Geolocation Estimation of Photos using a Hierarchical Model and Scene Classification

Eric Müller-Budack[1,2][0000−0002−6802−1241], Kader Pustu-Iren[1][0000−0003−2891−9783], and Ralph Ewerth[1,2][0000−0003−0918−6297]

[1] Leibniz Information Centre for Science and Technology (TIB), Hannover, Germany
[2] L3S Research Center, Leibniz Universität Hannover, Germany
{eric.mueller,kader.pustu,ralph.ewerth}@tib.eu

Abstract. While the successful estimation of a photo's geolocation enables a number of interesting applications, it is also a very challenging task. Due to the complexity of the problem, most existing approaches are restricted to specific areas, imagery, or worldwide landmarks. Only a few proposals predict GPS coordinates without any limitations. In this paper, we introduce several deep learning methods, which pursue the latter approach and treat geolocalization as a classification problem where the earth is subdivided into geographical cells. We propose to exploit hierarchical knowledge of multiple partitionings and additionally extract and take the photo's scene content into account, i.e., indoor, natural, or urban setting etc. As a result, contextual information at different spatial resolutions as well as more specific features for various environmental settings are incorporated in the learning process of the convolutional neural network. Experimental results on two benchmarks demonstrate the effectiveness of our approach outperforming the state of the art while using a significant lower number of training images and without relying on retrieval methods that require an appropriate reference dataset.

Keywords: Geolocation Estimation · Scene Classification · Deep Learning · Context-based Classification

1 Introduction

Predicting the geographical location of photos without any prior knowledge is a very challenging task, since images taken from all over the earth depict a huge amount of variations, e.g., different daytimes, objects, or camera settings. In addition, the images are often ambiguous and therefore provide only very few visual clues about their respective recording location. For these reasons, the majority of approaches simplifies photo geolocalization by restricting the problem to urban photos of, for example, well-known landmarks and cities [3,25,34,43,45,48] or natural areas like deserts or mountains [5,33,38]. Only a few frameworks treat the task at global-scale without relying on specific imagery [13,14,39,42] or any other prior assumptions. These approaches particularly benefit from the advancements in deep learning [15,16,21] and the increasing number of publicly available large-scale image collections from platforms such as *Flickr*. Due to the complexity

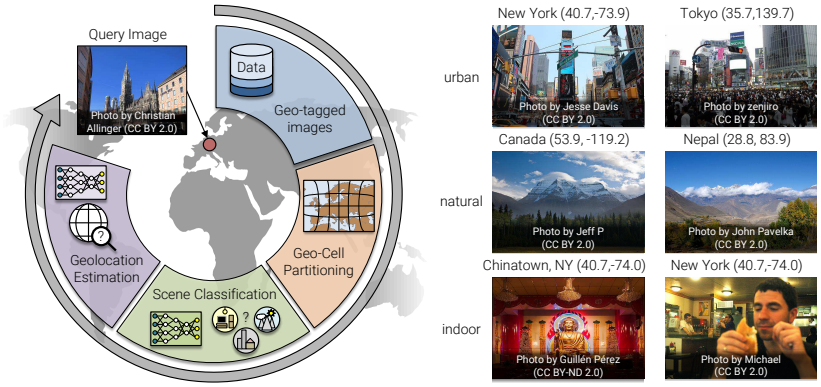

Fig. 1. Left: Workflow of the proposed geolocation estimation approach. Right: Sample images of different locations for specific scene concepts.

of the problem and the unbalanced distribution of photos taken from all over the earth, methods based on convolutional neural networks (CNNs) [39,42] treat photo geolocalization as a classification task subdividing the earth into geographical cells with a similar number of images. However, according to Vo et al. [39], even current CNNs are not able to memorize the visual appearance of the entire earth and to simultaneously learn a model for scene understanding. Moreover, geographical partitioning approaches [39,42] entail a trade-off problem. While a finer partitioning leads to a higher accuracy at city-scale (location error less than 1 km), a coarser subdivision increases the performance at country-scale (750 km). In our opinion, one main reason for these problems is the huge diversity caused by various environmental settings, which requires specific features to distinguish different locations. Referring to Figure 1, we argue that urban images mainly differ in, e.g., architecture, people, and specific objects like cars or street signs. On the contrary, natural scenes like forests or indoor scenarios are most likely defined by features encoding the flora and fauna or the style of the interior furnishings, respectively. Therefore, we claim that photo geolocalization can greatly benefit from contextual knowledge about the environmental scene, since the diversity in the data space could be drastically reduced.

In this paper, we address the aforementioned problems by (1) incorporating hierarchical knowledge at different spatial resolutions in a multi-partitioning approach, as well as (2) extracting and taking information about the respective type of environmental settings (e.g., *indoor*, *natural*, and *urban*) into account. We consider photo geolocalization as a classification task by subdividing the earth into geographical cells with a balanced number of images (similar to *PlaNet* [42]). There are several contributions. We combine the outputs from all scales to exploit the hierarchical information of a CNN that is trained simultaneously with labels from multiple partitionings to encode local and global information. Furthermore, we suggest two strategies to include information about the respective scene type: (a) deep networks that are trained separately with images

of distinctive scene categories, and (b) a multi-task network trained with both geographical and scene labels. This should enable the CNN to learn specific features for estimating the GPS (Global Positioning System) coordinate of images in different environmental surroundings. The workflow is illustrated in Figure 1.

To the best of our knowledge, this is the first approach that considers scene classification and exploits hierarchical (geo)information to improve unrestricted photo geolocalization. Furthermore, we have used a state of the art CNN architecture and our comprehensive experiments include an evaluation of the impact of different scene concepts. Experimental results on two different benchmarks demonstrate that our approach outperforms the state of the art without relying on image retrieval techniques (*Im2GPS* [13,14,39]), while using a significant lower number of training images compared to *PlaNet* [42] – making our approach more feasible.

The remainder of the paper is organized as follows. In Section 2, we review related work on photo geolocation estimation. The proposed framework to extract and utilize visual concepts of specific scenes and multiple earth partitionings to estimate the GPS coordinates of images is introduced in Section 3. Experimental results on two different benchmarks are presented and discussed in Section 4. Section 5 concludes the paper and outlines areas of future work.

2 Related Work

Related work on visual geolocalization can be roughly divided into two categories: (1) proposals which are restricted to specific environments or imagery, and (2) approaches at planet-scale without any restrictions. In this section, we focus on the second category since it is more closely related to our work. For a more comprehensive review, we refer to Brejcha and Čadík's survey [8].

Many proposals of the first category are introduced at city-scale resolution restricting the problem to specific cities or landmarks. These mainly apply retrieval techniques to match a query image against a reference dataset [3,12,18,20,29,34,46]. Approaches that focus on landmark recognition use either a pre-defined set of landmarks or cluster a given photo collection in an unsupervised manner to retrieve the most interesting areas for geolocalization [4,23,28,48]. Other proposals match query images against 3D models of cities [10,19,24,27,30]. However, the underlying data collections of these methods are restricted to popular scenes and urban environments and therefore lack accuracy when predicting photos that do not have (many) instance matches. For this reason, some approaches additionally make use of satellite aerial imagery to enhance the geolocalization in sparsely covered regions [35,40,44,45]. In this context, solutions are presented that match an aerial query image against a reference dataset containing satellite images in a wide baseline approach [2,6,43]. Some of these proposals [25,26] even address geolocation at planet-scale. But since these frameworks require a reference dataset that contains satellite images, we still consider them as restricted frameworks. Only a minority of proposals has been designed for natural geolocalization of images depicting beaches [9,41], deserts [38], or mountains [5,33].

All of the aforementioned proposals are restricted to well-covered regions, specific imagery, or environmental scenes. As a first attempt for planet-scale geolocation estimation, Hays and Efros [13] have introduced *Im2GPS*. They use a retrieval approach to match a given query image based on a combination of six global image descriptors to a reference dataset consisting of more than six million GPS-tagged images. The authors extend *Im2GPS* [14] by incorporating information on specific geometrical classes like sky and ground as well as an improved retrieval technique. Weyand et al. [42] have introduced *PlaNet*, where the task of geolocalization is treated as a classification problem. The earth is adaptively subdivided into geographical cells with a similar number of images that are used to train a convolutional neural network. This approach noticeably outperformed *Im2GPS*, which encouraged Vo et al. [39] to learn a feature representation with a CNN to improve the *Im2GPS* framework. Using the extracted features of a query photo, the (k)-nearest neighbors in the reference dataset based on kernel density estimation are retrieved. In this way, a multi-partitioning approach is introduced to simultaneously learn photo-geolocation at different spatial resolutions. However, in contrast to our work this approach does not make use of the hierarchical knowledge given by the predictions at each scale.

3 Hierarchical Geolocalization using Scene Classification

In this section, we present the proposed deep learning framework for geolocation estimation. According to *PlaNet* [42], we treat the task as a classification problem by subdividing the earth into geographical cells C that contain a similar number of images (Section 3.1). In contrast to previous work, we exploit contextual information of the environmental scenario solely using the visual content of a given photo to improve the localization accuracy. Therefore, we assign scene labels to all the images based on the 365 categories of the *Places2* dataset [49] (Section 3.2). Several approaches that are aimed at integrating the extracted information about the given type of scene and multiple geographical cell partitionings are introduced in Section 3.3. Finally, we explain how the proposed approaches are applied to estimate the GPS coordinates of images based on the predicted geo-cell probabilities $\hat{C}$ (Section 3.4). In this context, we introduce our hierarchical approach to combine the results of multiple spatial resolutions. An overview of the proposed framework is presented in Figure 2.

3.1 Adaptive Geo-Cell Partitioning

The *S2 geometry library*[3] is utilized to generate a set of non-overlapping geographical cells C. In more detail, the earth's surface is projected on an enclosing cube with six sides representing the initial cells. An adaptive hierarchical subdivision based on the GPS coordinates of the images is applied [42], where each cell is the node of a quad-tree. Starting at the root nodes, the respective quad-tree is subdivided recursively until all cells contain a maximum of τ_{max} images.

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

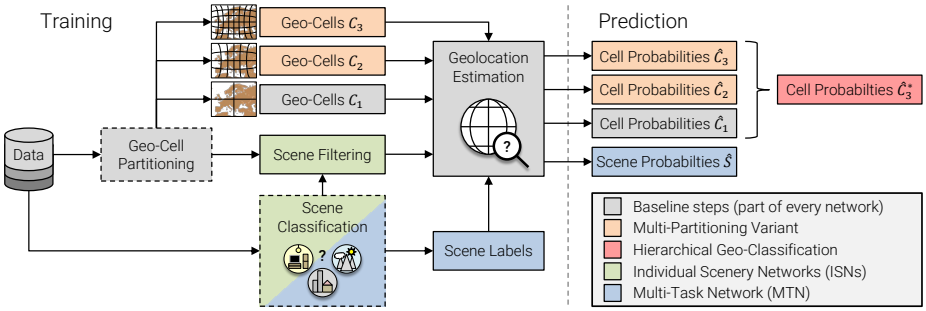

Fig. 2. Pipeline of the proposed geolocation estimation frameworks. Gray: Baseline steps that are part of every network. Additional steps are visualized in different colors. Dashed elements are applied to all images before the training process takes place.

Afterwards, all resulting cells with less than τ_{min} photos are discarded, because they most likely cover areas like poles or oceans which are hard to distinguish.

This approach has several advantages compared to a subdivision of the earth into cells with roughly equally areas. On the one hand side, an adaptive subdivision prevents dataset biases and allows to create classes with a similar number of images. On the other hand, fine cells in photographically well covered areas are generated. This enables a more accurate prediction of image locations which most likely depict interesting regions such as landmarks or cities.

3.2 Visual Scene Classification

To classify scenes and extract scene labels, the ResNet model [16] with 152 layers[4] of the *Places2* dataset [49] is applied. The model has been trained on more than 16 million training images from 365 different place categories. This fits nicely with our approach, since the resulting classifier already distinguishes images that depict specific environments. We predict the scene labels based on the scene set S_{365} of all training images using the maximum probability of the output vector. Based on the provided scene hierarchy[5], we additionally extract labels of the sets S_{16} and S_3 containing 16 and three superordinate scene categories, respectively. We add the probabilities of all classes which are assigned to the same superordinate category and generate the corresponding label. However, some scenes like *barn* are allocated to multiple superordinate categories (*outdoor, natural*; *outdoor, man-made*), because they visually overlap. For this reason, we first divide the probability of these classes by the number of assigned categories to maintain the normalization. Please note, that we use the terms *natural* for "*outdoor, natural*" and *urban* for "*outdoor, man-made*" in the rest of the paper.

[4] *Places2* ResNet152 model: https://github.com/CSAILVision/places365
[5] *Places2* scene hierarchy: http://places2.csail.mit.edu/download.html

3.3 Geolocation Estimation

In this section, several approaches based on convolutional neural networks for an unrestricted planet-scale geolocalization are introduced. First, we present a baseline approach which is trained without using scene information and multiple geographical partitionings. In the following, we describe how the information for different spatial resolutions as well as environmental concepts are integrated in the training process. In this context, two different approaches to utilize visual scene labels are proposed. An overview is provided in Figure 2.

Baseline: To evaluate the impact of the suggested approaches for geolocalization, we first present a baseline system that does not rely on information about the environmental setting and different spatial resolutions. Therefore, we generate a single geo-cell partitioning C as described in Section 3.1. For classification, we add a fully-connected layer on top of the global pooling layer of the ResNet architecture [16], where the number of output neurons corresponds to the number of geo-cells $|C|$. During training the cross-entropy geolocalization loss L_{geo}^{single} based on the probability distribution $\hat{C}$ and the ground-truth cell label encoded in a one-hot vector $\hat{C}_{GT}$ is minimized.

Multi-Partitioning Variant: We propose to simultaneously learn geolocation estimation at multiple spatial resolutions (according to Vo et al. [39]). In contrast to the baseline approach, we add a fully-connected layer for the geographical cells of all partitionings $P = \{C_1, \ldots, C_n\}$. The multi-partitioning classification loss L_{geo}^{multi} is calculated using the mean of the loss values L_{geo}^{single} for every partitioning. As a consequence, the CNN is able to learn geographical features at different scales resulting in a more discriminative classifier. However, in contrast to Vo et al. [39] we further exploit the hierarchical knowledge for the final prediction. The details are presented in Section 3.4.

Individual Scene Networks (ISNs): In a first attempt to incorporate context information about the environmental setting for photo geolocalization, individual networks for images depicting a specific scene are trained. For each photograph, we extract the scene probabilities using the scene classification presented in Section 3.2. During the training, every image with a scene probability greater than a threshold of τ_S is used as input for the respective *Individual Scene Network (ISN)*. Following this approach offers the advantage, that the network is solely trained on images depicting specific environmental scenarios. It greatly reduces the diversity in the underlying data space and enables the network to learn more specific features. On the contrary, it is necessary to train individual models for each scene concept, which is hard to manage if the number of different concepts $|S|$ becomes larger. For this reason, we suggest to fine-tune a model, which was initially trained without scene restriction, with images of the respective environmental category.

Multi-Task Network (MTN): Since the aforementioned method for geolocation estimation may become infeasible for a large amount of different environmental concepts, we aim for a more practicable approach using a network which treats photo geolocalization and scene recognition as a multi-task problem. In order to encourage the network to distinguish between images of different environmental scenes, we simultaneously train two classifiers for these complementary tasks. Adding another (complementary) task has proven to be efficient to improve the results of the main task [7,17,32,47]. More specifically, an additional fully-connected layer on top of the global pooling layer of the ResNet CNN architecture [16] is utilized. The number of output neurons of this layer corresponds to the amount of scene categories $|S|$. The weights of all other layers in the network are completely shared. In addition, the scene loss L_{scene} based on the ground-truth one-hot vector $\hat{S}_{GT}$ and the scene probabilities $\hat{S}$ is minimized using the cross-entropy loss. The total loss L_{total} of the *Multi-Task Network (MTN)* is defined by the sum of the geographical and scene loss.

3.4 Predicting Geolocations using Hierarchical Spatial Information

In order to estimate the GPS coordinate from the classification output, we apply the trained models from Section 3.3 on three evenly sampled crops of a given query image according to its orientation. Afterwards, the mean of the resulting class probabilities of each crop is calculated. Please note that an additional step for testing is necessary for the *Individual Scene Networks*. In this case, the scene label is first predicted using the maximum probability as described in Section 3.2 in order to feed the image into the respective *ISN* for geolocalization.

Standard Geo-Classification: Without relying on hierarchical information, we solely utilize the probabilities $\hat{C}$ of one given geo-cell partitioning C. In this respect, we assign the class label with the maximum probability to predict the geographical cell. Applying the multi-partitioning approach in Section 3.3 we are therefore able to obtain $|P|$ class probabilities at different spatial resolutions. In our opinion, the probabilities at all scales should be exploited to enhance the geolocalization and to combine the capabilities of all partitionings.

Hierarchical Geo-Classification: To ensure that every geographical cell in the finest representation can be uniquely connected to a larger parent area in an upper-level, a fixed threshold parameter τ_{min} for the adaptive subdivision (Section 3.1) is applied. Thus, we are able to generate a geographical hierarchy from the different spatial resolutions. Inspired by the hierarchical object classification approach from *YOLO9000* [31], we multiply the respective probabilities at each level of the hierarchy. Consequently, the prediction for the finest subdivision can be refined by incorporating the knowledge of coarser representations.

Class2GPS: Depending on the predicted class we extract the GPS coordinates of the given query image. In contrast to Weyand et al. [42], we use the mean

Table 1. Number of classes $|C|$ for each partitioning C with different thresholds τ_{min} and τ_{max}.

C	τ_{min}	τ_{max}	$\lvert C\rvert$
coarse	50	5,000	3,298
middle	50	2,000	7,202
fine	50	1,000	12,893

Table 2. Top-1 and Top-5 accuracy on the validation set of the *Places2* benchmark [49] for different scene hierarchies.

Hierarchy	**Top-1**	**Top-5**
S_3	91.5 %	—
S_{16}	72.1 %	97.1 %
S_{365}	45.7 %	77.3 %

location of all training images in the predicted cell instead of the geographical center. This is more precise for regions containing an interesting area where the majority of photos is taken. Imagine a geographical cell centered around an ocean and a city which is located at the cell boundary. In this example, the error using the geographical center would be very high, even if it is clear that the photo was most likely taken in the city.

4 Experimental Setup and Results

Training Data: We use a subset of the Yahoo Flickr Creative Commons 100 Million dataset (*YFCC100M*) [37] as input data for our approach. This subset was introduced for the MediaEval Placing Task 2016 (*MP-16*) [22] and includes around five million geo-tagged images[6] from Flickr without any restrictions. The dataset contains ambiguous photos of, e.g., indoor environments, food, and humans for which the location is difficult to predict. Like Vo et al. [39] we exclude images from the same authors as in the test datasets, which we use for evaluation. A ResNet model [15] is used which has been pre-trained on ImageNet [11] to avoid duplicate images by comparing the resulting feature vectors from the last pooling layer. Overall, our training dataset consists of $|I| = 4{,}723{,}695$ images.

Partitioning Parameters: As explained in Section 3.4, we choose a constant value of $\tau_{min} = 50$ (according to PlaNet [42]) as the minimum threshold for the adaptive subdivision, to enable the hierarchical classification approach. Our goal is to train the geolocation at multiple spatial resolutions. Therefore, the following maximum thresholds $\tau_{max} \in \{1{,}000; 2{,}000; 5{,}000\}$ are used. We select these thresholds because the *MP-16* dataset has approximately 16 times less images than PlaNet [42] and we therefore aim to produce around $\sqrt{16}$ less classes (PlaNet has 26,263 cells) at the middle representation. Since we want to show how fine and coarse representations can be efficiently combined, the other thresholds are specified to produce circa two times more and less classes than the middle representation. The resulting number of classes $|C|$ for different partitionings to train our deep learning approaches are shown in Table 1.

[6] Available at: http://multimedia-commons.s3-website-us-west-2.amazonaws.com

Scene Classification Parameters: The performance of the concept classification (Section 3.2) is evaluated on the *Places2* validation dataset [49] containing 36,500 images (100 for each scene). In Table 2 results for the different scene hierarchy levels are reported. The quality of the scene classification is very crucial for the *ISNs* presented in Section 3.3, because it defines the underlying data space. Since the top-1 accuracy of 91.5 % already provides a good basis, we focus on a set of three scene concepts $S_3 = \{indoor, natural, urban\}$. Furthermore, this limits the amount of *ISNs* to a feasible number of three concepts. We suggest to apply a small threshold of $\tau_S = 0.3$. Admittedly, this selection is somewhat arbitrary, but we intend to use images with similar scene probabilities as input for each *ISN*. This could be especially useful for images depicting rural areas, because they share visual information like architecture as well as flora and fauna that are beneficial for both environmental categories *urban* and *natural*. The scene filtering yields a total of around 1.80M, 1.42M, and 2.34M training images for the concepts *indoor*, *natural*, *urban*, respectively.

Network Training: The proposed approaches are trained using a ResNet architecture [16] with 101 convolutional layers. The weights are initialized by a pre-trained ImageNet model [11]. To avoid overfitting, the data is augmented by randomly selecting an area which covers at least 70 % of the image with an aspect ratio R between $3/4 \leq R \leq 4/3$. Furthermore, the input images are randomly flipped and subsequently cropped to 224×224 pixels. We use the Stochastic Gradient Descend (SGD) optimizer with an initial learning rate of 0.01, a momentum of 0.9, and a weight decay of 0.0001. The learning rate is exponentially lowered by a factor of 0.5 after every five training epochs. We initially train the networks for 15 epochs and a batch size of 128. We validate the CNNs on 25,600 images of the *YFCC100M* dataset [37].

As described in Section 3.3, it could be beneficial to fine-tune the *ISNs* based on a model which was initially trained without scene restriction. For a fair comparison, all models are therefore fine-tuned for five epochs or until the loss on the validation set converges. In this respect, the initial learning rate is decreased to 0.001. Finally, the best model on the validation set is used for conducting the experiments. The implementation is realized using the TensorFlow library [1] in Python. The trained models and all necessary data to reproduce our results are available at: https://github.com/TIBHannover/GeoEstimation

Test Setup: We evaluate our approaches on two public benchmarks datasets for geolocation estimation. The *Im2GPS* test dataset [13] contains 237 photos, where 5 % are depicting specific tourist sites and the remaining are only recognizable in a generic sense. Because this benchmark is very small, Vo et al. [39] introduced a new datasets called *Im2GPS3k* that contains 3,000 images from *Im2GPS* (2,997 images are provided with a GPS tag). The great circle distance (GCD) between the predicted and ground-truth image location is calculated for evaluation. As suggested by Hays and Efros [13], we report the geolocalization accuracy as the percentage of test images that are predicted within a certain distance to the

Table 3. Notation of the geolocalization approaches. T denotes whether the network was trained with a single/lone (L) or multiple (M) partition(s). $C \in \{c, m, f\}$ indicates which cell partition (coarse (c), middle (m), fine (f)) is used for classification. If C is denoted with a star (*) the hierarchical classification is utilized.

Notation	Description
base (T, C)	*Baseline* trained without scene information
ISNs (T, C, S_3)	*Individual Scene Networks* using the scene set S_3
MTN (T, C, S)	*Multi-Task Network* using a scene set $S \in \{S_3, S_{16}, S_{365}\}$

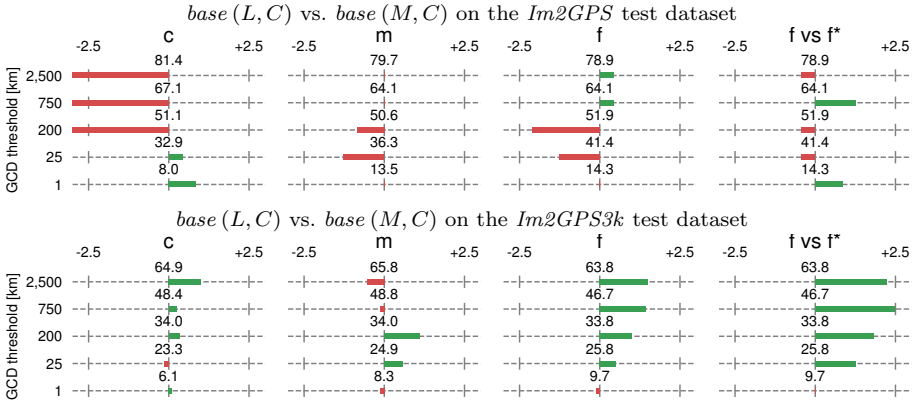

Fig. 3. Comparison of the geolocation approaches trained with and without multiple subdivisions for different geo-cell partitionings C. First mentioned approach *base* (L, C) is used as reference and its accuracy is denoted in the middle of the x-axis.

ground-truth location. The notations of the proposed approaches are presented in Table 3. The most significant results using the suggested multi-partitioning and scene concepts for geolocalization as well a comparison to the state of the art methods are given in the related Sections. A complete list of results is provided in the supplemental material.

4.1 Evaluating the Multi-Partitioning Approach

The results for the baseline and the multi-partitioning approach are displayed in Figure 3. Surprisingly, no significant improvement using multiple partitionings can be observed for the *Im2GPS* test dataset. But it is clearly visible that the results especially for the *fine* partitioning have improved for the *Im2GPS3k* dataset, which is more representative due to its larger size. This demonstrates that the network is able to incorporate features at different spatial resolutions and utilizes this knowledge to learn a more discriminative classifier. A similar observation was made in the latest *Im2GPS* approach [39]. However, by exploiting the hierarchical knowledge at different spatial resolutions the localization accuracy can be indeed further increased. Figure 4 shows that the geo-location of the

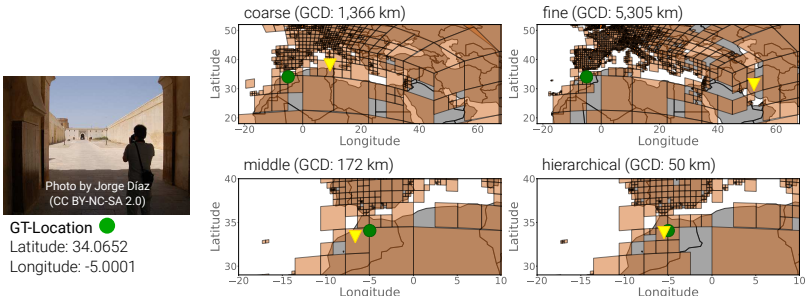

Fig. 4. Quantitative result using the prediction of the different partitioning output layers as well as the hierarchical result.

Table 4. Number of images on the evaluation datasets for different scene concepts in S_3.

Scenes	Im2GPS	Im2GPS3k
all	237	2,997
indoor	19	545
natural	80	845
urban	138	1,607

Table 5. Top-1 and Top-5 scene classification accuracies on the validation set of the *Places2* benchmark [49] for different *Multi-Task Networks*.

Network	Top-1	Top-5
$MTN\,(L, f, S_3)$	92.0 %	—
$MTN\,(L, f, S_{16})$	71.7 %	97.5 %
$MTN\,(L, f, S_{365})$	46.0 %	76.5 %

photo is predicted with a higher accuracy using the coarse and middle partitioning compared to the finest representation. But, the capabilities of the network in terms of spatial resolution are not fully exploited using coarser partitionings. The hierarchical information, however, leads to a more accurate prediction at the finest scale and consequently to a better estimation of the photo's GPS position. Referring to the supplemental material and the next section, it is worth mentioning that the *ISNs* greatly benefit from the knowledge at multiple spatial resolutions. The results on both datasets improve drastically while using the multi-partitioning approach.

4.2 Evaluating the Individual Scene Networks

We apply the scene classifier introduced in Section 3.2 to extract the scene labels for all test images to evaluate the results for specific environmental settings. The resulting number of images for every scene is presented in Table 4. Due to the low number of images in the *Im2GPS* test dataset, we analyze the performance of the *ISNs* on the *Im2GPS3k* dataset. However, referring to Table 6 and the supplemental material, similar observations can be made for *Im2GPS*. The geolocation results do not improve when restricting a single-partitioning network to specific concepts (Figure 5).On the other hand, using a multi-partitioning approach with scene restrictions noticeably improves the geolocation estimation, in particular for urban and indoor photos. One possible explanation is that the intra-class variation for coarser subdivision with more images in larger areas

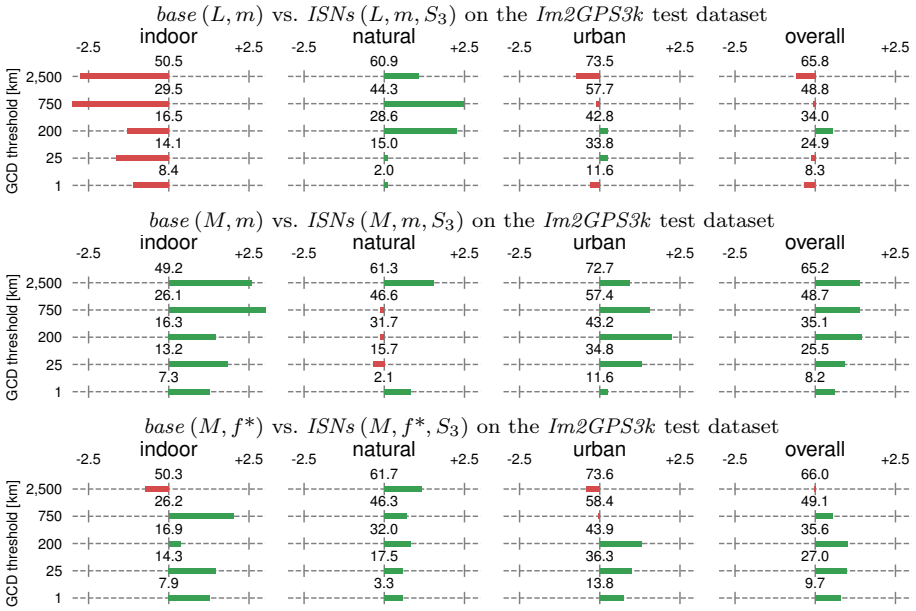

Fig. 5. Comparison of the *Individual Scene Networks* to the baseline approaches for different scene concepts. First mentioned approach is used as reference and its accuracy is denoted in the middle of the x-axis.

is reduced. Therefore, the network is able to learn specific features for the respective scene concept. The best results are achieved for urban images, which is intuitive since they often contain relevant cues for geolocation. It is also not surprising that the performance of indoor photos is the lowest among all scene concepts, since the images can be ambiguous. Weyand et al. (PlaNet) [42]) even consider indoor images as noise. Despite only 1.42M natural images are available to cover the huge diversity of very different scenes like beaches, mountains, and glaciers, we were able to improve the performance for this concept. We believe that the respective *ISN* mainly benefits from the hierarchical information, because it enables the encoding of more global features such as different climatic zones. Overall, the results show that geolocation estimation benefits from training with specific scene concepts and improves at nearly all GCD thresholds for every scene category.

4.3 Evaluating the Multi-Task Network

We investigate the performance of the *Multi-Task Network* regarding the geolocation estimation (Figure 6) and scene classification (Table 5). Despite the results demonstrate that the CNNs are able to learn both tasks simultaneously, geolocalization unfortunately does not benefit from learning an additional task no matter which model we analyze. This underlines that the more important fact

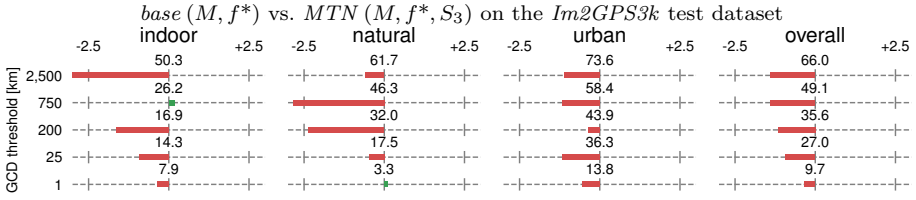

Fig. 6. Comparison of the *Multi-Task Network* to the baseline approach for different scene concepts S. First mentioned approach is used as reference and its accuracy is denoted in the middle of the x-axis.

for predicting the GPS coordinates of photos is to reduce the diversity in the underlying data space. Regarding scene classification, similar results compared to the provided model of the Places2 dataset (Table 2) are achieved.

4.4 Comparison to the State of the Art

We can directly compare the results of our system *base* (L, m) to $[L]\,7011C$ network from Im2GPS [39] and PlaNet (6.2M) [42], since they have a similar number of training images and geographical classes. In addition, PlaNet (91M) [42] can be considered as equivalent at larger scale. The multi-partitioning approach *base* (M, m) is comparable to $[M]\,7011C$ of Im2GPS [39]. The corresponding results on the *Im2GPS* and *Im2GPS3k* test datasets are presented in Table 6. It is obvious that our proposed approaches significantly outperform the current state of the art methods. Interestingly, already our baseline approach *base* (L, m) noticeably outperforms its equivalents. For this reason, we investigate the influence of the utilized ResNet architecture [16]. Therefore, we train the system *base* (L, m) with VGG16 network [36] used in the *Im2GPS* approach [39]. The result is denoted with *base-vgg* (L, m) and shows that the main improvement is explained by the more powerful ResNet architecture. The system *base-vgg*$_c$ (L, m) uses the geographical center of the predicted cell as location (like in *PlaNet* and *Im2GPS*) instead of the mean GPS coordinate of all images that we suggested in Section 3.4. This already noticeably improves the performance on street and city level. Compared to Weyand et al. [42] we have used a less noisy training dataset. As described in the previous sections, the geolocalization can be further increased by training the CNN with multiple partitionings and exploiting the hierarchical knowledge at all spatial resolutions. However, the best results are achieved when the *ISNs* are combined with the hierarchical approach that is trained with images of a specific visual scene concept.

5 Conclusions

In this paper, we have presented several deep learning approaches for planet-scale photo geolocation estimation. For this purpose, scene information was exploited

Table 6. Results on the *Im2GPS* (top) and *Im2GPS3k* (bottom) test sets. Percentage is the fraction of images localized within the given radius using the GCD distance.

Method	**Street 1 km**	**City 25 km**	**Region 200 km**	**Country 750 km**	**Continent 2,500 km**
Human [39]			3.8 %	13.9 %	39.3 %
Im2GPS [39]					
• $[L]$ 7011C	6.8 %	21.9 %	34.6 %	49.4 %	63.7 %
• $[L]$ kNN, $\sigma = 4$	12.2 %	33.3 %	44.3 %	57.4 %	71.3 %
• ... 28m database	14.4 %	33.3 %	47.7 %	61.6 %	73.4 %
PlaNet (6.2M) [42]	6.3 %	18.1 %	30.0 %	45.6 %	65.8 %
PlaNet (91M) [42]	8.4 %	24.5 %	37.6 %	53.6 %	71.3 %
base-vgg$_c$ (L, m)	7.6 %	22.8 %	35.0 %	50.6 %	66.7 %
base-vgg (L, m)	8.9 %	26.6 %	36.7 %	50.6 %	65.8 %
base (L, m)	13.5 %	36.3 %	50.6 %	64.1 %	79.7 %
base (M, m)	13.5 %	35.0 %	49.8 %	64.1 %	79.7 %
base (M, f^*)	15.2 %	40.9 %	51.5 %	65.4 %	78.5 %
ISNs (M, f^*, S_3)	**16.9 %**	**43.0 %**	**51.9 %**	**66.7 %**	**80.2 %**

Method	**Street 1 km**	**City 25 km**	**Region 200 km**	**Country 750 km**	**Continent 2,500 km**
Im2GPS [39]					
• $[L]$ 7011C	4.0 %	14.8 %	21.4 %	32.6 %	52.4 %
• $[M]$ 7011C	3.7 %	14.2 %	21.3 %	33.5 %	52.7 %
• kNN, $\sigma = 4$	7.2 %	19.4 %	26.9 %	38.9 %	55.9 %
base-vgg$_c$ (L, m)	4.2 %	14.6 %	22.2 %	34.4 %	54.2 %
base-vgg (L, m)	4.8 %	16.5 %	22.6 %	34.5 %	54.4 %
base (L, m)	8.3 %	24.9 %	34.0 %	48.8 %	65.8 %
base (M, m)	8.2 %	25.5 %	35.1 %	48.7 %	65.2 %
base (M, f^*)	9.7 %	27.0 %	35.6 %	49.2 %	**66.0 %**
ISNs (M, f^*, S_3)	**10.5 %**	**28.0 %**	**36.6 %**	**49.7 %**	**66.0 %**

to incorporate context about the environmental setting in the convolutional neural network model. We have integrated the extracted knowledge in a classification approach by subdividing the earth into geographical cells. Furthermore, a multi-partitioning approach was leveraged that combines the hierarchical information at different scales. Experimental results on two benchmarks have demonstrated that our framework improves the state of the art in estimating the GPS coordinates of photos. We have shown that the convolutional neural network is enabled to learn specific features for the different environmental settings and spatial resolutions, yielding a more discriminative classifier for geolocalization. Best results were achieved when the hierarchical approach was combined with scene classification. In contrast to previous work, the proposed framework does neither rely on an exemplary dataset for image retrieval nor on a training dataset that consists of several tens of millions images. In the future, we intend to investigate how other contextual information like specific objects, image styles, daytimes and seasons can be exploited to improve geolocalization.

Acknowledgement

This work is financially supported by the German Research Foundation (DFG: Deutsche Forschungsgemeinschaft, project number: EW 134/4-1).

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2016)
2. Altwaijry, H., Trulls, E., Hays, J., Fua, P., Belongie, S.: Learning to match aerial images with deep attentive architectures. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3539–3547. IEEE (2016)
3. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 5297–5307. IEEE (2016)
4. Avrithis, Y., Kalantidis, Y., Tolias, G., Spyrou, E.: Retrieving landmark and non-landmark images from community photo collections. In: International Conference on Multimedia. pp. 153–162. ACM (2010)
5. Baatz, G., Saurer, O., Köser, K., Pollefeys, M.: Large scale visual geo-localization of images in mountainous terrain. In: European Conference on Computer Vision. pp. 517–530. Springer (2012)
6. Bansal, M., Daniilidis, K., Sawhney, H.: Ultrawide baseline facade matching for geo-localization. In: Large-Scale Visual Geo-Localization, pp. 77–98. Springer (2016)
7. Bingel, J., Søgaard, A.: Identifying beneficial task relations for multi-task learning in deep neural networks. arXiv preprint arXiv:1702.08303 (2017)
8. Brejcha, J., Čadík, M.: State-of-the-art in visual geo-localization. Pattern Analysis and Applications **20**(3), 613–637 (2017)
9. Cao, L., Smith, J.R., Wen, Z., Yin, Z., Jin, X., Han, J.: Bluefinder: estimate where a beach photo was taken. In: International Conference on World Wide Web. pp. 469–470. ACM (2012)
10. Chen, D.M., Baatz, G., Köser, K., Tsai, S.S., Vedantham, R., Pylvänäinen, T., Roimela, K., Chen, X., Bach, J., Pollefeys, M., et al.: City-scale landmark identification on mobile devices. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 737–744. IEEE (2011)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. IEEE (2009)
12. Gordo, A., Almazán, J., Revaud, J., Larlus, D.: Deep image retrieval: Learning global representations for image search. In: European Conference on Computer Vision. pp. 241–257. Springer (2016)
13. Hays, J., Efros, A.A.: Im2gps: estimating geographic information from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008)
14. Hays, J., Efros, A.A.: Large-scale image geolocalization. In: Multimodal Location Estimation of Videos and Images, pp. 41–62. Springer (2015)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778. IEEE (2016)
16. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European Conference on Computer Vision. pp. 630–645. Springer (2016)
17. Jaderberg, M., Mnih, V., Czarnecki, W.M., Schaul, T., Leibo, J.Z., Silver, D., Kavukcuoglu, K.: Reinforcement learning with unsupervised auxiliary tasks. arXiv preprint arXiv:1611.05397 (2016)

18. Jin Kim, H., Dunn, E., Frahm, J.M.: Predicting good features for image geo-localization using per-bundle vlad. In: IEEE International Conference on Computer Vision. pp. 1170–1178. IEEE (2015)
19. Kendall, A., Grimes, M., Cipolla, R.: Posenet: A convolutional network for real-time 6-dof camera relocalization. In: IEEE International Conference on Computer Vision. pp. 2938–2946. IEEE (2015)
20. Kim, H.J., Dunn, E., Frahm, J.M.: Learned contextual feature reweighting for image geo-localization. In: IEEE International Conference on Computer Vision. pp. 2136–2145. IEEE (2017)
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1097–1105. NIPS (2012)
22. Larson, M., Soleymani, M., Gravier, G., Ionescu, B., Jones, G.J.: The benchmarking initiative for multimedia evaluation: Mediaeval 2016. IEEE MultiMedia **24**(1), 93–96 (2017)
23. Li, Y., Crandall, D.J., Huttenlocher, D.P.: Landmark classification in large-scale image collections. In: International Conference on Computer Vision. pp. 1957–1964. IEEE (2009)
24. Li, Y., Snavely, N., Huttenlocher, D.P., Fua, P.: Worldwide pose estimation using 3d point clouds. In: Large-Scale Visual Geo-Localization, pp. 147–163. Springer (2016)
25. Lin, T.Y., Belongie, S., Hays, J.: Cross-view image geolocalization. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 891–898. IEEE (2013)
26. Lin, T.Y., Cui, Y., Belongie, S., Hays, J.: Learning deep representations for ground-to-aerial geolocalization. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 5007–5015. IEEE (2015)
27. Liu, L., Li, H., Dai, Y.: Efficient global 2d-3d matching for camera localization in a large-scale 3d map. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2391–2400. IEEE (2017)
28. Quack, T., Leibe, B., Van Gool, L.: World-scale mining of objects and events from community photo collections. In: International Conference on Content-based Image and Video Retrieval. pp. 47–56. ACM (2008)
29. Radenović, F., Tolias, G., Chum, O.: Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In: European Conference on Computer Vision. pp. 3–20. Springer (2016)
30. Ramalingam, S., Bouaziz, S., Sturm, P., Brand, M.: Skyline2gps: Localization in urban canyons using omni-skylines. In: International Conference on Intelligent Robots and Systems. pp. 3816–3823. IEEE (2010)
31. Redmon, J., Farhadi, A.: Yolo9000: Better, faster, stronger. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 6517–6525. IEEE (2017)
32. Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017)
33. Saurer, O., Baatz, G., Köser, K., Pollefeys, M., et al.: Image based geo-localization in the alps. International Journal of Computer Vision **116**(3), 213–225 (2016)
34. Schindler, G., Brown, M., Szeliski, R.: City-scale location recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–7. IEEE (2007)
35. Shan, Q., Wu, C., Curless, B., Furukawa, Y., Hernandez, C., Seitz, S.M.: Accurate geo-registration by ground-to-aerial image matching. In: International Conference on 3D Vision. vol. 1, pp. 525–532. IEEE (2014)
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

37. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016)
38. Tzeng, E., Zhai, A., Clements, M., Townshend, R., Zakhor, A.: User-driven geolocation of untagged desert imagery using digital elevation models. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 237–244. IEEE (2013)
39. Vo, N., Jacobs, N., Hays, J.: Revisiting im2gps in the deep learning era. arXiv preprint arXiv:1705.04838 (2017)
40. Vo, N.N., Hays, J.: Localizing and orienting street views using overhead imagery. In: European Conference on Computer Vision. pp. 494–509. Springer (2016)
41. Wang, Y., Cao, L.: Discovering latent clusters from geotagged beach images. In: International Conference on Advances in Multimedia Modeling. pp. 133–142. Springer (2013)
42. Weyand, T., Kostrikov, I., Philbin, J.: Planet-photo geolocation with convolutional neural networks. In: European Conference on Computer Vision. pp. 37–55. Springer (2016)
43. Workman, S., Souvenir, R., Jacobs, N.: Wide-area image geolocalization with aerial reference imagery. In: IEEE International Conference on Computer Vision. pp. 3961–3969. IEEE (2015)
44. Zamir, A.R., Shah, M.: Accurate image localization based on google maps street view. In: European Conference on Computer Vision. pp. 255–268. Springer (2010)
45. Zamir, A.R., Shah, M.: Image geo-localization based on multiple nearest neighbor feature matching using generalized graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(8), 1546–1558 (2014)
46. Zemene, E., Tariku, Y., Idrees, H., Prati, A., Pelillo, M., Shah, M.: Large-scale image geo-localization using dominant sets. arXiv preprint arXiv:1702.01238 (2017)
47. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Learning deep representation for face alignment with auxiliary attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(5), 918–930 (2016)
48. Zheng, Y.T., Zhao, M., Song, Y., Adam, H., Buddemeier, U., Bissacco, A., Brucher, F., Chua, T.S., Neven, H.: Tour the world: building a web-scale landmark recognition engine. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1085–1092. IEEE (2009)
49. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)