogeiton, David Picard, and Loic Landrieu. 2025. Around the World in 80 Timesteps: A Generative Approach to Global Visual Geolocation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 23016–23026.

[6] Lukas Haas, Silas Alberti, and Michal Skreta. 2023. Learning generalized zero-shot learners for open-domain image geolocalization. *arXiv preprint arXiv:2302.00275* (2023).

[7] Mike Izbicki, Evangelos E Papalexakis, and Vassilis J Tsotras. 2020. Exploiting the earth's spherical geometry to geolocate images. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*. Springer, 3–19.

[8] Hao Li, Fabian Deuser, Wenping Yin, Xuanshu Luo, Paul Walther, Gengchen Mai, Wei Huang, and Martin Werner. 2025. Cross-view geolocalization and disaster mapping with street-view and VHR satellite imagery: A case study of Hurricane IAN. *ISPRS Journal of Photogrammetry and Remote Sensing* 220 (2025), 841–854.

[9] Georges Matheron. 1963. Principles of geostatistics. *Economic geology* 58, 8 (1963), 1246–1266.

[10] Niluthpol Chowdhury Mithun, Kshitij S Minhas, Han-Pang Chiu, Taragay Oskiper, Mikhail Sizintsev, Supun Samarasekera, and Rakesh Kumar. 2023. Cross-view visual geo-localization for outdoor augmented reality. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, 493–502.

[11] Eric Muller-Budack, Kader Pustu-Iren, and Ralph Ewerth. 2018. Geolocation estimation of photos using a hierarchical model and scene classification. In *Proceedings of the European conference on computer vision (ECCV)*. 563–579.

[12] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). 8748–8763.

[14] Bojan Šavrič, Tom Patterson, and Bernhard Jenny. 2019. The equal earth map projection. *International Journal of Geographical Information Science* 33, 3 (2019), 454–465.

[15] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. 2020. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems* 33 (2020), 7537–7547.

[16] Waldo R Tobler. 1970. A computer movie simulating urban growth in the Detroit region. *Economic geography* 46, sup1 (1970), 234–240.

[17] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. 2023. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems* 36 (2023), 8690–8701.

[18] Nam Vo, Nathan Jacobs, and James Hays. 2017. Revisiting im2gps in the deep learning era. In *Proceedings of the IEEE international conference on computer vision*. 2621–2630.

[19] Zhangyu Wang, Gengchen Mai, Krzysztof Janowicz, and Ni Lao. 2024. MC-GTA: Metric-constrained model-based clustering using goodness-of-fit tests with autocorrelations. *arXiv preprint arXiv:2405.18395* (2024).

[20] Tobias Weyand, Ilya Kostrikov, and James Philbin. 2016. Planet-photo geolocation with convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer, 37–55.