OpenStreetView-5M: The Many Roads to Global Visual Geolocation

Guillaume Astruc* [1,2,5] Nicolas Dufour* [1,6] Ioannis Siglidis* [1] Constantin Aronssohn [1]
Nacim Bouia [1] Stephanie Fu [1,4] Romain Loiseau [1,2] Van Nguyen Nguyen [1]
Charles Raude [1] Elliot Vincent [1,3] Lintao XU [1] Hongyu Zhou [1] Loic Landrieu[1]

[1] LIGM, Ecole des Ponts, CNRS, UGE [2] UGE, IGN, ENSG, LASTIG [3] Inria Paris [4] UC Berkeley
[5] CESBIO, Univ de Toulouse, CNES/CNRS/IRD/INRAE/UPS [6] LIX, CNRS, Ecole Polytechnique, IP Paris

Abstract

Determining the location of an image anywhere on Earth is a complex visual task, which makes it particularly relevant for evaluating computer vision algorithms. Yet, the absence of standard, large-scale, open-access datasets with reliably localizable *images has limited its potential. To address this issue, we introduce OpenStreetView-5M, a large-scale, open-access dataset comprising over* 5.1 *million geo-referenced street view images, covering* 225 *countries and territories. In contrast to existing benchmarks, we enforce a strict train/test separation, allowing us to evaluate the relevance of learned geographical features beyond mere memorization. To demonstrate the utility of our dataset, we conduct an extensive benchmark of various state-of-the-art image encoders, spatial representations, and training strategies. All associated codes and models can be found at* `github.com/gastruc/osv5m`.

1. Introduction

While natural image classification is the standard for evaluating computer vision methods [11, 49, 59], global geolocation offers a compelling alternative task. In contrast to classification, where the focus is often a single object, geolocation involves detecting and combining various visual clues, like road signage, architectural patterns, climate, and vegetation. Predicting a single GPS coordinate or location label from these observations necessitates a rich representation of both the Earth's culture and geography; see Figure 1 for some examples. Furthermore, the abundance of geo-tagged street-view images depicting complex scenes with a clear and consistent point of view makes this task appropriate for training and evaluating modern vision models.

Despite this potential, few supervised approaches are trained and evaluated for the task of geolocation. We attribute this to the limitations of existing geolocation datasets: (i) Large and open geolocation datasets contain a significant portion of noisy and non-localizable images [19, 25, 60]; (ii) Street view datasets are better suited for the task but are both proprietary and expensive to download [7, 10, 16, 18, 33, 53]. To address these issues, we introduce OpenStreetView-5M (OSV-5M), an open-access dataset of 5.1 million high-

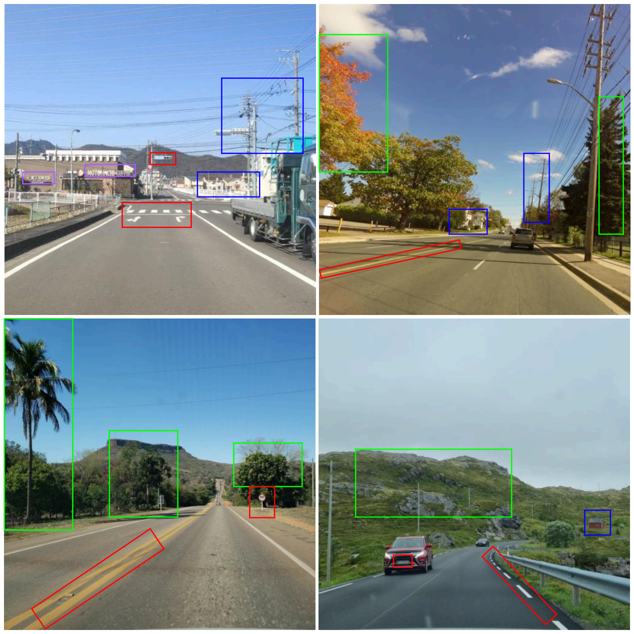

drivephotograph, and_eng, gclem, bootprint, Mapillary, licensed under CC-BY-SA.

Figure 1. **Global Visual Geolocation.** Predicting the location of an image taken anywhere in the world from just pixels requires detecting a combination of clues of various abstraction levels [36]. Can you guess where these images were taken?[1]

*Denotes equal contributions.

[1] **From top left to bottom right:** Nagoya, Japan; Ontario, Canada; Mato Grosso, Brazil; Lofoten, Norway.

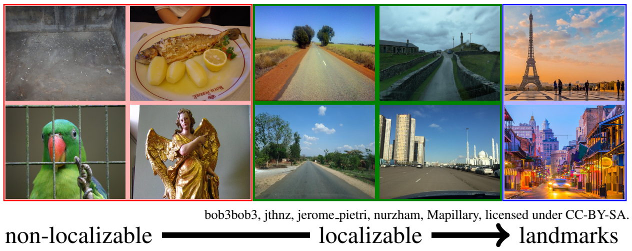

bob3bob3, jthnz, jerome_pietri, nurzham, Mapillary, licensed under CC-BY-SA.
non-localizable
localizable
landmarks

Figure 2. **Localizable vs Non-Localizable.** Images from our dataset (green) occupy the space between weakly localizable images (red) like the ones from the test set of Im2GPS3k [60] and landmark images used to advertise CV conferences (blue).

quality and crowd-sourced street view images. Our ambition is to make both street view images and global geolocation new standards for measuring progress in deep learning.

Automating visual geolocation has significant potential benefits, with direct applications in fields such as journalism, forensics, as well as historical and cultural studies. Learning robust geographical representations may also be valuable for various deep learning challenges, including self-supervised learning and generative modeling, or the development of more interpretable AI systems. Thanks to its size and scope, and its strict train/test split, OSV-5M serves as a robust and reliable benchmark for computer vision models. To demonstrate this, we design an extensive evaluation experiment to measure the impacts of various factors such as pretraining strategies, model scale, spatial representations, fine-tuning approaches, contrastive losses, and auxiliary tasks.

2. Related Work

In this section, we detail the notion of image localizability (Section 2.1), the main existing geolocation datasets (Section 2.2), and geolocation methods (Section 2.3).

2.1. Localizability

As noted by Izbicki et al. [25], images exhibit a range of localizability, an inherently perceptual concept, see Figure 2. Non-localizable images lack information that connects them to a specific location or are of too low quality to properly analyze. Weakly localizable images only contain vague or indirect hints, such as people, animals, and objects in indoor scenes. Localizable images should contain enough information to allow for an informed guess relative to their location. For example, street view images are generally localizable as they typically contain salient features indicative of the local environment such as climate, nature, architecture, or utility and regulatory infrastructure. At the far end of the spectrum, landmark images showcase emblematic monuments or iconic landscapes, making their location instantly identifiable to most viewers.

According to this criteria, a visual inspection suggests that 35% of the images in Im2GPS3k, a dataset commonly used to benchmark geolocation methods [60], are non-localizable. When used for evaluation, this may lead to unreliable errors or promote methods that have memorized biases of the training distribution. When used for training, non-localizable images can lead to sub-optimal representations or encourage spurious correlations. OSV-5M predominantly comprises localizable street view images whose accurate geolocation requires robust geographical representations.

2.2. Geolocation Datasets

We motivate the need for OSV-5M by reviewing existing geolocation datasets from the two main sources of geotagged images: web-scraped and street view images, see Table 1.

Web-Scraped. Image hosting platforms like Flickr provide a near-endless source of geotagged images, which has been used to create large open datasets, like YFC100M [56]. Most images correspond to personal or amateur photographs representing food, art, and images of pets and friends, and are either weakly or non-localizable. Even strongly localizable images are typically taken in tourist spots, injecting an often Western cultural bias towards recognizable landmarks [29]. The provided location metadata can be occasionally missing or inaccurate, and the online nature of these images implies they can be easily removed, hindering reproducibility[2]. For evaluation purposes, cleaner subsets have been proposed that improve both the image distribution coverage and annotation quality [54, 60], but remain still heavily biased and predominantly non-localizable. Despite their small scope and size, these datasets are currently the primary means of evaluating geolocation models.

Street View. Conversely, street view images tend to be strongly localizable. Captured through panoramic cameras or dash-cams, they depict in high quality a vehicle's surroundings, which corresponds mostly to outdoor scenes with rich geographical cues. Google famously provides a global street view coverage, which is, however, expensive to acquire for academic purposes ($1000 for 150k images) and cannot be shared. Existing open datasets from this source either only consist of dense samples from 3 US cities meant for navigation [38, 67], or are inaccessible [10, 16, 33].

Luckily, crowd-sourced platforms such as Mapillary [4] offer a global and diverse source of open-access street view images for various environments, from dense cities and suburbs to remote and inhabited landscapes. These images have been used to construct several benchmarks for multiple tasks other than geolocation, including depth estimation [6], semantic segmentation [41], traffic sign detection and classification [14], place recognition [61] and visual localization [26]. With 5.1M Mappilary images taken across the globe, OSV-5M is the largest open-access street-view image dataset

[2] 60% of the 2014 YFCC-split [39] was deleted by 2020 [25]!

Table 1. **Geolocation Datasets.** OpenStreetView-5M contains strongly localizable street views with access, scope, and size comparable to web-scraped databases.

Image Source	size	open-access	scope
Web-scraped			
Im2GPS [19]	237	✔	biased
Im2GPS3k [60]	2997	✔	biased
YFCC4k [60]	4536	✔	biased
YFCC26k [54]	26k	✔	biased
MP-16 [32]	4.7M	✔	biased
Moussely *et al.* [39]	14M/6M[2]	✔	global
YFCC100M [56]	100M	✔	biased
PlaNet [62]	125M	✗	biased
Street view			
Google-WS-15k [10]	15k	✗	global
GMCP [67]	105k	✗	3 cities
StreetCLIP [16]	1M	✗	unknown
OpenStreetView-5M	5.1M	✔	global

and the only one designed for global geolocation. OSV-5M has a similar order of magnitude to popular YFCC-based geolocation train sets [32, 39], and comes with a clean test set that is 33 times bigger than the current largest street-view image test benchmark [10] (which is not openly accessible).

2.3. Geolocation Methods

Place recognition [68] and visual localization [12, 30, 44, 45, 50] are popular tasks that consist in finding the pose of images in a known scene. In contrast, visual geolocation predicts 2D coordinates or discrete locations (*e.g.*, countries), and aims for lower accuracy and the ability to generalize to unseen areas [20]. Existing geolocation approaches can be categorized by whether they treat geolocation as an image retrieval problem, a classification problem, or both.

Image Retrieval-Based Approaches. A straightforward method for image localization is to find the most similar image in a large image database and predict its location [19]. The first successful approaches involved retrieving the nearest image in a space of handcrafted features such as color histograms [19], gist features [42], or textons [35]. It was later improved with SIFT features and support vector machines [21]. Deep features further boosted the performance of these approaches [60]. While such models typically exhibit high performance given a large and dense enough image database, they do not involve representation learning. Consequently, unless provided with robust features, they may perform poorly in sparsely represented or dynamically changing environments.

Classification-Based Approaches. Geolocation can also be approached as a classification problem by discretizing latitude and longitude coordinates. The choice of partition is critical, ranging from regular [62], adaptive [10], semantic-driven [55], combinatorial [52], administrative [17, 46], and hierarchical [10, 60] partitions. Classification-based methods must strike a delicate balance between the quantity and size of cells; if the discretization is too coarse, the performance will be limited, while too many small cells may not have enough samples for learning-based methods. Furthermore, a typical classification loss such as cross-entropy does not incorporate the distance between regions: confusing two adjacent cells is equivalent to mistaking the continent.

Hybrid Approaches. Retrieval and classification approaches can be combined to overcome the limitations of discretization. This can be achieved using ranking losses [60] or contrastive objectives [31]. Haas *et al.* [17] follow a classification-then-regression approach based on prototype networks. Finally, Izbicki *et al.* [25] go beyond single-location prediction by estimating probability distributions based on spherical Gaussians.

3. OpenStreetView-5M

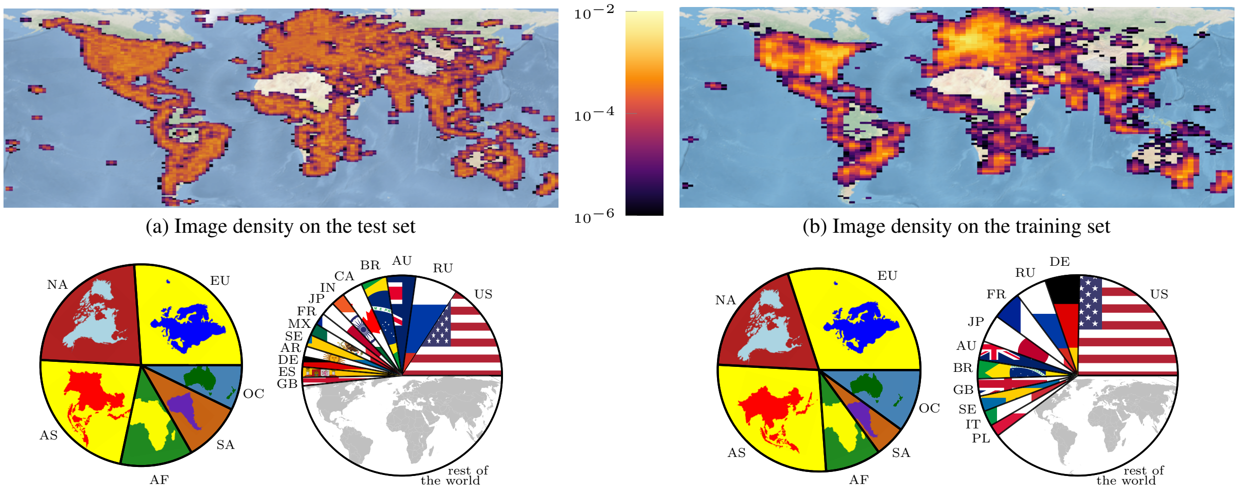
OpenStreetView-5M establishes a new open benchmark for geolocation by providing a large, open, and clean dataset. The Appendix details the construction of the dataset. As detailed below, OpenStreetView-5M improves upon several limitations of current geolocation datasets.

Scale. Deep neural networks have historically been selected over other machine learning methods because they benefit from larger amounts of data. OSV-5M consists of 4,894,685 training and 210,122 test images, with a height of 512 pixels and an average width of 792 ± 127 pixels.

Scope. Many geolocation datasets are restricted to a few cities [38, 67] or are significantly biased towards the Western world [29]. In contrast, OpenStreetView-5M images are uniformly sampled on the globe, covering 70k cities and 225 countries and territories, as shown in Figure 3. The distribution of test images across countries has a normalized entropy of 0.78 [63, Eq. 19], suggesting high diversity. Our train set has a normalized entropy of 0.67, which is comparable to the entropy of the distribution of the countries' area (0.71).

Access. OpenStreetView-5M is based on the crowd-sourced street view images of Mapillary [4] which follow the CC-BY-SA license: free of use with attribution [2].

Quality Evaluation. We estimate through manual inspection of 4500 images that 96.1% (±0.57%) of the images in the OpenStreetView-5M dataset are localizable, with a 95% confidence level [24, Chap. 8]. Among the weakly or non-localizable images, 70% (2.7% total) are low-quality: under- or over-exposed, blurry, or rotated; 30% (1.2% total) are poorly framed, indoor, or in tunnels.



(a) Image density on the test set

(b) Image density on the training set

(c) Continent and country distributions of the test set

(d) Continent and country distributions of the training set

Figure 3. **OpenStreetView-5M.** Image density and proportions per country and continent for the train and test sets. To ensure an unbiased evaluation, we prioritize the uniformity of the test set's distribution across the globe over the training set distribution.

Spatial Separation. Without carefully enforcing the spatial separation between train and test images, geolocation can reduce to place-recognition. As our goal is to assess the capacity of models to learn robust geographical representations, we ensure that no image in the OSV-5M training set lies within a 1km radius of any image in the test set.

Sequence Separation. Street-view images are typically acquired by a limited number of camera sensors mounted on the top or front of a small fleet of vehicles assigned to a given region. This correlation between location, cars, and sensors can be exploited to simplify the geolocation task. Notoriously, players of the web-based geolocation game GeoGuessr [3] can locate images from Ghana by spotting a piece of duct tape placed on the corner of the roof rack of the Google Street View car [5]. OpenStreetView-5M tries to avoid this pitfall by ensuring that no image sequence (a continuous series of images acquired by the same user) appears in both training and test sets. While this might not prevent images taken with the same vehicle on different days from being in both sets, it limits such occurrences.

Metadata. Rich metadata beyond geographical coordinates can improve the robustness and versatility of geolocation models. Each image in our dataset is associated with four tiers of administrative data: country, region (*e.g.*, state), area (*e.g.*, county), and the nearest city [6]. Note that areas are not defined for one-third of the dataset. We also associate each image with a set of additional information: land cover, climate, soil type, the driving side, and distance to the sea where the image was taken. See the Appendix for more details on these attributes.

4. Benchmark

We use OSV-5M to benchmark supervised deep learning approaches in the context of visual geolocation. We first present our evaluation metrics (Section 4.1) and framework (Section 4.2). We then explore several design choices, starting with the image encoder backbone (Section 4.3), the prediction objective (Section 4.4), the fine-tuning strategy (Section 4.5), and the choices of contrastive losses (Section 4.6). In each experiment, we select the top-performing designs and integrate them into a *combined model*, which we evaluate and analyze in Section 4.7.

4.1. Evaluation Metrics.

We denote the space of images by $\mathcal{I}$ and the span of longitude and latitude coordinates by $\mathcal{C} = [-180, 180] \times [-90, 90]$. Our objective is to design a model that maps an image from $\mathcal{I}$ to its corresponding location in $\mathcal{C}$. We measure the accuracy of predicted location across geolocation models with three complementary sets of metrics:

- *Haversine distance* [58] δ, between predicted and ground truth image locations;
- *Geoscore*, based on the famous GeoGuessr game [3], defined as $5000 \exp(-\delta/1492.7)$ [17];
- Accuracy of predicted locations across administrative boundaries: country, region, area, and city.

While the average distance between predictions and ground truth is sensitive to outliers (*i.e.*, a few poor predictions can significantly undermine an otherwise high-performing algorithm), the accuracy metric based on administrative borders can avoid this issue. However, this metric

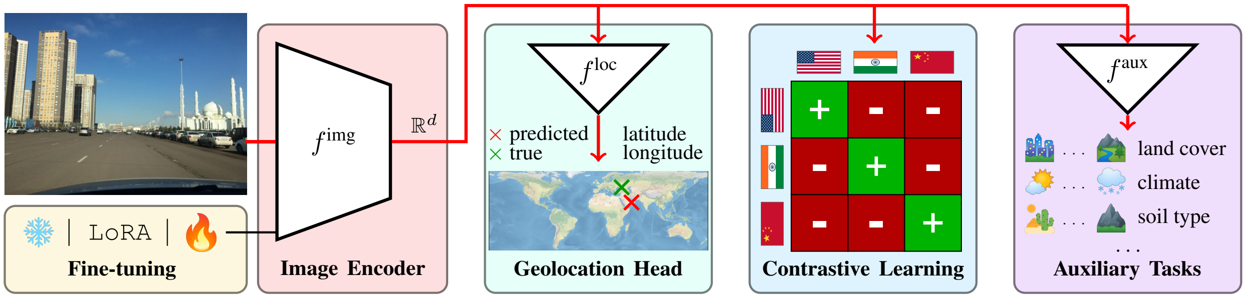


Figure 4. **Visual Geolocation Model.** We propose a simple and versatile framework for visual geolocation and explore the impact of various components of this approach in train-test performance on OpenStreetView-5M. Starting from the left, the input image is converted to a vector representation by an image encoder f^{img} (red). Then a geolocation head f^{loc} maps this vector to a set of geographical predictions (mint). Then a contrastive objective is potentially added (cyan), as well as auxiliary targets to learn better representations for geolocation (lila). We also consider various parameter fine-tuning strategies for training our image encoder, by freezing all or part of f^{img} (yellow).

Table 2. **Impact of Image Encoder.** Several pretrained backbones are evaluated in OpenStreetView-5M. We outline the influence of various architectures, pretraining strategies, and datasets. Best scores are highlighted in **bold**. We denote closed datasets with †.

	Architecture	Size ($\times 10^6$)	Pretraining		Train. time (in h)	Geoscore ↑	Distance ↓	Classification accuracy ↑			
			Objective	Dataset				Country	Region	Area	City
1	ViT-B-32	88	CLIP	LAION-2B	22	2052	2992	35.7	7.0	0.5	0.3
2	ResNet50	23	Classification	ImageNet-1k	45	1260	4171	20.8	3.0	0.2	0.1
3	ViT-L-14	300	DINOv2	DINOv2†	316	2530	2233	46.9	10.7	0.7	0.3
4	ViT-L-14	300	CLIP	LAION-2B	206	2474	2358	44.8	10.6	0.8	0.2
5	ViT-L-14	300	CLIP	DATA_COMP	206	2719	1964	50.6	12.8	1.0	0.4
6	ViT-L-14	300	CLIP	Meta-CLIP	206	2724	1888	49.7	12.7	1.1	0.4
7	ViT-L-14	300	CLIP	OpenAI†	206	2888	1688	53.3	14.6	1.2	0.5
8	ViT-L-14	300	StreetCLIP	OpenAI† + GSV†	206	**3028**	**1481**	**56.5**	**16.3**	**1.5**	**0.7**
9	ViT-bigG-14	1800	CLIP	LAION-2B	900	2878	1766	53.4	15.0	1.3	0.5

can be too lenient for large divisions or arbitrarily punitive for small ones. The Geoscore offers a compromise by rewarding precise predictions without being overly sensitive to large but rare errors.

4.2. Framework

The models evaluated in this benchmark follow a consistent architecture, represented in Figure 4. All considered networks contain the two following modules:

- the image encoder $f^{\text{img}} : \mathcal{I} \mapsto \mathbb{R}^d$, which maps an image to a d-dimensional vector;
- the geolocation head $f^{\text{loc}} : \mathbb{R}^d \mapsto \mathcal{C}$, which maps this vector to geographic coordinates.

Implementation details. Unless stated otherwise, f^{img} is always a pretrained and frozen CLIP ViT-B/32 model [48] with $d = 768$ and f^{loc} is a Multilayer Perceptron (MLP) with GroupNorms [65]. This base model directly regresses geographical coordinates and uses the L_1 norm as loss function. The model is trained with a batch size of 512 images for 30 epochs (260k iterations) with a fixed learning rate of 2×10^{-4}. Throughout the paper we will denote in blue the frozen base model, in orange its fine-tuned version, and in green the model combining all top-performing designs.

4.3. Image Encoder

We first benchmark various architectures for the image encoder module f^{img}, with varying backnones, and pretraining strategies and datasets:

- *Architecture.* We test a standard ResNet50 [22], and modern ViTs [13] of multiple sizes (B-32, L-14, and bigG-14).
- *Pretraining.* We consider different types of pretraining objectives, including classification on ImageNet, self-supervized pretraining DINOv2 [43], text supervision CLIP [48], as well as StreetCLIP [16], which is finetuned specifically for geolocation.
- *Dataset.* We consider several pretraining datasets, including LAION-2B [51], DATA_COMP [15], Meta-CLIP [66], and the proprietary datasets of DINOv2, OpenAI, and StreetCLIP [16].

Analysis. Our experimental results are presented in Table 2. Here, we summarize several key takeaways:

- *Model Size.* As shown in Rows 1, 2, 4, and 9 of Table 2, there is a direct correlation between the size of the image encoder and its geolocation performance. The large ViT, bigG-14 model with 1.8 billion parameters (Row 9) improves significantly on the performance of its smaller versions. As

Table 3. **Prediction Modules.** We report the performance of various prediction models and objectives. QuadTrees, hierarchical supervision, and hybrid models all significantly improve on direct regression or classification with administrative borders. We underline the accuracy for divisions that the method is specifically trained to categorize.

		Number classes	Geo ↑ score	Dis ↓ tance	Classification accuracy ↑			
					country	region	area	city
Reg.	Coord.	-	2052	2992	35.7	7.0	0.5	0.3
	Sin/cos	-	1192	4797	13.6	2.1	0.1	0.0
Classification	Country	222	2263	2981	56.3	-	-	-
	Region	2.8k	2683	2858	57.0	30.2	-	-
	Area	9.3k	1935	4454	36.3	19.7	8.8	-
	City	69.8k	2600	3217	52.2	28.5	7.3	4.9
	+ hierarchy	69.8k	2868	2768	58.2	34.3	**9.6**	**6.0**
	QuadTree	11.0k	2772	2832	54.8	27.7	5.4	2.8
	+ hierarchy	11.0k	2890	2654	57.4	29.9	5.9	2.9
Hybrid		11.0k	**3036**	**2518**	**60.8**	**36.3**	9.5	5.7

the size of models correlates with their training time, we select ViT-L-14 as the best compromise.

- *Pretraining.* As seen in rows 3, 7, and 8, CLIP pretraining leads to better results than DINO or image classification. We thus focus on the latter for further comparisons.
- *Dataset.* Rows 4 to 8 show the significant impact of the choice of pretraining datasets. The geolocation-oriented StreetCLIP (row 8) leads to the best results, followed by OpenAI's CLIP (row 7). However, both datasets are not open access. We choose DATA_COMP (row 5) as the best open-source dataset for its slightly better country classification rate compared to Meta-CLIP (row 6).

4.4. Prediction Head

We examine three different possible supervision schemes for the geolocation head f^{loc}: regression, classification (including hierarchical classification), and a hybrid approach.

Regression. We start with the most straightforward approach: f^{loc} directly regresses coordinates in $\mathcal{C}$. We train an MLP supervised with the L_1 loss between true and predicted coordinates. To account for the periodicity of the latitude, we also test an approach where we regress instead the cosine and sine of the longitude and latitude and then recover the real coordinates with trigonometry [34].

Table 4. **Parameter Fine-tuning Strategies.** We compare the performance of different parameter fine-tuning strategies, in terms of performance, number of parameters, and training time.

	Param. (M)	Train. time	Geo ↑ score	Dis ↓ tance	Classification accuracy ↑			
					country	region	area	city
Frozen	0.6	22	2052	2992	35.7	7.0	0.5	0.3
LoRA-32	2.4	44	2101	2760	36.7	6.4	0.4	0.0
Last block	7.7	26	2587	2372	46.7	12.9	1.0	0.5
Fine-tuning	88.0	132	**2893**	**2085**	**54.9**	**19.1**	**1.6**	**0.8**

Classification. We divide the train set into a set $\mathcal{K}$ of K *divisions*, such as countries, regions, areas, and cities, which amount to $K = 222$, 2.8k, 9.3k, and 69.8k, respectively. As some administrative borders can have vastly different sizes, we also consider an adaptive partition with a QuadTree of depth 10 and maximum leaf size of 1000, corresponding to 11k cells. We then train a classifier $f^{\text{classif}} : \mathbb{R}^d \mapsto \mathbb{R}^K$ which maps an image representation to the probability that the image was taken in each division. Then, to predict the final geographic location, we define f^{lookup}, which associates each division with the average location of its training images: $f^{\text{lookup}} : \mathcal{K} \mapsto \mathcal{C}$. The predicted geolocation can be summarized as: $f^{\text{loc}} = f^{\text{lookup}} \circ \arg\max f^{\text{classif}}$.

In our implementation, f^{classif} is an MLP trained with cross-entropy, while f^{lookup} is a look-up table obtained directly from the training set.

Hierarchical Supervision. We can exploit the nested nature of the administrative divisions and QuadTree cells to supervise all levels simultaneously [40, 60]. More precisely, we predict a probability vector *at the finest resolution* (either city or maximum depth of the QuadTree), which we aggregate recursively to obtain predictions at all levels. We can now supervise with a cross-entropy term for each level.

Hybrid Approach. Inspired by approaches that combine both classification and retrieval [17, 60], we perform regression and classification in a two-step approach. Given the output of our QuadTree classifier $f^{\text{classif}} : \mathbb{R}^d \to \mathbb{R}^K$, we define $f^{\text{relative}} : \mathbb{R}^d \to [-1, 1]^{2K}$ that outputs the relative coordinates of the predicted location inside each cell k. We scale these values such that $(0, 0)$ points to the centroid of the training images in the cell and $[-1, 1]^2$ spans the entire bounding box. Using the cell prediction of the classifier f^{classif} and the relative position from f^{relative}, we can predict the location of the image with sub-cell precision.

We train f^{classif} with the cross-entropy, and f^{relative} with the L_2 loss between the predicted and true relative coordinates on the division that contains the true location.

Analysis. We report the performance of different prediction heads in Table 3, and make the following observations:

- *Regression.* Predicting sines and cosines does not improve the regression model's performance. We hypothesize that this is due to the non-linearity of the trigonometric formula.
- *Classification.* Classification methods generally perform well in Geoscore and starkly improve their respective classification rates, *e.g.* $+23.2\%$ region accuracy for the region classifier compared to the regression model. However, their influence on the average error distance is smaller. Coarse partitions, like countries, are limited by the low precision of f^{lookup}. Inversely, overly refined partitions such as cities lead to a more challenging classification setting where most labels

Table 5. **Contrastive Learning.** We report the impact of adding a contrastive objective to our model, defined by various notions of positive matches between images.

	Pairs	Geoscore ↑	Distance ↓	Classification accuracy ↑			
				country	region	area	city
	no contrastive	2893	2085	54.9	19.1	1.6	0.8
geographic	country	2903	**2005**	**66.8**	13.7	0.7	0.3
	region	**3028**	2131	60.0	**33.3**	2.9	1.0
	area	2376	2886	43.7	18.9	**3.7**	1.2
	city	2912	2209	56.3	24.5	3.2	1.2
	cell	2891	2310	55.9	25.4	3.5	**1.3**
	text-based	2812	2171	66.0	13.0	0.7	0.2

have only a few training examples. QuadTree-constructed labels achieve performance close to the administrative division-based classifier across all levels, *e.g.* 54.8% *vs.* 56.3% for countries and 27.7% *vs.* 30.2% for regions. This compounds into an overall better performance, which shows that adapting the granularity of the label distribution according to the image density appears to be a successful heuristic.

- *Hierarchical & Hybrid.* Supervising on all levels simultaneously significantly improves the prediction. Hybrid methods bridge the gap between classification and regression, yielding high precision without relying on very fine-grained partitions. These results validate the underlying spatial hierarchical nature of geographical data [57]. We select both hybrid and hierarchical designs for the combined model.

4.5. Parameter Fine-tuning

We evaluate different fine-tuning strategies to quantify the impact of learning dedicated features for geolocation. In all configurations, we learn f^{loc} from random weight, and f^{img} is fine-tuned as follows:

- *Frozen.* f^{img} is initialized with pretrained weights and remains frozen.
- *LoRA-32.* We fine-tune f^{img} using Low Rank Adaptation [23] and a rank of 32 (more values in supplementary).
- *Last block.* We unfreeze the last transformer block of f^{img}, responsible for producing the image embedding.
- *Fine-tuning.* We fine-tune all parameters of f^{img}.

Analysis. In Table 4, we report the impact of different fine-tuning strategies. Training only the last transformer block instead of using LoRA leads to a ten times larger Geoscore improvement in only half the training time. This suggests that pretrained models can extract relevant patch embeddings, while image encoding must be significantly adapted for geolocation. Fine-tuning the entire network leads to an even larger improvement but a five-fold increase in training time. However, the resulting performance is comparable to the frozen ViT-bigG-14 shown in Table 2 and trains 9 times faster. We select the fine-tuning configuration as the top-performing approach and denote it in orange.

Table 6. **Combined Model.** We report the improvements brought by each top-performing design choice and their combination and compare them with baselines and competing approaches.

	Geo ↑ score	Dis ↓ tance	Classification accuracy ↑			
			country	region	area	city
Base model	2052	2992	35.7	7.0	0.5	0.3
ViT-L-14 DC	+ 667	- 1028	+14.9	+ 5.8	+0.5	+ 0.1
QuadTree	+ 720	- 160	+19.1	+20.7	+5.4	+ 2.5
Hybrid	+ 264	- 314	+ 6.0	+ 8.6	+4.5	+ 2.9
Hierarchical	+ 118	- 178	+ 2.6	+ 0.2	+0.5	+ 0.1
Fine-tuning	+ 841	- 907	+19.2	+12.1	+1.1	+ 0.5
Region contrast.	+ 135	+ 46	- 5.1	+14.2	+2.1	+ 0.2
Combined model	+1309	- 1178	+32.3	+32.4	+9.8	+ 5.6
	3361	**1814**	**68.0**	**39.4**	10.3	5.9
Random	328	8724	20.0	2.0	0.0	0.0
Human Evaluation	1009	6407	48.9	12.2	3.0	0.0
GeoEstimator [40]	3331	2308	66.8	**39.4**	**18.4**	4.2
StreetCLIP 0-shot [16]	2273	2854	38.4	20.8	9.9	**14.8**

4.6. Contrastive Objectives

Contrastive learning builds positive and negative sample pairs from the training set and pushes representations of positive pairs close to each other while contrasting negative ones [8, 9]. Positive pairs can be formed within the same modality, such as different views of an object, or across modalities, such as images and captions. In the geolocation context, we propose two approaches to construct such pairs:

- *Geographic.* We match images if they are within the same administrative division: countries, regions, areas, cities, or QuadTree cells. We modify the dataloader to ensure each image is part of at least one positive pair. Contrary to Haas *et al.* [16], we use the multi-positive MIL-NCE loss [37] as our contrastive objective to account for images in several positive pairs, *e.g.* in the same country.
- *Text-Based.* Similar to Haas *et al.* [16], we pair each image with a textual description of its location formed as the following string: “An image of the city of $CITY, in the area of $AREA, in the region of $REGION, in $COUNTRY.”.

Analysis. In Table 5, we measure the impact on the fine-tuned model of different approaches for constructing contrastive pairs. We observe a consistent improvement in terms of performance when building positive pairs with regions, which may be the division most likely to present unique and homogeneous visual and cultural identities. In contrast, areas appear to hurt the performance when used contrastively. Overall, contrastive learning yields a much higher country and region classification rate compared to the classification-based approaches of Table 3, suggesting that encouraging geographically consistent representations is advantageous for geolocation. We also observe that using text as a proxy when geographically consistent pairs exist is not beneficial.

4.7. Combined Model

Summarizing our previous exploration and analysis, we combine the most impactful design choices for each experiment

Table 7. **Nearest Neighbors.** We report the performance of nearest neighbor retrieval using different encoders.

	Geo ↑ score	Dis ↓ tance	Classification accuracy ↑			
			country	region	area	city
CLIP-VIT-B32-LAION	2511	3455	49.3	29.6	1.9	13.1
DINOv2	2994	2542	61.1	37.1	22.9	16.4
CLIP-VIT-L14-DATACOMP	3201	2047	64.5	38.4	23.3	16.6
CLIP-VIT-L14-OpenAI	3545	1458	72.8	44.4	27.5	19.3
StreetCLIP	**3597**	**1386**	**73.4**	**45.8**	**28.4**	**19.9**
Combined model	2734	2608	54.9	24.5	13.6	9.4

into a strong geolocation model, denoted in green: ViT-L-14 backbone pretrained on DATA_COMP, QuadTree partition with hybrid prediction and hierarchical supervision, fully fine-tuned with a region-contrastive loss. As shown in Table 6, this model starkly improves on the base model, with an increase of +1309 in Geoscore, an average distance reduced by 45%, and significantly better accuracy at all levels of administrative divisions.

Analysis. In Table 6, we compare the performance of our combined model to a random baseline (select the location of a random image in the training set) and a human performance obtained by asking 80 annotators to guess the locations of the same 50 images randomly sampled from the test set [36].Despite the difficulty of the task, the average annotator's performance is significantly better than chance. Our baseline model, and more substantially our combined model, far surpasses the accuracy of annotators. We also evaluate two state-of-the-art geolocation models: StreetCLIP [16] evaluated in zero-shot using the text string given in Section 4.6, and the GeoEstimator model [40] fine-tuned on our training set. As both models are designed for geolocation, they yield good performance. Owing to its bespoke geocells, GeoEstimator reaches the highest accuracy for area classification, illustrating the benefit of architectures with built-in geographical priors. See the appendix for further experiments, notably on the impact of auxiliary variables.

Nearest Neighbor. We perform retrieval by matching each image from the test set with an image from the train set based on the cosine distance between the features of each image encoder. We perform approximate matching with the FAISS algorithm [28] through the AutoFAISS package [1], without re-ranking [27, 47]. As reported in Table 7, retrieval methods trained through contrastive learning exhibit high performance. However, the supervision of our combined model based on geographic coordinates and cells does not enhance its retrieval performance. In fact, its retrieval score is lower than that of its pretrained image encoder. These findings are consistent with observations that fine-tuning self-supervised models decreases retrieval performance [64].

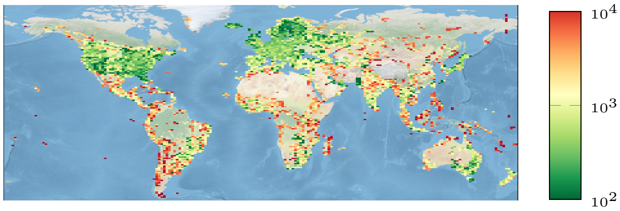


Figure 5. **Spatial Distribution of Errors.** We plot the average prediction error of the combined model in km across the globe.

Error Distribution. We report in Figure 5 a heatmap of the average error distance. Areas sparsely populated with training images, such as South America, have a significantly higher error rate. We report a Pearson correlation coefficient of −0.25 between image density and error, suggesting that image density is not the only factor in the mistakes of our proposed model. See Figure 6 for a visualization of the error distribution. Over half of the combined model's predictions are within 250km of the true image locations.

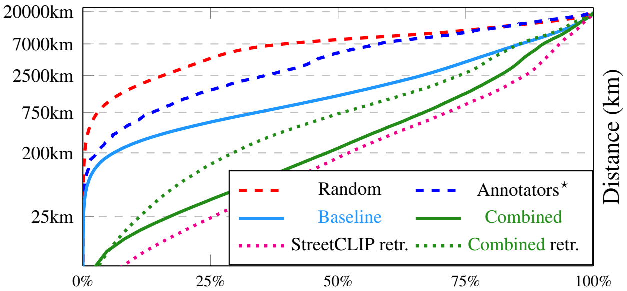


Figure 6. **Error Distribution.** Proportion of predictions within a set distance in the test set. * evaluated on 50 images only.

5. Conclusion

We introduced a new open-access street view dataset of unprecedented size and quality, enabling the consistent training and evaluation of global geolocation models for the first time. Through an extensive experimental framework, we demonstrate that our dataset is a competitive benchmark for developing and evaluating general and bespoke state-of-the-art computer vision approaches for geolocation. Through its scale and quality, we expect OSV-5M to also be useful for self-supervised learning and generative modeling, valuable tasks beyond the scope of visual geolocation.

Acknowledgements. OSV-5M was made possible through the generous support of the Mapillary team, which helped us navigate their vast street view image database. Our work was supported by the ANR project READY3D ANR-19-CE23-0007, and the HPC resources of IDRIS under the allocation AD011014719 made by GENCI. We thank Valérie Gouet for her valuable feedback.

References

[1] AutoFAISS. https://github.com/criteo/autofaiss. Accessed: 2023-10-10. 8

[2] CC BY-SA 2.0 DEED: Attribution-ShareAlike 2.0 Generic. https://creativecommons.org/licenses/by-sa/2.0/deed.en. Accessed: 2023-10-10. 3

[3] GeoGuessr. https://www.geoguessr.com/. Accessed: 2023-10-10. 4

[4] Mapillary. https://www.mapillary.com/. Accessed: 2023-10-10. 2, 3

[5] Plonkit guide to Ghana. https://www.plonkit.net/ghana. Accessed: 2023-10-10. 4

[6] Manuel López Antequera, Pau Gargallo, Markus Hofinger, Samuel Rota Bulò, Yubin Kuang, and Peter Kontschieder. Mapillary planet-scale depth dataset. In *ECCV*, 2020. 2

[7] David M Chen, Georges Baatz, Kevin Köser, Sam S Tsai, Ramakrishna Vedantham, Timo Pylvänäinen, Kimmo Roimela, Xin Chen, Jeff Bach, Marc Pollefeys, et al. City-scale landmark identification on mobile devices. In *CVPR*, 2011. 1

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICLR*, 2020. 7

[9] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005. 7

[10] Brandon Clark, Alec Kerrigan, Parth Parag Kulkarni, Vicente Vivanco Cepeda, and Mubarak Shah. Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes. In *CVPR*, 2023. 1, 2, 3

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Image Net: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[12] Mingyu Ding, Zhe Wang, Jiankai Sun, Jianping Shi, and Ping Luo. CamNet: Coarse-to-fine retrieval for camera relocalization. In *ICCV*, 2019. 3

[13] 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 *ICLR*, 2020. 5

[14] Christian Ertler, Jerneja Mislej, Tobias Ollmann, Lorenzo Porzi, Gerhard Neuhold, and Yubin Kuang. The Mapillary traffic sign dataset for detection and classification on a global scale. In *ECCV*, 2020. 2

[15] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. DataComp: In search of the next generation of multimodal datasets. *NeurIPS Dataset and Benchmark*, 2023. 5

[16] Lukas Haas, Silas Alberti, and Michal Skreta. Learning generalized zero-shot learners for open-domain image geolocalization, 2023. 1, 2, 3, 5, 7, 8

[17] Lukas Haas, Silas Alberti, and Michal Skreta. PIGEON: Predicting image geolocations. *arXiv preprint arXiv:2307.05845*, 2023. 3, 4, 6

[18] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-NetVLAD: multi-scale fusion of locally-global descriptors for place recognition. In *CVPR*, 2021. 1

[19] James Hays and Alexei A Efros. Im2gps: estimating geographic information from a single image. In *CVPR*, 2008. 1, 3

[20] James Hays and Alexei A Efros. Large-scale image geolocalization. *Multimodal location estimation of videos and images*, 2015. 3

[21] James Hays and Alexei A Efros. Large-scale image geolocalization. *Multimodal location estimation of videos and images*, 2015. 3

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[23] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2021. 7

[24] Barbara Illowsky and Susan Dean. Introductory statistics. 2018. 3

[25] Mike Izbicki, Evangelos E Papalexakis, and Vassilis J Tsotras. Exploiting the Earth's spherical geometry to geolocate images. In *MLKDD*, 2020. 1, 2, 3

[26] Ara Jafarzadeh, Manuel López Antequera, Pau Gargallo, Yubin Kuang, Carl Toft, Fredrik Kahl, and Torsten Sattler. CrowdDriven: A new challenging dataset for outdoor visual localization. In *ICCV*. 2

[27] Herve Jegou, Hedi Harzallah, and Cordelia Schmid. A contextual dissimilarity measure for accurate and efficient image search. In *CVPR*, 2007. 8

[28] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 2019. 8

[29] Sebastian Kalkowski, Christian Schulze, Andreas Dengel, and Damian Borth. Real-time analysis and visualization of the YFCC100M dataset. In *Workshop on community-organized multimodal mining: opportunities for novel solutions*, 2015. 2, 3

[30] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A convolutional network for real-time 6-dof camera relocalization. In *ICCV*, 2015. 3

[31] Giorgos Kordopatis-Zilos, Panagiotis Galopoulos, Symeon Papadopoulos, and Ioannis Kompatsiaris. Leveraging EfficientNet and contrastive learning for accurate global-scale location estimation. In *International Conference on Multimedia Retrieval*, 2021. 3

[32] Martha Larson, Mohammad Soleymani, Guillaume Gravier, Bogdan Ionescu, and Gareth JF Jones. The benchmarking initiative for multimedia evaluation: MediaEval 2016. *IEEE MultiMedia*, 2017. 3

[33] Grace Luo, Giscard Biamby, Trevor Darrell, Daniel Fried, and Anna Rohrbach. g^3: Geolocation via guidebook grounding. *Findings of EMNLP*, 2022. 1, 2

[34] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification.

In *CVPR*, 2019. 6

[35] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, 2001. 3

[36] Sneha Mehta, Chris North, and Kurt Luther. An exploratory study of human performance in image geolocation tasks. In *HCOMP 2016 GroupSight Workshop on Human Computation for Image and Video Analysis*, volume 308, 2016. 1, 8

[37] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, 2020. 7

[38] Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, Denis Teplyashin, Karl Moritz Hermann, Mateusz Malinowski, Matthew Koichi Grimes, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, et al. The StreetLearn environment and dataset. *arXiv preprint arXiv:1903.01292*, 2019. 2, 3

[39] Hatem Mousselly-Sergieh, Daniel Watzinger, Bastian Huber, Mario Döller, Elöd Egyed-Zsigmond, and Harald Kosch. World-wide scale geotagged image dataset for automatic image annotation and reverse geotagging. In *ACM multimedia systems*, 2014. 2, 3

[40] Eric Müller-Budack, Kader Pustu-Iren, and Ralph Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. In *ECCV*. 6, 7, 8

[41] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The Mapillary Vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 2

[42] Aude Oliva and Antonio Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 2006. 3

[43] Maxime Oquab, Timothée Darcet, Théo 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. *TMLR*, 2023. 5

[44] Nathan Piasco, Désiré Sidibé, Valérie Gouet-Brunet, and Cédric Demonceaux. Improving image description with auxiliary modality for visual localization in challenging conditions. *International Journal of Computer Vision*, 2021. 3

[45] Noé Pion, Martin Humenberger, Gabriela Csurka, Yohann Cabon, and Torsten Sattler. Benchmarking image retrieval for visual localization. In *3DV*, 2020. 3

[46] Shraman Pramanick, Ewa M Nowara, Joshua Gleason, Carlos D Castillo, and Rama Chellappa. Where in the world is this image? Transformer-based geo-localization in the wild. In *ECCV*, 2022. 3

[47] Danfeng Qin, Stephan Gammeter, Lukas Bossard, Till Quack, and Luc Van Gool. Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. In *CVPR*, 2011. 8

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 5

[49] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. In *NeurIPS Datasets and Benchmarks Track*, 2021. 1

[50] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6-DoF outdoor visual localization in changing conditions. In *CVPR*, 2018. 3

[51] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. 2022. 5

[52] Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. CPlaNet: Enhancing image geolocalization by combinatorial partitioning of maps. In *ECCV*, 2018. 3

[53] Zachary Seymour, Karan Sikka, Han-Pang Chiu, Supun Samarasekera, and Rakesh Kumar. Semantically-aware attentive neural embeddings for image-based visual localization. *arXiv preprint arXiv:1812.03402*, 2018. 1

[54] Jonas Theiner, Eric Müller-Budack, and Ralph Ewerth. Interpretable semantic photo geolocation. In *WACV*, 2022. 2, 3

[55] Jonas Theiner, Eric Müller-Budack, and Ralph Ewerth. Interpretable semantic photo geolocation. In *WACV*, 2022. 3

[56] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 2016. 2, 3

[57] Waldo R Tobler. A computer movie simulating urban growth in the Detroit region. *Economic geography*, 1970. 7

[58] Glen Van Brummelen. *Heavenly mathematics: The forgotten art of spherical trigonometry*. Princeton University Press, 2012. 4

[59] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *CVPR*, 2018. 1

[60] Nam Vo, Nathan Jacobs, and James Hays. Revisiting IMG2GPS in the deep learning era. In *ICCV*, 2017. 1, 2, 3, 6

[61] Frederik Warburg, Soren Hauberg, Manuel Lopez-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *CVPR*, 2020. 2

[62] Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet-photo geolocation with convolutional neural networks. In *ECCV*, 2016. 3

[63] Allen R Wilcox. Indices of qualitative variation. Technical report, Oak Ridge National Lab., Tenn., 1967. 3

[64] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *CVPR*, 2022. 8

[65] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018. 5

[66] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke

Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. *ICLR*, 2024. 5

[67] A.R. Zamir and M. Shah. Image geo-localization based on multiple nearest neighbor feature matching using generalized graphs. 2014. 2, 3

[68] Xiwu Zhang, Lei Wang, and Yan Su. Visual place recognition: A survey from deep learning perspective. *Pattern Recognition*, 2021. 3