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

Supplementary Material

This supplementary material starts by providing further details on the construction and analysis of our dataset OpenStreetView-5M in Section A, showcasing indicative samples in Figure A.1. Then, we provide additional experiments in Section B and qualitative results in Figure B. Finally, Section C further implementation details can be found and Section D outlines our Datasheet [9] for OpenStreetView-5M.

A. OpenStreetView-5M Dataset

OpenStreetView-5M is designed to achieve an open, large-scale, balanced, and global geographical coverage. Through the Mapillary API and the support of the Mapillary team, we gained access to the locations of all 1.8B images [2]. To provide a more manageable and better distributed dataset, we design a specific construction approach, presented in this section. The code to reproduce the treatment can be found at github.com/gastruc/osv5m.

A.1. Construction Approach

Sampling. We start by ensuring that regions with high image density are not disproportionately represented. We define a 100×100m grid across the entire world and randomly choose one image per cell. Then, both the training and test sets are sampled with a weight proportional to the local image density raised to the power of -0.75. Such a strategy balances density-based sampling (which tends to be biased towards urban centers) and area-based sampling (which might favor larger countries). We eliminate images from the test set that are either located within a 1km radius of any train image or share a sequence ID.

Handcrafted Filters. We apply a series of handcrafted filters to remove low-quality images

- *Blurriness.* Blurry images indicate low quality and potentially low localizability. We remove images whose average logarithmic magnitude spectrum is below 120dB.

- *Radiometry.* Certain images hosted on Mapillary are too dark to be meaningfully analyzed, while other have a distinct encoding errors giving them a purple tint. To remove those, we first filter out images whose average brightness (average value over pixels and RGB channels) is below 50. To handle purple images, we remove images for which over 50% of pixels meet the following criteria: $R > 60$ & $G > 60$ & $B < 50$.

- *Exposition.* The exposure of Dash-cam images can be badly exposed, for example, when they face the sun. To filter them, we remove images for which 70% of pixels have a brightness over 250 (overexposed) or under 5 (underexposed).

Rotation-Based Filtering. We perform a learning-based filtering based on a pretrained and frozen RotNet network [10]. This model learns self-supervised image representations by training for the pretext task of predicting a random rotation applied to an input image. Although it it used as a pretext task in the original paper, it becomes useful for filtering out images downloaded from Mapillary's website that are incorrectly rotated. We use the pretrained network to infer the rotation of various images and then use the following filtering strategy depending on RotNet's prediction:

- *0°* **(96% of images)** For normal street view images the cues that signify an absence of rotation are multiple: the sky is up, and cars and pedestrians are upward. We keep these images unchanged.

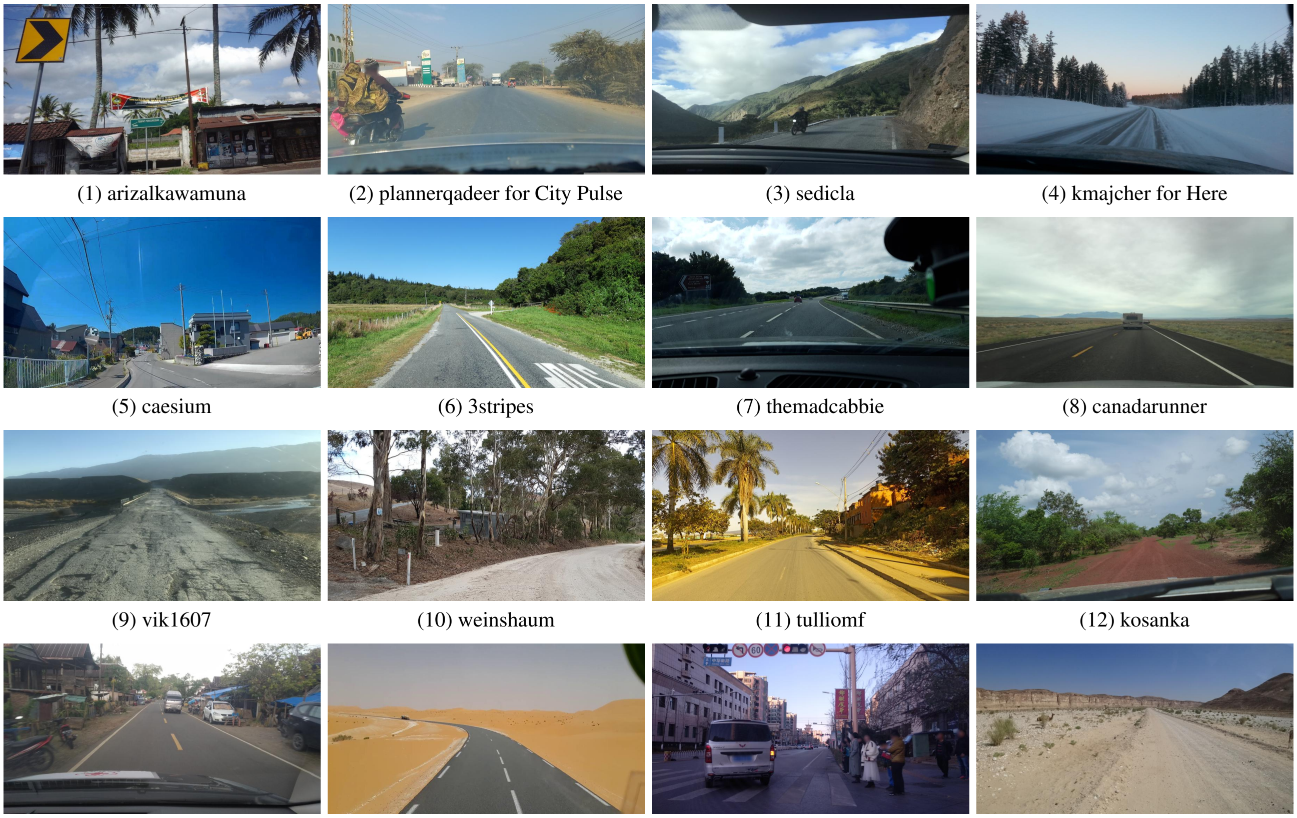
- *180°* **(4%)** Over 90% images predicted to be rotated by 180° are, in fact, actually upside down. We rotate all these images by a half-turn. For the images in the test, we perform an additional visual inspection to remove the small proportion of non-localizable images not removed by the previous filters.

- *90° or 270°* **(0.2%)** Images predicted as rotated by a quarter-turn are in the vast majority taken indoors or in tunnels. We remove all such images from both the train and test set.

A.2. Discussion

Why Not Just Subsample YFCC100M? The wide adoption of YFCC100M, with its nearly 50 million geotagged images,, might question the need for creating yet another geotagged image dataset. However, several compelling reasons justify creating OpenStreetView-5M instead of subsampling YFCC100M:

- *Data Distribution.* The images shared on Flickr do no aim to capture our world in an objective way, but instead focus on aesthetic and cultural value. For example, recognizable landmarks like the Eiffel Tower or the Louvre, are a cultural symbol of the city of Paris, yet they lack any information that is useful in identifying other cities as French or even other streets as Parisian. Additionally, many images are renders or infographics. In contrast, OSV-5M only features dashcam pictures, that offer a consistent front-view perspective, that is more objective as it doesn't focus on something specific, and



(1) arizalkawamuna (2) plannerqadeer for City Pulse (3) sedicla (4) kmajcher for Here

(5) caesium (6) 3stripes (7) themadcabbie (8) canadarunner

(9) vik1607 (10) weinshaum (11) tulliomf (12) kosanka

(13) benjidad (14) cut (15) mapillario (16) vbombaerts

Figure A. **Images from OSV-5M.** The true locations can be found on the next page. The Mapillary users are credited in the subcaptions.

thus may be more beneficial for learning visual geographical representations.

- *Localizability.* From a manual inspection of 1000 images we find that fewer than 10% (±1.3%, 95% confidence) of YFCC100M's images are perceptually localizable. In stark contrast, OSV-5M boosts this perceptual localizability to a rate of 96.1% (±0.57, 95% confidence), making it a more suitable candidate for a standard evaluation benchmark for global geolocation.
- *Geographical Bias.* Images in the YFCC100M dataset exhibit a high cultural bias towards the Western world, with over 35% of images from the US and nearly 70% from North America and Europe [16]. OSV-5M offers a more equitable global representation, as detailed in Figure 2 of the main paper.
- *Selection Challenges.* Subsampling YFCC100M based on metadata alone is ambiguous: 30% of images lack titles, 68% lack descriptions, 30% lack tags, and 50% lack geotagging. The tags "travel" and "nature" cover fewer than 2 million images. Using instead automated selection methods may inadvertently propagate existing biases, such as filtering street views of non-Western countries.
- *Persistence.* As happens with a lot of large research dataset, YFCC comes only as a collection of image URLs that need to be downloaded directly from Flickr. Such a dataset construction approach, even if the only feasible choice for very large datasets, is very volatile and can prevent future reproducibility. For example, 60% of the 2014 YFCC-split [19] was deleted by 2020 [14]. While YFCC100M used to be hosted on Yahoo's Webscope, this option is no longer available [3]. Instead users need to create an AWS account, that requires a credit card to acquire API credentials for downloading the data through a designated S3 bucket [5]. Even if no charge is applied, this setting may be prohibitive for academics or residents of certain countries. Also, due to the sensitive nature of the Flickr data, users need to make a formal request to download the dataset, something that isn't needed for our dataset. Instead, OSV-5M ensures persistence, open and easy access for long-term and broad usage.

To summarize, YFCC100M is a vast and unstructured set of images, a subset of which may be well suited for localization and place recognition. However, the ambiguous localizability, geographical content, metadata, persistence, and access to its images highlight the need for a dedicated

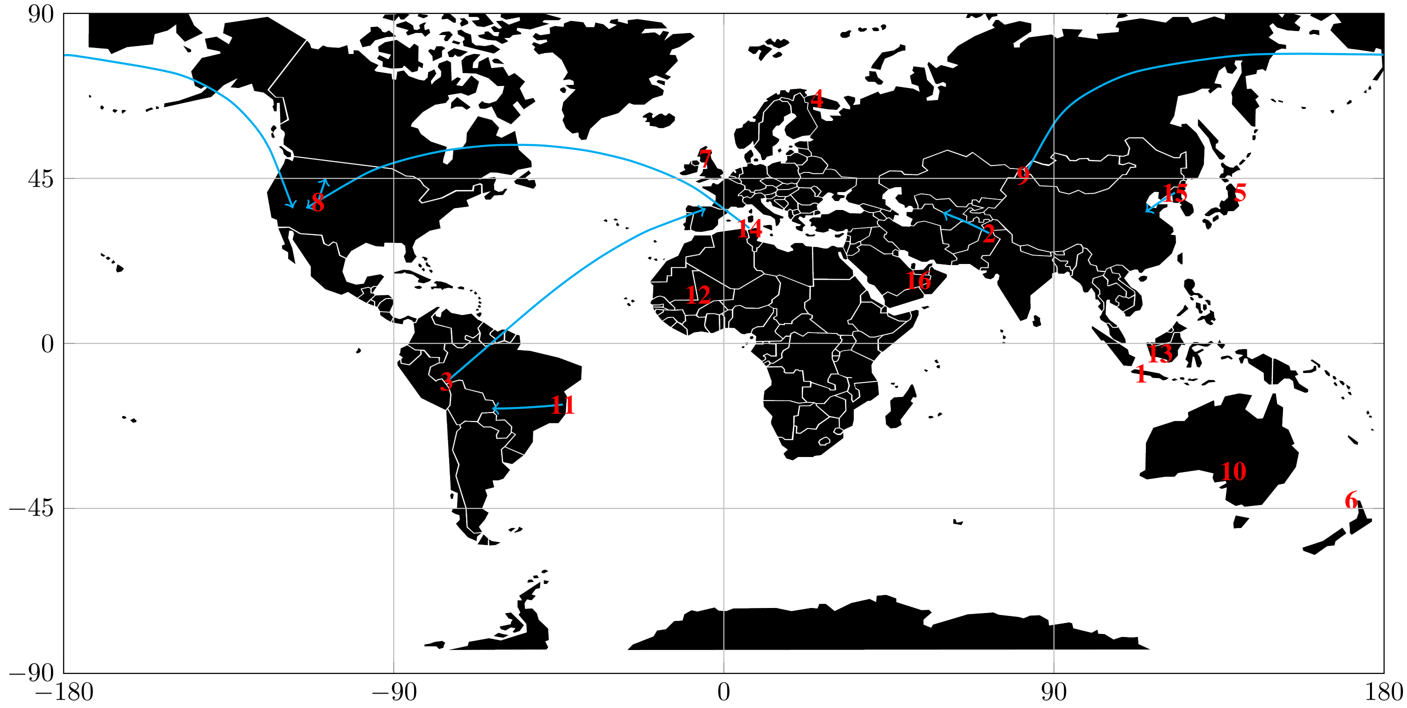


Figure B. **True Locations.** Location of the images of Figure A.1. With blue we visualize errors of the combined model that are superior to 500 km. Most of the images (9 out of 16) are predicted within 500km of where they were taken. We observe that two difficult images (9 and 14) are erroneously mapped to the US, which could be explained by the geographical bias of the training set.

dataset like OSV-5M, specifically designed for the task of global visual geolocation.

Visible GeoTags. Due to the diversity in user input data, we found that a small percentage of images ($< 5\%$) have a visible overlayed text on the bottom part that tags their location. This should be taken into consideration when constructing a benchmark for a future dataset. However, due to the standard ViTs resampling of images to 224×224, these coordinates become indecipherable, as demonstrated in Figure C. We implement for our data loader the option to add a Gaussian blur with a width of 2 to the bottom 14 rows. When training and/or testing a baseline model with this blur, we observe only small and inconclusive differences in score: training without blurring but testing with it yielded slightly better results than both training and testing without the blur, yet training and testing with the blur produced inferior outcomes. This indicates that (i) the network is not able to read the coordinates, and (ii) the bottom rows do not contain critical geographic information. However, we recommend using the blur for methods that use higher-resolution models to obscure any potential location-specific details in the text.

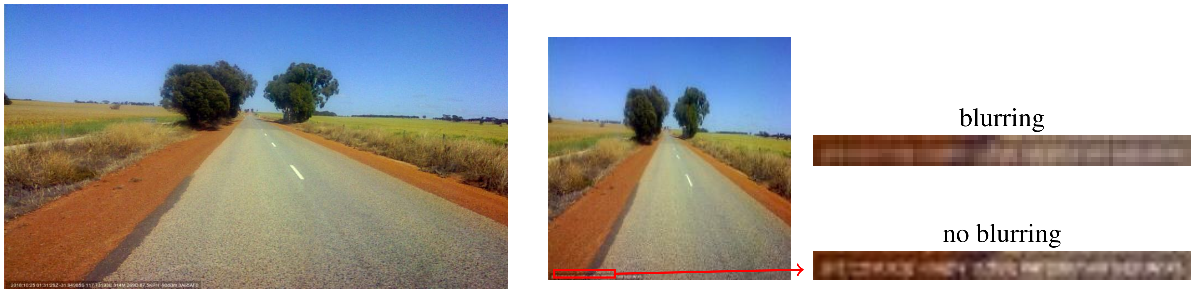
Limitations. We list three main limitations of our OSV-5M dataset:

(i) *Geographical Bias in Training.* Due to our reliance crowd-sourced from Mapillary users, the distribution of locations is biased towards Western countries. We designed our test set to explicitly balance this distribution, but the training set remains affected by the number of selected images.

(ii) *User separation.* We successfully separated images from the same sequence between training and test sets. However, we could not separate images uploaded by the same user on different days, as the required metadata was not available at the time of the dataset construction.

(iii) *Resolution.* The dataset provides images with a vertical resolution of 512 pixels. This restricts the ability to zoom in and read distant texts, for example in street signs, potentially obscuring valuable visual cues. However, through our metadata users can access higher-resolution versions of all our images on the mapillary website.

Training SOTA methods on OSV-5M. Many state-of-the-art geolocation methods [7, 11, 12, 23] either rely on private datasets or lack publicly available code, that prevents their evaluation. In our main paper we evaluated the performance of the pretrained StreetCLIP model both for zero-shot retrieval (Tab 6 and Fig 6) and as a pretrained image encoder (Tab 2), yet the implementation required to fine-tune the model is not publicly available. Similarly, the complete



(a) Full resolution image

(b) Image rescaled in dataloader.

Figure C. **Visible Geotagging.** A small minority of images ($< 5\%$) have visually overlayed geotags at their bottom left corner (a). For those images as resized by our data loader to 224×224 and as optionally blurred, we empirically measure to not provide any important information that the network can use to improve its performance .

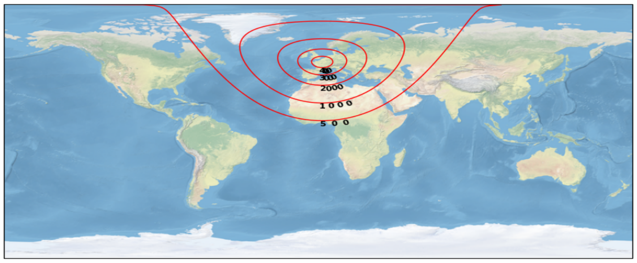


Figure D. **Geoscore.** From a point centered in Paris, red contours highlight level sets of the score along the earth's spherical geometry.

training code of Translocator [21] is also not available. We managed to train the publicly available ISN model [8] on OSV-5M, achieving good performance which we attribute to its bespoke geocell module. The aforementioned difficulty in training and evaluating SOTA models show the clear need for open-source datasets and implementations of visual geolocation approaches, that our paper directly addresses.

Geoscore. In our paper geoscore is introduced as a better evaluation method as it strikes a balance between rewarding precision and not being oversensitive to outlier predictions. Let us consider, for example, a model which produces nine accurate predictions but fails on the tenth image, choosing New Zealand instead of Ireland, a 20 000km mistake. Contrast this with another model which consistently mispredicts by 2 000km. Solely examining the mean error might misleadingly favor the latter model, when the first one has a higher geographic proficiency. In terms of geoscore, the model with one major error would achieve an average score close to 4500, while the one that is consistently off would score 1300. In that way, geoscore provides a more intuitive way to compare the performance of models on our dataset. See Figure D for an illustration of Geoscore.

B. Additional Experiments

This section presents further results and analysis of our proposed framework.

Auxiliary Supervision. We start by evaluating the performance gained by learning to predict various auxiliary information. Based on their coordinates, we associate to each image of our dataset the following meta-data, according to its latitude and longtitude coordinates:

- *Land Cover.* Relying on the Global Land Cover Share Database [17], we classify each image of our dataset into one of 11 land cover types, such as artificial, forest, or crops.
- *Climate.* We use recent Köppen-Geiger climate classification maps [6] to associate each image with a climate type among 31, such as tropical rainforest, arid steppe, or temperate with dry winter.
- *Soil Type.* Thanks to the Digital World Soil Map [22], we characterize the local soil with a 15 class nomenclature, such as acrisols, fluvisols, or ferralsols.
- *Driving Side.* We also add a binary indicator for whether a country uses left or right-hand traffic.
- *Distance to the Sea.* For all locations we compute the distance to their nearest sea.

The maps we used to extract land cover, climate, and soil types come in a resolution of 1 km (or 30 arc-seconds).

We use an MLP f^{aux} to predict the image's metadata in addition to its coordinates. All categorical variables are supervised with the unweighted sum of cross-entropy terms, while the distance to the sea is supervised with the L1 loss. Adding auxiliary tasks encourages the model to focus on relevant geographical cues. As seen in Table A, we only observe a modest impact, indicating that the large train set of OSV-5M allows our model to already learn good latent variables for geolocation. It should be noted that our model can perform accurate predictions for complex geographic variables in the test set, which may have some useful appli-

Table A. **Auxiliary Variables.** We report the impact on geolocation performance of learning to predict various auxiliary variables. We also report the performance on the test set for each variable as the overall accuracy or the average error.

	Num of classes.	Perf. test	Geo ↑ score	Dis ↓ tance	Classification accuracy ↑			
					country	region	area	city
no auxiliary	-	-	2893	2085	**54.9**	19.1	1.6	**0.8**
land cover	11	54.8	2821	2102	52.2	16.9	1.4	0.7
climate	31	58.3	2898	2022	53.7	18.8	**1.7**	**0.8**
soil type	15	47.7	2826	2111	52.4	17.6	1.5	0.7
driving side	1	94.6	2896	2025	54.5	18.7	1.6	0.7
dist to sea	-	543km	2870	2053	52.5	18.7	1.5	0.7
all	-	-	**2910**	**1987**	54.0	**19.8**	1.6	**0.8**

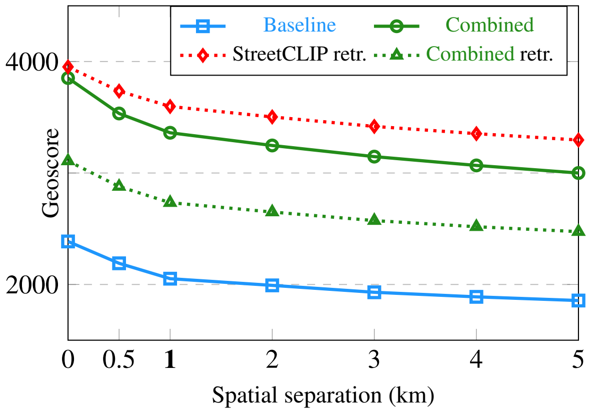


Figure E. **Spatial Separation.** We report the performance of different approaches for test sets defined by various separation radii for the train set.

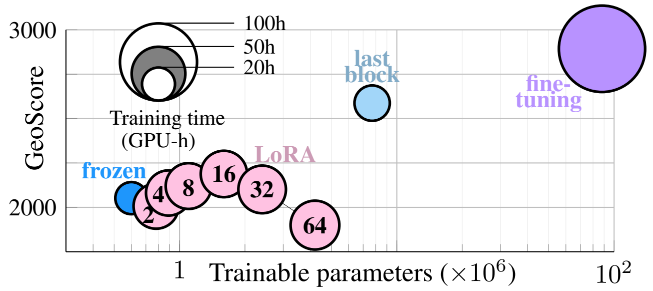
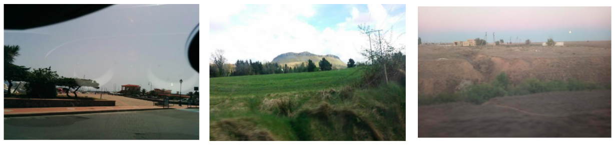


Figure F. **Effect of LoRA Bottleneck Width.** We report the performance of finetuning with LoRA of different bottleneck widths, in comparison to finetuning the last block, or the whole network. For each experiment, the marks' radius are proprtional to the training time.

cations in itself.

Spatial Separation. We study the impact of the radius of spatial separation between the train and the test set. We do this by creating test sets along different radii of separation from the training set: 0m (488k images), 500m (294k), 1km (210k), 2km (166k), 3km (136k), 4km (117k) and 5km (107k). As observed in Figure E, all methods, including retrieval-based approaches, are equally affected by this phenomenon, indicating that, as expected, the problem of global geolocation becomes harder as the separation radius increases. This allows us to define different versions of our test set tiered by difficulty. In particular, if we remove the separation between train and test makes the task becomes significantly easier: 3952 geoscore for StreetCLIP in retrieval mode and 3852 for our best model, corresponding to an average distance error of 1191km.



(a) **GT:** Sardinia
Comb.: Senegal
Base: Mali
SCLIP: Italy

(b) **GT**: Irland
Comb.: Lesotho
Base: Australia
SCLIP: USA

(c) **GT**: Russia
Comb.: Erythrea
Base: Saudi Arabia
SCLIP: Turkmenistan

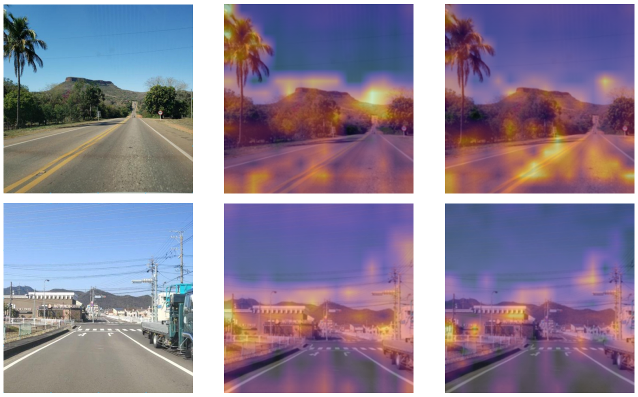
Figure G. **Erroneous Predictions.** Images that are consistently predicted wrongly despite being sampled from areas with relatively high density of training images.

LoRA. Fig F shows the results with different widths of the LoRA bottleneck, ranging from 2 to 64. We share similar observations with the LoRA paper [13, 7.2]: higher ranks do not increase or even slightly decrease performance. Unfreezing the last transformer block remains more efficient in terms of training time, and fine-tuning the entire model leads to even better performance.

Erroneous Predictions. In Fig G we illustrate some sources of geolocation errors not related to the density of training images. These include landscapes that are: (i) similar between very distant countries (Fig G (a,b)), or (ii) any key information is far away from the camera (Fig G (b,c)), or are (iii) monotonous and nearly featureless (Fig G (c)).

Humans and Baselines. We compare in Table B our models against two random baselines: selecting randomly a location on the map or the location of a random image from the training set. We also construct an Annotator Ensemble Oracle by selecting the most accurate prediction for each image from all annotators. Our baseline model, and more substantially our combined model, far surpasses the accuracy of individual annotators, but is still outmatched by the Annotator Ensemble Oracle.

Attention Maps. We represent in Figure H the self-attention maps of the `[CLS]` token of the last layer of the



(a) Input Image (b) Mean attention (c) Selected head

Figure H. **Attention Maps.** We visualize the self-attention maps of the `[CLS]` token of the last layer of the image encoder of the combined model. We show the mean across all heads in (b), and manually selected an interesting layer in (c).

Table B. **Annotator Performance.** We report the average performance of 80 annotators on a subset of 50 images.

	Geo ↑ score	Dis ↓ tance	Classification accuracy ↑		
			continent	country	region
Annot. performance	1009	6407	48.9	12.2	3.0
Annot. ensemble oracle	**3919**	**443**	**98.0**	**70.0**	28.0
Random location	120	10273	16.0	0.0	0.0
Random image	328	8724	20.0	2.0	0.0
Base model	2235	3247	74.0	36.0	8.0
Combined model	3333	1948	86.0	**70.0**	**34.0**

combined model of images from the teaser. We observe that the network focuses on regions of interest containing useful geographical cues, such as the double yellow road line—a specific trait to certain countries—or vegetation and buildings.

C. Implementation Details

In this section, we detail our architecture, loss, metrics, and the retrieval algorithm.

Architecture. All considered networks have a base image encoder $\mathcal{I} \mapsto \mathbb{R}^d$, with a d which depends on each architecture ($d = 768$ for the model ViT-B-32, and $d = 1024$ for all the other encoders). We then add one or several heads to map the image representation to geographical information:

- *Regression* f^{loc}. This network directly predicts the longitude and latitude of an image with a MLP of size $d \mapsto d \mapsto 64 \mapsto 2$ with group norms of 4 groups [24] and without normalizing the last layer.

- *Regression* f^{loc} *sin/cos.* For this variation, we predict the cosine and sine of both coordinates with an MLP: $d \mapsto d \mapsto 64 \mapsto 4$ with a normalization that ensures that the squared sum between coordinate $0, 1$ and $2, 3$ is 1. We then use the **atan2** function to recover the corresponding coordinates.

- *Classification* $f^{classif}$. To predict in which of the K geographic divisions an image was taken, we use an MLP: $d \mapsto d \mapsto 512 \mapsto K$.

- *Hybrid* $f^{relative}$. In the hybrid model, we predict both the division and the position of the image within this cell. The relative position is predicted *for all cells* with an MLP $\phi^{\text{relative}} : d \mapsto d \mapsto 512 \mapsto \mathbb{R}^2 K$ with a specialized normalization for the last layer, explained below. During inference, we select the relative prediction of the cell with the highest prediction score for f^{classif}. During training, we only supervise the relative prediction that corresponds to the true cell.

For this network, a key implementation detail is the normalization of the last layer of ϕ^{relative}. We require that for each cell a prediction of $(0, 0)$ should correspond to the centroid $h^\star, w^\star \in \mathcal{C}^2$ of the training set images in the cell, and that a range of prediction of $[-1, 1]^2$ covers the entire bounding box of size h, w. As illustrated in Figure I, we denote by $x^\star, y^\star \in [0, 1]^2$ the relative position of the centroid in the cell and by x, y the prediction of the MLP ϕ^{aux}. The output of f^{relative} is defined as follows:

$$w^\star + w \begin{cases} -xx^\star & \text{if } x \leq 0 \\ x(1 - x^\star) & \text{else} \end{cases}, \tag{1}$$

$$h^\star + h \begin{cases} -yy^\star & \text{if } y \leq 0 \\ y(1 - y^\star) & \text{else} \end{cases}. \tag{2}$$

This normalization allows the network ϕ^{relative} to easily predict the centroid of the cell, which facilitates learning the distribution of images of that cell. This is particularly crucial for cells with an off-centered centroid, as it provides increased precision in high density areas. In practice, removing this normalization decreases the performance of the hybrid model by 59 points of geoscore, or 22% from the benefit brought by using a hybrid model over pure classification.

- *Auxiliary* f^{aux}. Finally, the auxiliary network is an MLP $d \mapsto d \mapsto 64 \mapsto A$, where A corresponds to the number of auxiliary task predictions: 11 for land cover, 31 for climate, 15 for soil type, 1 for the driving side, and 1 for the distance to the nearest sea. For all classification tasks (*i.e.* everything except the distance to the sea), we softmax the output logits.

Contrastive Learning. We use the MIL-NCE loss [18] as our contrastive objective, which extends the InfoNCE loss [20] to cases where each sample can have multiple positive

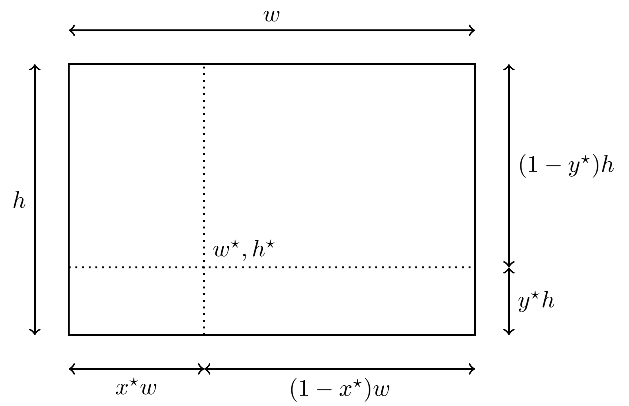


Figure I. **Hybrid Model.** The normalization of the hybrid model requires special considerations to ensure that the output (x, y) of ϕ^{aux} is such $(0, 0)$ maps to the cell's centroid $w^\star, h^\star$, and that $[-1, 1]^2$ maps the entire cell.

matches.

$$\sum_{i\in\mathcal{B}} \log\left(\frac{\sum_{p\in\mathcal{P}_i} e^{f^{\text{img}}(i)^\intercal f^{\text{img}}(p)/T}}{\sum_{p\in\mathcal{P}_i} e^{f^{\text{img}}(i)^\intercal f^{\text{img}}(p)/T}+\sum_{n\in\mathcal{B}\setminus\mathcal{P}_i} e^{f^{\text{img}}(i)^\intercal f^{\text{img}}(n)/T}}\right), \tag{3}$$

with $\mathcal{P}_i \subset \mathcal{B}$ the set of image positively paired with i and T a temperature parameter set as 0.1. If an image has only one positive match, this equation becomes the InfoNCE loss [20].

Nearest Neighbors Retrieval To perform nearest neighbor retrieval, we create a HNSW32 indexe using the FAISS library [15] through the autofaiss package (https://github.com/criteo/autofaiss). This approach achieves fewer than 200 self-consistency errors per million with over 90% compression rate.

During retrieval, our training set is divided into five parts, each requiring 15 minutes for index computation and collectively consuming 15.6GB of storage for StreetCLIP embeddings, our most resource-intensive model. This setup enables us to predict locations for 12,000 to 32,000 test images per second, depending on the model size.

Although retrieval methods demonstrate high performance and have been made efficient with approximate methods, it is important to note that they are not a learning technique, as they rely on already geographically relevant representations that are already learned.

D. Datasheet for Dataset

D.1. Motivation

Q1 **For what purpose was the dataset created?** Was there a specific task in mind? Was there a particular gap that needed to be filled? Please provide a description.

- OpenStreetView-5M (OSV-5M) is the first global scale, open-access, large dataset of street view images. Its goal is to enable the training and evaluation of modern computer vision approaches for global visual geolocation, which would depend until now on proprietary or expensive APIs such as Google Street View. More broadly, OSV-5M can be used to evaluate and improve representation learning.

Q2 **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

- The dataset was created as part of a the "IMAGINE Summer Hackathon", an internal event of the LIGM/ENPC/UGE laboratory. All images of OSV-5M come from the Mapillary website, which is a platform where users upload georeferenced images.

Q3 **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

- This work was partially supported by the ANR project READY3D ANR-19-CE23-0007 and used the HPC resources of IDRIS under the allocation 2022-AD011012096R1 made by GENCI.

Q4 **Any other comments?**

- All the images of OSV-5M are already openly accessible through Mapillary's heavily moderated database. We only selected a small fraction distributed across the globe, and added metadata from public sources.

D.2. Composition

Q5 **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?**

- OSV-5M is composed of street view images depicting various street scenes, captured by dash-cams of different vehicles from across the world.

Q6 **How many instances are there in total (of each type, if appropriate)?**

- The training set contains 4,894,685 images, and the test set 210,122.

Q7 **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**

- OSV-5M is a small subset of 5.1M images from the 1.8 billion images hosted on the Mapillary website.

Q8 **What data does each instance consist of?**

- Each instance consists of a georeferenced street view image with a height of 512 pixels.

Q9 **Is there a label or target associated with each instance?**

- **Yes.** Each image is associated with the following targets: longitude and latitude, administrative division (country, region, sub-region, closest city), and labels corresponding to the local land cover, soil, and climate type at a resolution of 30 arc seconds (1km). We also add the distance to the nearest sea and the driving side of the country.

Q10 **Is any information missing from individual instances?**

- **Yes.** Sub-regions are not defined for all countries, about 30% of the instances do not have a value for this field.

Q11 **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?**

- **No.** The data is organized as a collection of images with no particular order or relations. However, the metadata allows a user to organize them based on different geographical criteria.

Q12 **Are there recommended data splits (e.g., training, development/validation, testing)?**

- **Yes.** We provide an official training and test set. Our implementation also proposes a validation split.

Q13 **Are there any errors, sources of noise, or redundancies in the dataset?**

- **Yes.** We have heavily filtered the dataset using semi-automatic methods to discard low-quality images and wrong localization, as presented in Section A. We have estimated through the manual inspection of 4500 images that 96.1% (±0.57% with a 95% confidence level) of the images in OpenStreetView-5M are perceptually localizable, *i.e.* provide a clear enough overview of their surroundings.

Q14 **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?**

- **No.** OSV-5M is self-contained and will be stored and distributed on huggingface.co.

Q15 **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?**

- **No.** OSV-5M relies on crowdsourced data, whose license is respected by providing usernames for each image, which is include in our metadata.

Q16 **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** *If so, please describe why.*

- **Highly unlikely**: OSV-5M contains 5 million images of streets that come from Mapillary, which imposes a strong crowd-sourced moderation policy.

Q17 **Does the dataset relate to people?**

- **Yes.** Many of the images of OSV-5M contain vehicles and some contain pedestrians, yet Mappilary performs highly accurate privacy blurring.[1]

Q18 **Does the dataset identify any subpopulations (e.g., by age, gender)?**

- **No.** The metadata contains no information about the people present in the photography beyond, who are also privacy blurred.[1]

Q19 **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?**

- **No.** The license plates and faces of pedestrians have been privacy blurred by Mapillary using an automatic algorithm with over 99% recall for faces and 99.9% recall for license plates.[1] Furthermore, users can signal images that violate privacy.

  We also manually inspected 4500 images and observed no confidentiality leak. With a confidence of 95% we can assume that fewer than 0.067% of the dataset contains leaks.

Q20 **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**

- **No.**

Q21 **Any other comments?**

- **No.**

D.3. Collection Process

Q22 **How was the data associated with each instance acquired?**

The images of Mapillary are taken and uploaded by users of the Mapillary platform. We downloaded the images directly from Mapillary's API. Additional metadata was collected from the following open-access sources: (i) land cover: Global Land Cover Share Database [17] (ii) climate: Köppen-Geiger climate classification maps [6], (iii) soil type: Digital World Soil Map [22] (iv) administrative division: reverse geocoder [4].

[1] See https://blog.mapillary.com/update/2018/04/19/accurate-privacy-blurring-at-scale.html

Q23 **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?**

- We used Mapillary's web API and a Python script running on a standard workstation.

Q24 **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

- We first defined a 100×100m grid across the entire world and sampled one image per cell among the 1.8B images of Mapillary. We then sample the train and test sets with a weight proportional to the local image density raised to the power of -0.75. We then filter the images based on both learned and handcrafted filters, as described in Section A.

Q25 **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

- The images are crowdsourced by Mapillary users who agree on Mapillary's terms of use. To the best of our knowledge, users are not compensated.

Q26 **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?**

- The images used in OSV-5M were uploaded between January 2011 and August 2023.

Q27 **Were any ethical review processes conducted (e.g., by an institutional review board)?**

- **No.**

Q28 **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

- **N.A.** The images were downloaded through Mapillary's API.

Q29 **Were the individuals in question notified about the data collection?**

- **No.** We followed the terms of use of Mapillary.

Q30 **Did the individuals in question consent to the collection and use of their data?**

- **Yes.** Following the Mapillary terms of use, a user agrees for their data to be be used respecting the CC BY-SA 2.0 DEED license.

Q31 **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**

- **N.A.**

Q32 **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?**

- **No.** However, users of OSV-5M can signal potential issues with the images to the corresponding authors. Flagged images will be removed and Mapillary will be further contacted.

Q33 **Any other comments?**

- All the images of OSV-5M are already openly accessible through Mapillary's heavily moderated database. We only added additional metadata from public sources.

D.4. Preprocessing, Cleaning, and/or Labeling

Q34 **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**

- **Yes.** We removed the images based on learned and handcrafted filters, as described in Section A. In particular, we removed images that were classified as blurry, too dark or purple, or badly exposed. We also used a pretrained model [10] to detect and remove images with potential spurious orientation.

Q35 **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** *If so, please provide a link or other access point to the "raw" data.*

- **Yes.** The removed images are saved on a local server but are not public. Note that all these images, including the filtered ones, are still available on Mapillary's website.

Q36 **Is the software used to preprocess/clean/label the instances available?**

- **Yes.** The script used for cleaning the dataset will be released alongside the dataset.

Q37 **Any other comments?**

- **No.**

D.5. Uses

Q38 **Has the dataset been used for any tasks already?**

- **Yes.** To train and evaluate geolocation models, the subject of the paper.

Q39 **Is there a repository that links to any or all papers or systems that use the dataset?**

- **No.** But once we release the dataset we will maintain an updated list on the project page.

Q40 **What (other) tasks could the dataset be used for?**

- The images of OSV-5M can be used for both self-supervised learning and generative modeling, both as a pretraining or fine-tuning dataset. The metadata beyond geolocation can be used as targets for separate tasks.

Q41 **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**

- The density-based sampling leads to a spatial distribution that may not fit other datasets and tasks.

Q42 **Are there tasks for which the dataset should not be used?**

- **Yes.**The same limitations that apply for Mapillary data (CC BY-SA 2.0 DEED), also apply to our dataset.
- **Privacy Concerns.** Despite being heavily moderated, the dataset may contain images of individuals or private residences. Usage must avoid applications that can infringe on personal privacy or exercise surveillance and open-source intelligence (OSINT).
- **Cultural and Ethical Sensitivity.** The dataset spans a wide range of cultures and countries, each with its own set of ethical norms and cultural sensitivities. We strongly advise against using OSV-5M in a way that might propagate stereotypes, misrepresent cultures, or otherwise harm the dignity and representation of the featured communities.
- **Manipulation and Misrepresentation.** The dataset should not be used to create misleading representations of locations or to manipulate images in a way that distorts or misrepresents the reality of the places and the depicted people.

Q43 **Any other comments?**

- **No.**

D.6. Distribution

Q44 **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?**

- **Yes.** The dataset will be open-access and accessible to the research community.

Q45 **How will the dataset be distributed (e.g., tarball on website, API, GitHub)?**

- The data will be hosted on huggingface.co.

Q46 **When will the dataset be distributed?**

- The dataset will be distributed upon the publication of the preprint on arXiv, which should be in Q2 of 2024.

Q47 **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** *If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

- **Yes.** The dataset inherits from Mappilary CC-BY-SA license: free of use with attribution to the authors of the images [1].

Q48 **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**

- **No.**

Q49 **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**

- **No.**

Q50 **Any other comments?**

- **No.**

D.7. Maintenance

Q51 **Who will be supporting/hosting/maintaining the dataset?**

- The authors will maintain the dataset. The dataset will be hosted on huggingface.co.

Q52 **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

- A dedicated email will be created.

Q53 **Is there an erratum?**

- **No.** There is no erratum for our initial release. Errata will be documented as future releases on the dataset website.

Q54 **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**

- **Yes.**

Q55 **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?**

- **N.A.**

Q56 **Will older versions of the dataset continue to be supported/hosted/maintained?**

- **Yes.** We are dedicated to providing ongoing support for the OSV-5M dataset.

Q57 **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**

- **Yes.** The data is free of use under Mappilary CC-BY-SA license. User making explicit use of our proposed split should cite our paper.

Q58 **Any other comments?**

- **No.**

References

[1] 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. 10

[2] Mapillary. https://www.mapillary.com/. Accessed: 2023-10-10. 1

[3] Multimediacommons - yfc100m core dataset. https://multimediacommons.wordpress.com/yfcc100m-core-dataset/. Accessed: 2023-10-10. 2

[4] Reverse Geocoder. pypi.org/project/reverse_geocoder/. Accessed: 2023-10-10. 8

[5] Yfcc100m. https://gitlab.com/jfolz/yfcc100m. Accessed: 2023-10-10. 2

[6] Hylke E Beck, Niklaus E Zimmermann, Tim R McVicar, Noemi Vergopolan, Alexis Berg, and Eric F Wood. Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Scientific data*, 2018. 4, 8

[7] 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. 3

[8] Müller-Budack etal. Geolocation estimation of photos using a hierarchical model and scene classification. In *ECCV*. 4

[9] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 2021. 1

[10] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *ICLR*, 2018. 1, 9

[11] Lukas Haas, Silas Alberti, and Michal Skreta. Learning generalized zero-shot learners for open-domain image geolocalization, 2023. 3

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

[13] 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. 5

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

[15] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 2019. 7

[16] 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

[17] John Latham, Renato Cumani, Ilaria Rosati, and Mario Bloise. Global land cover share (GLC-SHARE) database beta-release version 1.0-2014. *FAO: Rome, Italy*, 2014. 4, 8

[18] 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. 6

[19] 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

[20] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 6, 7

[21] 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. 4

[22] Pedro A Sanchez, Sonya Ahamed, Florence Carré, Alfred E Hartemink, Jonathan Hempel, Jeroen Huising, Philippe Lagacherie, Alex B McBratney, Neil J McKenzie, Maria De Lourdes Mendonça-Santos, et al. Digital soil map of the world. *Science*, 2009. 4, 8

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

[24] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018. 6