PIGEON: Predicting Image Geolocations

Lukas Haas Michal Skreta Silas Alberti Chelsea Finn

Stanford University

Abstract

Planet-scale image geolocalization remains a challenging problem due to the diversity of images originating from anywhere in the world. Although approaches based on vision transformers have made significant progress in geolocalization accuracy, success in prior literature is constrained to narrow distributions of images of landmarks, and performance has not generalized to unseen places. We present a new geolocalization system that combines semantic geocell creation, multi-task contrastive pretraining, and a novel loss function. Additionally, our work is the first to perform retrieval over location clusters for guess refinements. We train two models for evaluations on street-level data and general-purpose image geolocalization; the first model, PIGEON, is trained on data from the game of GeoGuessr and is capable of placing over 40% of its guesses within 25 kilometers of the target location globally. We also develop a bot and deploy PIGEON in a blind experiment against humans, ranking in the top 0.01% of players. We further challenge one of the world's foremost professional GeoGuessr players to a series of six matches with millions of viewers, winning all six games. Our second model, PIGEOTTO, differs in that it is trained on a dataset of images from Flickr and Wikipedia, achieving state-of-the-art results on a wide range of image geolocalization benchmarks, outperforming the previous SOTA by up to 7.7 percentage points on the city accuracy level and up to 38.8 percentage points on the country level. Our findings suggest that PIGEOTTO is the first image geolocalization model that effectively generalizes to unseen places and that our approach can pave the way for highly accurate, planet-scale image geolocalization systems. Our code is available on GitHub.[1]

1. Introduction

The online game GeoGuessr has recently reached 65 million players [22], attracting a worldwide crowd of users trying to solve a single problem: given a Street View image taken somewhere in the world, identify its location. The problem of uncovering geographical coordinates from visual data is more formally known in computer vision as image geolocalization, and, just like the game of GeoGuessr, remains notoriously challenging. The scale and diversity of our planet, seasonal appearance disturbance, and climate change impacts are some among the many reasons why image geolocalization remains an unsolved problem.

Over the past decade, researchers have advanced the field by casting image geolocalization as a classification task [38], developing hierarchical approaches to problem modeling [7, 25, 27], as well as leveraging vision transformers [7, 27] and contrastive pretraining [23]. Yet despite this progress, the most capable models have been highly dependent on distributional alignments between training and testing data, failing to generalize to more diverse datasets that predominantly include unseen locations [7].

In this work, we present a two-pronged multi-task modeling approach that both exhibits world-leading performance in the game of GeoGuessr and achieves state-of-the-art performance on a wide range of image geolocalization benchmark datasets. First, we present **PIGEON**, a model trained exclusively on planet-scale Street View data, taking a four-image panorama as input. PIGEON is the first computer vision model to reliably beat the most experienced players in the game GeoGuessr, comfortably ranking within the top 0.01% of players while also beating one of the world's best professional players in six out of six games with millions of viewers. Our model achieves impressive image geolocalization results on outdoor street-level images globally, placing 40.4% of its geographic coordinate predictions within a 25-kilometer radius of the correct location.

Subsequently, we evolve our model to **PIGEOTTO** which differs from PIGEON in that it takes a single image as input and is trained on a larger, highly diverse dataset of over 4 million photos derived from Flickr and Wikipedia and no Street View data. PIGEOTTO achieves state-of-the-art results across a wide range of benchmark datasets, including IM2GPS [14], IM2GPS3k [37], YFCC4k [37], YFCC26k [25], and GWS15k [7]. The model slashes the

[1] https://github.com/LukasHaas/PIGEON.

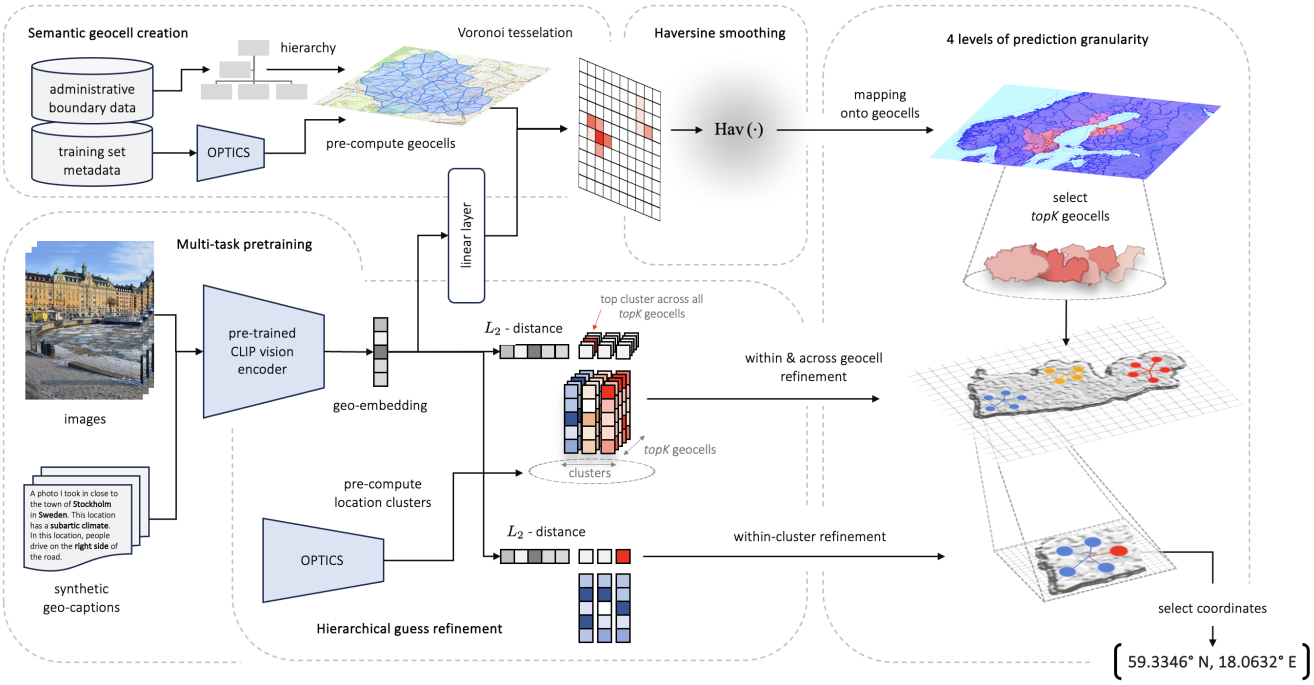


Figure 1. **Prediction pipeline and main contributions of PIGEON**. Administrative boundary and training set metadata are hierarchically ranked, clustered, and Voronoi tessellated to create semantic geocells. The geocell labels are then used to create continuous labels via haversine smoothing. Additionally, we pretrain CLIP via geographic synthetic captions in a multi-task setting. The pretrained CLIP model together with an OPTICS clustering model are employed to generate location cluster representations. During inference, an image embedding is computed and first passed to a linear layer to create geocell predictions and to identify the *topK* geocell candidates. The embedding is also used in our refinement pipeline to refine predictions within and across geocells. This is achieved by minimizing the embedding L_2-distance between the inference image embedding and all location cluster representations across the *topK* geocells. Finally, predictions are refined within the top identified cluster to generate geographic coordinates as outputs.

median distance error roughly in half on three benchmark datasets and more than five times reduces the median error on GWS15k which includes images from predominantly unseen locations. PIGEOTTO is the first model that is robust to location and image distribution shifts by picking up general locational cues in images as evidenced by the often double-digit percentage-point increase in performance on larger evaluation radii. By performing well on out-of-distribution datasets, PIGEOTTO closes a major gap in prior literature that is essential for solving the problem of image geolocalization.

As PIGEON and PIGEOTTO only differ in the training data and hyperparameter settings, the efficacy of our approach has important implications for planet-scale image geolocalization. Our contributions of semantic geocells, multi-task contrastive pretraining, a new loss function, and downstream guess refinement all contribute to minimizing distance errors, as shown in our ablation studies in Section 4. Still, it is important that future research addresses the safety of image geolocalization technologies, ensuring responsible progress in developing computer vision systems.

2. Related work

2.1. Image geolocalization problem setting

Image geolocalization refers to the problem of mapping an image to coordinates that identify where it was taken. This problem, especially if planet-scale, remains a very challenging area of computer vision. Not only does a global problem formulation render the problem intractable, but accurate image geolocalization is also difficult due to changes in daytime, weather, seasons, time, illumination, climate, traffic, viewing angle, and many more factors.

The first modern attempt at planet-scale image geolocalization is attributed to IM2GPS (2008) [14], a retrieval-based approach based on hand-crafted features. Dependence on nearest-neighbor retrieval methods [43] using hand-crafted visual features [8] meant that an enormous database of reference images would be necessary for accurate planet-scale geolocalization, which is infeasible. Consequently, subsequent work decided to restrict the geographic scope, focusing instead on specific cities [40] like Orlando and Pittsburgh [42] or San Francisco [5]; specific countries like the United States [32]; and even mountain ranges [3, 29, 34], deserts [35], and beaches [6].

2.2. Vision transformers and multi-task learning

With the advent of deep learning, methods in image geolocalization shifted from hand-crafted features to end-to-end learning [24]. In 2016, Google released the PlaNet [38] paper that first applied convolutional neural networks (CNNs) [19] to geolocalization. It also first cast the problem as a classification task across "geocells" as a response to research demonstrating that it was difficult for deep learning models to directly predict geographic coordinates via regression [9, 33]. This was due to the subtleties in geographic data distributions and the complex interdependence between latitudes and longitudes. The improvements realized with deep learning led researchers to revisit IM2GPS [37], apply CNNs to massive datasets of mobile images [16], and deploy their models in the game of GeoGuessr against human players [23, 32]. Prior literature has also combined classification and retrieval approaches [18]; our work modernizes this approach via a hierarchical retrieval mechanism over location clusters, equivalent to prototypical networks [31] with fixed parameters.

Following the success of transformers [36] in natural language processing, the transformer architecture found its application in computer vision. Pretrained vision transformers (ViT) [17] and multi-modal derivatives such as OpenAI's CLIP [28] and GPT-4V [26] have successfully been deployed to image geolocalization [1, 23, 26, 27, 40, 44]. Our approach is novel in that in pretrains CLIP specifically for the task of image geolocalization in a multi-task fashion via auxiliary geographic, demographic, and climate data. Auxiliary data had previously been shown to aid in image geolocalization [14, 27], but our work is the first to use auxiliary data for contrastive pretraining, retaining CLIP's exceptional in-domain generalized zero-shot capabilities that are critical for geolocalization performance [13].

2.3. Geocell partitioning

With image geolocalization framed as a classification problem, the chosen method of partitioning the world into geographical classes, or "geocells", can have an enormous effect on downstream performance. Previous approaches rely on geocells that are either plainly rectangular, rectangular while respecting the curvature of the Earth and being roughly balanced in class size [25] (as is the case of Google's S2 library[2]), or geocells that are effectively arbitrary as a result of combinatorial partitioning, initializing cells randomly but adjusting their *shapes* based on the training dataset distribution [30]. Hierarchical approaches to geocell creation like in individual scene networks (ISNs) [25, 33] can help preserve semantic information and exploit the hierarchical knowledge at different geospatial resolutions, for instance by categorizing the geocells at the city, region, and country levels.

While the semantic construction of geocells has been found to be of high importance to image geolocalization [33], even recently published papers continue to use the S2 library [7, 18, 27]. One of the possible reasons for this design choice is that for larger datasets, even the most granular semantic geocells contain too many data points, causing the classification problem to be very imbalanced. Our work addresses this limitation with a novel semantic geocell creation method, combining hierarchical approaches with clustering based on the training data distribution and Voronoi tesselation as the missing link between the two. For the first time, our approach renders semantic geocells useful for any dataset size and geographic distribution.

2.4. Additional work

Other notable academic work cites the efficacy of cross-view image geolocalization, especially for rural regions with sparse, ground-level geo-tagged photos. Cross-view approaches can combine land cover attributes and ground-level and overhead imagery to increase robustness through transfer learning [21, 41, 44]. Using land maps in particular is an important avenue for future research; in our work, however, we aim to demonstrate our models' performance relying solely on ground-level images from diverse settings.

3. Predicting image geolocations

Our image geolocalization system consists of both parametric and non-parametric components. This section first explains our data pre-processing pipeline and then walks through how we frame geolocalization as a distance-aware classification problem. We then delineate our pretraining and training stages, and finally describe how we refine location predictions to improve street-level guess performance.

3.1. Geocell creation

Contemporary methods all frame image geolocalization as a classification exercise, relying on geocells to discretize the Earth's surface into a set number of classes. Our work experiments with two types of geocell creation methods.

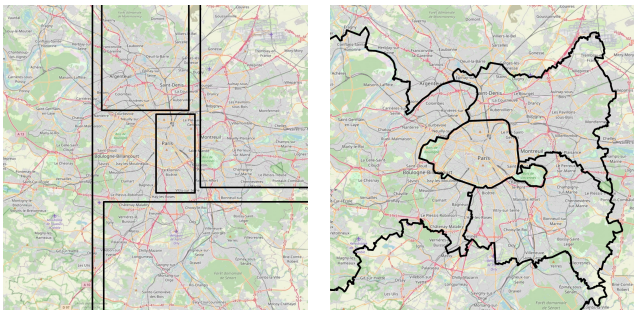
Naive geocells. We first employ naive, rectangular geocells inspired by the S2 library, subdividing every geocell until roughly balanced class sizes are reached. In contrast to S2 partitioning, our rectangular geocells are not of equal geographic size, creating even more balanced classes.

Semantic geocells. One limitation of the S2 library and our naive geocells is that the geocell boundaries are completely arbitrary and thus meaningless in the context of

[2] https://code.google.com/archive/p/s2-geometry-library.

image geolocalization. Ideally, each geocell should capture the distinctive characteristics of its enclosed geographic area. Political and administrative boundaries serve this purpose well as they often not only capture country or region-specific information (i.e. road markings and street signs) but also follow natural boundaries, such as the flow of rivers and mountain ranges which encode geological information.

Similar to Theiner et al. [33], we rely on planet-scale open-source administrative data for our semantic geocell design, drawing on non-overlapping political shapefiles of three levels of administrative boundaries (country, admin 1, and admin 2 levels) obtained from GADM [10]. Starting at the most granular level (admin 2), our algorithm merges adjacent admin 2 level polygons such that each geocell contains at least a minimum number of training samples. Our method attempts to preserve the hierarchy given by admin 1 level boundaries, never merges cells across country borders (defined by distinct ISO country codes) and, in contrast to Theiner et al. [33], allows for more granular hierarchies. Figure 2 shows an example of our semantic geocell design preserving the semantics of urban and surrounding Paris.



(a) With naive, rectangular geocells. (b) With our semantic geocells.

Figure 2. Geocell specifications around Paris, France.

OPTICS clustering & Voronoi tessellation. We further address a major limitation in the semantic geocell design of Theiner et al. [33] which is that some admin 2 areas are not fine-grained enough to result in a balanced classification dataset. This is especially the case for large training datasets where the number of training examples for a single, urban admin 2 area might greatly exceed the minimum class size, requiring admin 2 areas to be meaningfully split further. An important observation is that the geographic distribution of our training data already gives us an indication of how to meaningfully subdivide our geocells because it clusters around popular places and landmarks. We extract these clusters using the OPTICS clustering algorithm [2]. Finally, we assign all yet unassigned data points to their nearest clusters and employ Voronoi tessellation to define contiguous geocells for every extracted cluster.

3.2. Hierarchical image geolocalization using distance-based label smoothing

By discretizing the problem of image geolocalization, a trade-off is created between the granularity of geocells and predictive accuracy. More granular geocells enable fine-grained predictions but also result in the classification problem becoming more difficult due to a higher cardinality. Prior literature addresses this problem by generating separate geolocalization predictions across multiple levels of geographic granularity, refining guesses at every subsequent level [7, 25, 27]. Pramanick et al. [27] and Clark et al. [7] further propose architectures that share some model parameters between different hierarchy levels, improving geolocalization performance. Surprisingly, all prior work suffers from the same limitation: models figuratively guess in the blind as they do not know which geocells are located next to each other, learning their representations in isolation.

Our approach addresses this major limitation and improves upon prior work by sharing *all* parameters between multiple, implicit levels of geographic hierarchies. We achieve this through a new loss function that relates adjacent geocells to each other, biasing the label based on the haversine distance which calculates the distance between two points on the Earth's surface in kilometers. Given two points, $\mathbf{p}_1 = (\lambda_1, \phi_1)$ and $\mathbf{p}_2 = (\lambda_2, \phi_2)$ with longitude λ and latitude ϕ, and the earth's radius r in kilometers, we define the haversine distance $\text{Hav}(\mathbf{p}_1, \mathbf{p}_2)$ as follows:

$$\text{Hav}(\mathbf{p}_1, \mathbf{p}_2) = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \tag{1}$$

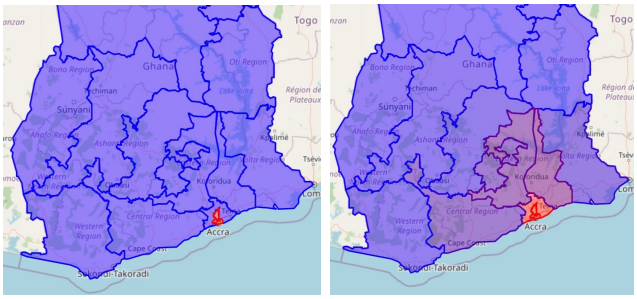
We then "haversine smooth" the original one-hot geocell classification label using this distance metric according to the following equation for a given sample n and geocell i:

$$y_{n,i} = \exp(-\left[\text{Hav}(\mathbf{g}_i, \mathbf{x}_n) - \text{Hav}(\mathbf{g}_n, \mathbf{x}_n)\right]/\tau) \tag{2}$$

where $\mathbf{g}_i$ are the centroid coordinates of the geocell polygon of cell i, $\mathbf{g}_n$ are the centroid coordinates of the true geocell, $\mathbf{x}_n$ are the true coordinates of the example for which the label is computed, and τ is a temperature parameter which is set to 75 for PIGEON and to 65 for PIGEOTTO in our experiments. It is important to note that our "haversine smoothing" is distinct from classical "label smoothing" because labels are not decayed using a constant factor but based on both the distance to the correct geocell and the true location. Since for every training example, multiple geocells will have a target $y_{n,i}$ that is significantly larger than zero, our model simultaneously learns to predict the correct geocell as well as an even coarser level of geographic granularity. We design the following loss function based on haversine smoothing for a particular training sample n:

$$\mathcal{L}_n = -\sum_{g_i \in G} \log\left(p_{n,i}\right) \cdot \exp\left(-\frac{\text{Hav}(\mathbf{g}_i, \mathbf{x}_n) - \text{Hav}(\mathbf{g}_n, \mathbf{x}_n)}{\tau}\right) \tag{3}$$

where $p_{n,i}$ is the probability our model assigns to geocell i for sample n. An added benefit of using the loss of Equation (3) is that it aids generalization because hierarchy definitions vary across every training sample. Additionally, if a sample lies close to the boundary of two geocells, this fact will be reflected through approximately equal target labels for these two geocells. This is especially helpful for larger, often rural, geocells. Furthermore, because every target label $y_{n,i}$ is now continuous and the difficulty of the classification problem can be freely adjusted using τ, an arbitrary number of geocells can be employed as long as geocells are still contextually meaningful and contain a minimum number of samples. Finally, we observe that our classification loss is now directly based on the distance to the true location $\mathbf{x}_n$ of a given sample while circumventing the regression difficulties encountered in prior literature [9, 33].



(a) Without haversine smoothing. (b) With haversine smoothing.

Figure 3. Impact of applying haversine smoothing over neighboring geocells for a location in Accra, Ghana.

3.3. Contrastive pretraining for geolocalization

To generate visual representations to then project onto our geocells, our architecture uses OpenAI's CLIP ViT-L/14 336 model as a backbone which is a multi-modal model that was pretrained on a dataset of 400 million images and captions [28]. The reason why we employ CLIP is that it has been shown to perform exceptionally well in generalized zero-shot learning setups [28], which is a desirable property for image geolocalization of both *seen* and *unseen* places.

In our experiments, we add a linear layer on top of CLIP's vision encoder to predict geocells. For model versions with multiple image inputs (i.e. four-image panorama for PIGEON), we average the embeddings of all images. Averaging embeddings resulted in a superior performance compared to combining multiple embeddings via multi-head attention or additional transformer layers.

In Haas et al. [13], the authors demonstrate that continuing the pretraining of CLIP using domain-specific, synthetic captions derived from caption templates improves the generalized zero-shot performance on image geolocalization tasks. We further improve upon their method through the continued pretraining of CLIP in a *multi-task* fashion.

To this end, we augment our training datasets with geographic, climate, and directional auxiliary data, used to create synthetic image captions by sampling caption components from different category templates and concatenating them. For PIGEOTTO, we use caption components based on the location, climate, and traffic direction. Meanwhile, for PIGEON, the Street View metadata allows us to additionally infer compass directions and the season, the latter included to avoid shortcut learning [12] (i.e. snow $\rightarrow$ polar latitudes). Examples of caption components include:

- **Location**: "A photo I took in the region of Gauteng in South Africa."
- **Climate**: "This location has a temperate oceanic climate."
- **Compass direction**: "This photo is facing north."
- **Season (month)**: "This photo was taken in December."
- **Traffic**: "In this location, people drive on the left side of the road."

All the above caption components contain information relevant for the geolocalization of an image. Consequently, our continued contrastive pretraining creates an implicit multi-task setting and ensures the model learns rich representations of the data while learning features that are relevant to the task of image geolocalization.

3.4. Multi-task learning with climate data

We also experiment with making our multi-task setup explicit by creating task-specific prediction heads for auxiliary labels, and adapt our loss function according to Equation (4), where $\mathcal{L}_{n,\text{loc}}$ corresponds to the loss in Equation (3). Our multi-task setup further includes a cross-entropy classification task ($\mathcal{L}_{n,\text{climate}}$) of the 28 different Köppen-Geiger climate zones [4], a cross-entropy month (season) classification task ($\mathcal{L}_{n,\text{month}}$), and six mean squared error (MSE) regression tasks (combined into $\mathcal{L}_{n,\text{reg}}$) that attempt to predict values related to the temperature, precipitation, elevation, and population density of a given location.

$$\mathcal{L}_n = \mathcal{L}_{n,\text{loc}} + \alpha\mathcal{L}_{n,\text{climate}} + \beta\mathcal{L}_{n,\text{month}} + \gamma\mathcal{L}_{n,\text{reg}} \quad (4)$$

We unfreeze the last CLIP layer to allow for parameter sharing across tasks with the goal of observing a positive transfer from our auxiliary tasks to our geolocalization problem and to learn more general image representations reducing the risk of overfitting to the training dataset. Adjusting α, β, and γ, our loss function weighs the geolocalization task as much as all auxiliary tasks combined considering each task's loss magnitude. A novel contribution of our work is that we use a total of eight auxiliary prediction tasks instead of just two compared to prior research [27].

3.5. Refinement via location cluster retrieval

To further refine our model's guesses within a geocell and to improve street- and city-level performance, instead of simply predicting the mean latitude and longitude of all points within a geocell [27], we perform intra-geocell refinement.

To this end, we design a hierarchical retrieval mechanism over location clusters akin to prototypical networks [31] with fixed parameters. We again use the OPTICS clustering algorithm [2] to cluster all points within a geocell g and thus propose location clusters C_g whose representation is the average of all corresponding image embeddings. To compute all image embeddings, we use our pretrained CLIP model $f(\cdot)$ described in Section 3.3, mapping each image l in a cluster c to its embedding $f(l)$.

$$c^* = \arg\min_{c \in C_g} \left\| f(x) - \frac{1}{|c|} \sum_{l \in c} f(l) \right\|_2 \tag{5}$$

During inference, we predict the location cluster c^* of an input image x by selecting the cluster with the minimum Euclidean image embedding distance to the input image embedding $f(x)$. Once the cluster c^* is determined, we further refine our guess by choosing the single best location within the cluster, again via minimizing the Euclidean embedding distance. The retrieval over location clusters and within-cluster refinement add two additional levels of prediction hierarchy to our system, with the number of unique potential guesses equaling the training dataset size.

While hierarchical refinement via retrieval is in itself a novel idea, our work goes one step further. Instead of refining a geolocalization prediction within a single cell, our mechanism optimizes across multiple cells which further increases performance. During inference, our geocell classification model outputs the *topK* predicted geocells (5 for PIGEON, 40 for PIGEOTTO) as well as the model's associated probabilities for these cells. The refinement model then picks the most likely location within each of the *topK* proposed geocells, after which a softmax is computed across the *topK* Euclidean image embedding distances. We use a temperature softmax with a temperature that is carefully calibrated on the validation datasets to balance probabilities across different geocells. Finally, these refinement probabilities are multiplied with the initial *topK* geocell probabilities to determine a final location cluster and within-cluster refinement is performed as illustrated in Figure 1.

4. Experimental results and analysis

4.1. Experimental setting

Training PIGEON and PIGEOTTO. Based on our technical methodology outlined in Section 3, we train two models for distinct downstream evaluation purposes.

First, inspired by GeoGuessr, we train PIGEON (Predicting Image Geolocations). We collect an original dataset of 100,000 randomly sampled locations from GeoGuessr and download a set of four images spanning an entire "panorama" in a given location, or a 360-degree view, for a total of 400,000 training images. For each location, we start with a random compass direction and take four images separated by 90 degrees, carefully creating non-overlapping image patches.

Second, motivated by PIGEON's image geolocalization capabilities, we train PIGEOTTO (Predicting Image Geolocations with Omni-Terrain Training Optimizations). Unlike PIGEON, PIGEOTTO is not a Street View photo localizer but rather a general image geolocator. To that end, we access the MediaEval 2016 dataset [20] consisting of geo-tagged Flickr images from all over the world and obtain 4,166,186 images, considering that some images have become unavailable since 2016. Additionally, recognizing the importance of geolocating landmarks for general image geolocalization capabilities, we add 340,579 images from the Google Landmarks v2 dataset [39] to our training mix which are all derived from Wikipedia. Importantly, there is no overlap in the training data we use between PIGEON and PIGEOTTO, as the models serve different downstream purposes. Unlike PIGEON, PIGEOTTO takes a single image per location as input, as obtaining a four-image panorama is often infeasible in general image geolocalization settings.

Evaluation datasets and metrics. Our work defines the median distance error to the correct location as the primary and composite metric. In line with the prior literature on image geolocalization, we further evaluate the "% @ km" statistic in our analysis as a more fine-grained metric. For a given dataset, the "% @ km" statistic determines the percentage of guesses that fall within a given kilometer-based distance from the ground-truth location. Just as in the prior work, we evaluate five distance radii: 1 km (roughly street-level accuracy), 25 km (city-level), 200 km (region-level), 750 km (country-level), and 2,500 km (continent-level).

For PIGEON, we run evaluations on a holdout dataset collected from GeoGuessr consisting of 5,000 Street View locations. We separately conduct extensive blind experiments in GeoGuessr deploying PIGEON against human players with varying degrees of expertise as well as a separate match against a world-class professional player. To quantify which parts of our modeling setup impact performance, we further run eight separate ablation studies.

For PIGEOTTO, we focus our evaluations squarely on the benchmark datasets that are established in the literature. Namely, we look at IM2GPS [14], IM2GPS3k [37], YFCC4k [37] and YFCC26k [25] (based on the MediaEval 2016 dataset [20]), and GWS15k [7]. As the last dataset has not been publicly released by the time of this writing, we reconstruct the dataset by exactly replicating the dataset generation procedure outlined in Clark et al. [7].

4.2. Street View evaluation with PIGEON

We present the results of our evaluations of PIGEON and ablations of our contributions in Table 1 and Table 2.

Table 1. Cumulative ablation study of our image geolocalization system on a holdout dataset of 5,000 Street View locations.

Ablation	Country Accuracy %	Mean Error *km*	Median Error *km*	GeoGuessr Score *points*
PIGEON	**91.96**	**251.6**	**44.35**	4,525
– Freezing Last CLIP Layer After Pretraining	91.82	255.1	45.47	**4,531**
– Hierarchical Guess Refinement	91.14	251.9	50.01	4,522
– Contrastive CLIP Pretraining	89.36	316.9	55.51	4,464
– Semantic Geocells	87.96	299.9	60.63	4,454
– Multi-task Prediction Heads	87.90	312.7	61.81	4,442
– Fine-tuning Last CLIP Layer	87.64	315.7	60.81	4,442
– Four-image Panorama	74.74	877.4	131.1	3,986
– Haversine Smoothing	72.12	990.0	148.0	3,890

Table 2. Cumulative ablation study using five common distance radii on a holdout dataset of 5,000 Street View locations.

Ablation	Distance (% @ km)				
	Street 1 km	*City* 25 km	*Region* 200 km	*Country* 750 km	*Continent* 2,500 km
PIGEON	**5.36**	**40.36**	78.28	94.52	**98.56**
– Freezing Last CLIP Layer After Pretraining	4.84	39.86	**78.98**	94.76	98.48
– Hierarchical Guess Refinement	1.32	34.96	78.48	**94.82**	98.48
– Contrastive CLIP Pretraining	1.24	34.54	76.36	93.36	97.94
– Semantic Geocells	1.18	33.22	75.42	93.42	98.16
– Multi-task Prediction Heads	1.10	32.74	75.14	93.00	97.98
– Fine-tuning Last CLIP Layer	1.10	32.50	75.32	92.92	98.00
– Four-image Panorama	0.92	24.18	59.04	82.84	92.76
– Haversine Smoothing	1.28	24.08	55.38	80.20	92.00

As evidenced by our results, each subsequent ablation deteriorates most metrics, pointing to the synergistic nature of the ensemble of methods in our geolocalization system.

Starting from the very bottom of both tables, corresponding to a simple CLIP vision encoder plus a geocell prediction head, we can see that with the introduction of haversine smoothing, the mean distance error decreases by 112.6 kilometers from 990.0 to 877.4 kilometers. The bulkiest performance lift, however, comes from the introduction of a four-image panorama instead of a single image, increasing our country accuracy by 12.9 percentage points and more than halving our median kilometer error from 131.1 to 60.8 kilometers. While fine-tuning the last CLIP layer and sharing parameters in a multi-task setting slightly improves the performance of our model, the uplift is much more palpable with the introduction of our semantic geocells, reducing the median error from 60.6 to 55.5 kilometers. When we additionally pretrain CLIP via our synthetic captions, we gain another 1.7 percentage points in long-range country accuracy. Complemented by our hierarchical location cluster refinement, we improve short-range street-level accuracy from 1.3% to 4.8%. Finally, we freeze the last CLIP layer again and thus prevent parameter sharing between our geocell and multi-task prediction heads, given that our pretraining procedure already incorporates multi-task training. This results in PIGEON's final metrics of a 92.0% country accuracy and a median distance error of 44.4 kilometers.

Beyond our ablations, we compare PIGEON's performance to humans in the game of GeoGuessr. To do so, we develop a Chrome extension bot that has access to PIGEON as an API and deploy our system in a blind experiment across 458 matches, each consisting of multiple rounds. PIGEON comfortably outperforms players in GeoGuessr's Champion Division, consisting of the top 0.01% of human players. The results are shown in Figure 4, underscoring PIGEON's ability to beat players of all skill levels. Notably, top GeoGuessr players perform orders of magnitudes better than the players evaluated in Seo et al. [30].

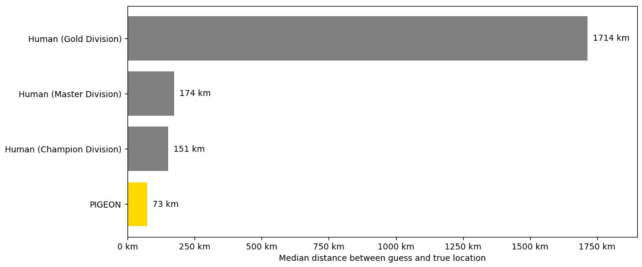


Figure 4. Geolocalization error of PIGEON against human players of various in-game skill levels across 458 multi-round matches. The Champion Division consists of the top 0.01% of players. PIGEON's error is higher than in Table 1 because GeoGuessr round difficulties are adjusted dynamically, increasing with every round.

For our final evaluation, we challenge one of the world's foremost professional GeoGuessr players to a match and win six out of six planet-scale, multi-round games.[3] PIGEON is the first model to reliably beat a GeoGuessr professional.

4.3. Benchmark evaluation with PIGEOTTO

The results of our evaluations of PIGEOTTO on benchmark datasets are displayed in Table 3. PIGEOTTO achieves state-of-the-art (SOTA) performance on every single benchmark dataset and on the majority of distance-based granularities. On IM2GPS, it is able to improve the state of the art on both country-level and continent-level accuracy by 2 percentage points or more. Its relative underperformance on smaller granularities can be attributed to the landmark-only nature of IM2GPS and its small size of 237 images. On a larger and more general dataset, IM2GPS3k, PIGEOTTO performs much better, achieving SOTA performance on all but the street-level metric, with an impressive 11.4 percentage-point improvement on the country level and a much lower median error of 147.3 kilometers. Meanwhile, on YFCC4k and YFCC26k, PIGEOTTO is able to outperform the current state of the art on 9 out of 10 metrics, including by 12.2 percentage points on the country level on YFCC4k and by 13.6 percentage points on YFCC26k, more than halving the previous SOTA median error. Finally, we see very significant improvements on the most recently released benchmark, GWS15k, consisting entirely of

[3] https://www.youtube.com/watch?v=ts5lPDV--cU.

Table 3. Comparison of PIGEOTTO's results against other models on benchmark datasets. PIGEOTTO reduces the median kilometer error by 2-5x on benchmarks not solely focused on landmarks.

Benchmark	Method	Median Error km	Distance (% @ km) *Street* 1 km	*City* 25 km	*Region* 200 km	*Country* 750 km	*Continent* 2,500 km
IM2GPS [14]	PlaNet [38]	> 200	8.4	24.5	37.6	53.6	71.3
	CPlaNet [30]	> 200	16.5	37.1	46.4	62.0	78.5
	ISNs(M,f^*,S_3) [25]	> 25	16.9	43.0	51.9	66.7	80.2
	Translocator [27]	> 25	19.9	48.1	64.6	75.6	86.7
	GeoDecoder [7]	~ **25**	**22.1**	**50.2**	**69.0**	80.0	89.1
	PIGEOTTO (Ours)	70.5	14.8	40.9	63.3	**82.3**	**91.1**
	Δ (% points)		-7.3	-9.3	-5.7	+2.3	+2.0
IM2GPS3k [37]	PlaNet [38]	> 750	8.5	24.8	34.3	48.4	64.6
	CPlaNet [30]	> 750	10.2	26.5	34.6	48.6	64.6
	ISNs(M,f^*,S_3) [25]	~ 750	10.5	28.0	36.6	49.7	66.0
	Translocator [27]	> 200	11.8	31.1	46.7	58.9	80.1
	GeoDecoder [7]	> 200	**12.8**	33.5	45.9	61.0	76.1
	PIGEOTTO (Ours)	**147.3**	11.3	**36.7**	**53.8**	**72.4**	**85.3**
	Δ (% points)		-1.5	+3.2	+7.9	+11.4	+9.2
YFCC4k [37]	PlaNet [38]	> 750	5.6	14.3	22.2	36.4	55.8
	CPlaNet [30]	> 750	7.9	14.8	21.9	36.4	55.5
	ISNs(M,f^*,S_3) [25]	> 750	6.7	16.5	24.2	37.5	54.9
	Translocator [27]	> 750	8.4	18.6	27.0	41.1	60.4
	GeoDecoder [7]	~ 750	10.3	**24.4**	33.9	50.0	68.7
	PIGEOTTO (Ours)	**383.0**	**10.4**	23.7	**40.6**	**62.2**	**77.7**
	Δ (% points)		+0.1	-0.7	+6.7	+12.2	+9.0
YFCC26k [25]	PlaNet [38]	> 2,500	4.4	11.0	16.9	28.5	47.7
	ISNs(M,f^*,S_3) [25]	~ 2,500	5.3	12.3	19.0	31.9	50.7
	Translocator [27]	> 750	7.2	17.8	28.0	41.3	60.6
	GeoDecoder [7]	~ 750	10.1	23.9	34.1	49.6	69.0
	PIGEOTTO (Ours)	**333.3**	**10.5**	**25.8**	**42.7**	**63.2**	**79.0**
	Δ (% points)		+0.4	+1.9	+8.6	+13.6	+10.0
GWS15k [7]	ISNs(M,f^*,S_3) [25]	> 2,500	0.05	0.6	4.2	15.5	38.5
	Translocator [27]	> 2,500	0.5	1.1	8.0	25.5	48.3
	GeoDecoder [7]	~ 2,500	**0.7**	1.5	8.7	26.9	50.5
	PIGEOTTO (Ours)	**415.4**	**0.7**	**9.2**	**31.2**	**65.7**	**85.1**
	Δ (% points)		+0.0	+7.7	+22.5	+38.8	+34.6

Street View images. Crucially, GWS15k is the most difficult dataset in the benchmark set. If we define images to be taken in the same location if they are less than 100 meters apart, 92% of locations in GWS15k are not taken in the same location as any MediaEval 2016 [20] training data on which prior SOTA models and our system were trained. For comparison, this number ranges from 23% to 42% for the other four benchmark datasets, underscoring the unique difficulty of GWS15k. Noting that PIGEOTTO was not trained on any Street View images, this suggests that PIGEOTTO is truly planet-scale in nature, exhibits robust behavior to distribution shifts, and is the first geolocalization model that effectively generalizes to unseen places.

5. Ethical considerations

Image geolocalization represents a sub-discipline of computer vision that comes with both potential benefits to society as well as with risks of misuse. While prior work in the field addresses ethical implications scantily, we believe that the potential misuse and negative downstream implications of image geolocalization systems afford a separate discussion section in this paper.

On the one hand, accurate geo-tagging of images opens up possibilities for various beneficial applications, far beyond the game of GeoGuessr, including helping to understand changes to particular locations over time. Image geolocalization has found use cases in autonomous driving, navigation, geography education, open-source intelligence, and visual investigations in journalism.

On the other hand, however, applications of image geolocalization may come with risks, especially if the precision of such systems significantly improves in the future. To our knowledge, this is the first state-of-the-art image geolocalization paper in the last five years that is not funded by military contracts. Recently published work has been supported by grants from the Department of Defense [27] and the US Army [7]. Any attempts to develop image geolocalization technology for military use cases should come under particular scrutiny. There are also privacy risks involved; for instance, some methods using Street View images have been shown to be capable of inferring local income, race, education, and voting patterns [11].

Image geolocalization technologies come with dual-use risks [15], and efforts need to be made to minimize harmful consequences. To that end, we decide not to release model weights publicly and only release our code for academic validation. While a major limitation of today's image geolocalization technologies (including ours) is that they are unable to make street-level predictions reliably, researchers ought to carefully consider the risk of potential misuse of their work as such technologies get increasingly precise.

6. Conclusion

We propose a novel deep multi-task approach for planet-scale image geolocalization that achieves state-of-the-art benchmark results while being robust to distribution shifts.

To confirm the efficacy of our approach, we train and evaluate two distinct image geolocalization models. First, we gather a global Street View dataset to train PIGEON, a multi-task model that places into the top 0.01% of human players in the game of GeoGuessr. On a holdout dataset of 5,000 Street View locations, 40.4% of PIGEON's predictions of geographic coordinates land within a 25-kilometer radius of the ground-truth location. Subsequently, we assemble a planet-scale dataset of over 4 million images derived from Flickr and Wikipedia to train the more general PIGEOTTO, improving the state of the art on a wide range of geolocalization benchmark datasets by a large margin.

Going forward, it remains to be seen whether applied image geolocalization technologies will be truly planet-scale or focused on a well-defined narrow distribution. In any case, our findings about the importance of semantic geocell creation, multimodal contrastive pretraining, and precise intra-geocell refinement, among others, point to important building blocks for such systems. Nevertheless, deployment of any downstream image geolocalization technology will need to balance potential benefits with possible risks, ensuring the responsible development of future computer vision systems.

References

[1] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating CLIP: Towards Characterization of Broader Capabilities and Downstream Implications, 2021. 3

[2] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. OPTICS: Ordering Points to Identify the Clustering Structure. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, page 49–60, New York, NY, USA, 1999. Association for Computing Machinery. 4, 6

[3] Georges Baatz, Olivier Saurer, Kevin Köser, and Marc Pollefeys. Large Scale Visual Geo-Localization of Images in Mountainous Terrain. 2

[4] 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*, 5(1):180214, 2018. 5, 3, 4

[5] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking Visual Geo-localization for Large-Scale Applications, 2022. 2

[6] Liangliang Cao, John R. Smith, Zhen Wen, Zhijun Yin, Xin Jin, and Jiawei Han. BlueFinder: Estimate Where a Beach Photo Was Taken. In *Proceedings of the 21st International Conference on World Wide Web*, page 469–470, New York, NY, USA, 2012. Association for Computing Machinery. 2

[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, 2023. 1, 3, 4, 6, 8

[8] David Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. Mapping the World's Photos. In *WWW '09: Proceedings of the 18th International Conference on World Wide Web*, pages 761–880, 2009. 2

[9] Alexandre de Brébisson, Étienne Simon, Alex Auvolat, Pascal Vincent, and Yoshua Bengio. Artificial Neural Networks Applied to Taxi Destination Prediction, 2015. 3, 5

[10] GADM. GADM Version 4.1, 2022. 4, 1, 3

[11] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences*, 114(50): 13108–13113, 2017. 8

[12] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 5

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

[14] James Hays and Alexei A. Efros. IM2GPS: estimating geographic information from a single image. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008. 1, 2, 3, 6, 8

[15] Peter Henderson, Eric Mitchell, Christopher D. Manning, Dan Jurafsky, and Chelsea Finn. Self-destructing models: Increasing the costs of harmful dual uses of foundation models, 2023. 8

[16] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, 2017. 3

[17] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. 2021. 3

[18] Giorgos Kordopatis-Zilos, Panagiotis Galopoulos, Symeon Papadopoulos, and Ioannis Kompatsiaris. Leveraging EfficientNet and Contrastive Learning for Accurate Global-scale Location Estimation, 2021. 3

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2012. 3

[20] Martha Larson, Mohammad Soleymani, Guillaume Gravier, Bogdan Ionescu, and Gareth J.F. Jones. The Benchmarking Initiative for Multimedia Evaluation: MediaEval 2016. *IEEE MultiMedia*, 24(1):93–96, 2017. 6, 8, 2, 3

[21] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-View Image Geolocalization. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2013. 3

[22] Jessica Lucas. A Geography Game Has Its First Superstar. Can It Survive Its First Player Revolt?, 2023. 1

[23] Grace Luo, Giscard Biamby, Trevor Darrell, Daniel Fried, and Anna Rohrbach. G^3: Geolocation via Guidebook Grounding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5841–5853, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. 1, 3

[24] Carlo Masone and Barbara Caputo. A Survey on Deep Visual Place Recognition. *IEEE Access*, 9:19516–19547, 2021. 3

[25] Eric Müller-Budack, Kader Pustu-Iren, and Ralph Ewerth. Geolocation Estimation of Photos Using a Hierarchical Model and Scene Classification. In *Computer Vision – ECCV 2018*, pages 575–592, Cham, 2018. Springer International Publishing. 1, 3, 4, 6, 8, 9

[26] OpenAI. GPT-4V(ision) System Card, 2023. 3

[27] 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, 2022. 1, 3, 4, 5, 8

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, 2021. 3, 5

[29] Olivier Saurer, Georges Baatz, Kevin Köser, L'ubor Ladický, and Marc Pollefeys. Image Based Geo-localization in the

Alps. *International Journal of Computer Vision*, 116(3): 213–225, 2016. 2

[30] Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. CPlaNet: Enhancing Image Geolocalization by Combinatorial Partitioning of Maps, 2018. 3, 7, 8

[31] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical Networks for Few-shot Learning. *CoRR*, abs/1703.05175, 2017. 3, 6

[32] Sudharshan Suresh, Nathaniel Chodosh, and Montiel Abello. DeepGeo: Photo Localization with Deep Neural Network, 2018. 2, 3

[33] Jonas Theiner, Eric Mueller-Budack, and Ralph Ewerth. Interpretable Semantic Photo Geolocation, 2021. 3, 4, 5, 1

[34] Jan Tomešek, Martin Čadík, and Jan Brejcha. CrossLocate: Cross-Modal Large-Scale Visual Geo-Localization in Natural Environments Using Rendered Modalities. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3174–3183, 2022. 2

[35] Eric Tzeng, Andrew Zhai, Matthew Clements, Raphael Townshend, and Avideh Zakhor. User-Driven Geolocation of Untagged Desert Imagery Using Digital Elevation Models. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 237–244, 2013. 2

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, 2017. 3

[37] Nam Vo, Nathan Jacobs, and James Hays. Revisiting IM2GPS in the Deep Learning Era, 2017. 1, 3, 6, 8

[38] Tobias Weyand, Ilya Kostrikov, and James Philbin. PlaNet - Photo Geolocation with Convolutional Neural Networks. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 3, 8

[39] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google Landmarks Dataset v2 – A Large-Scale Benchmark for Instance-Level Recognition and Retrieval, 2020. 6, 3

[40] Meiliu Wu and Qunying Huang. IM2City: Image Geo-Localization via Multi-Modal Learning. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, page 50–61, New York, NY, USA, 2022. Association for Computing Machinery. 2, 3

[41] Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view Geo-localization with Layer-to-Layer Transformer. In *Advances in Neural Information Processing Systems*, pages 29009–29020. Curran Associates, Inc., 2021. 3

[42] Amir Roshan Zamir and Mubarak Shah. Accurate Image Localization Based on Google Maps Street View. In *Computer Vision – ECCV 2010*, pages 255–268, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. 2

[43] Amir Roshan Zamir and Mubarak Shah. Image Geo-Localization Based on Multiple Nearest Neighbor Feature Matching UsingGeneralized Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1546–1558, 2014. 2

[44] Sijie Zhu, Mubarak Shah, and Chen Chen. TransGeo: Transformer Is All You Need for Cross-view Image Geo-localization, 2022. 3