

# Interpretable Semantic Photo Geolocalization

Jonas Theiner<sup>1</sup> Eric Müller-Budack<sup>2</sup> Ralph Ewerth<sup>1,2</sup> <sup>1</sup>L3S Research Center, Leibniz University Hannover, Hannover, Germany<sup>2</sup>Leibniz Information Centre for Science and Technology (TIB), Hannover, Germany

theiner@l3s.de

{eric.mueller, ralph.ewerth}@tib.eu

## Abstract

*Planet-scale photo geolocalization is the complex task of estimating the location depicted in an image solely based on its visual content. Due to the success of convolutional neural networks (CNNs), current approaches achieve super-human performance. However, previous work has exclusively focused on optimizing geolocalization accuracy. Moreover, due to the black-box property of deep learning systems, their predictions are difficult to validate for humans. State-of-the-art methods treat the task as a classification problem, where the choice of the classes, that is the partitioning of the world map, is the key for success. In this paper, we present two contributions in order to improve the interpretability of a geolocalization model: (1) We propose a novel, semantic partitioning method which intuitively leads to an improved understanding of the predictions, while at the same time state-of-the-art results are achieved for geolocation accuracy on benchmark test sets; (2) We introduce a novel metric to assess the importance of semantic visual concepts for a certain prediction to provide additional interpretable information, which allows for a large-scale analysis of already trained models.*

## 1. Introduction

Image geolocalization is the challenging task of predicting the location of a photo in form of a GPS coordinate only by its visual features. Almost all approaches for image geolocalization rely on Convolutional Neural Networks (CNNs) [15, 20, 22], which have (also) achieved near-human or super-human performance in computer vision tasks such as object recognition [39], object detection [25], or semantic image segmentation [8]. Usually, these networks are trained end-to-end with a huge amount of labeled data. This strategy and the typically large size of the networks turn them into a kind of black-box-systems, whose reasoning and predictions are not comprehensible – making it necessary to develop methods to understand

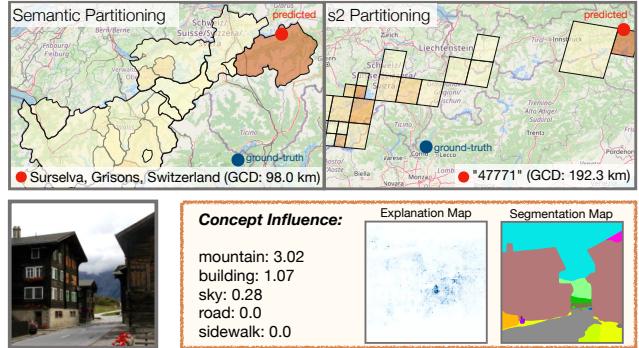


Figure 1. Example output of the proposed semantic partitioning for more explainable geolocation estimation (left) compared to an existing partitioning approach (right) and with state-of-the-art performance. Additionally, we measure the influence of visible concepts on the prediction of a model (*concept influence*).

their decisions [10, 12]. This need also exists for CNN-based geolocalization approaches, which have focused exclusively on maximizing localization accuracy, but lack interpretability and explainability. This is particularly a requirement for geolocalizations systems since humans are far worse at estimating locations than current deep learning approaches [55].

While many approaches restrict the problem of photo geolocalization to a part of the earth (e.g., landmarks or mountains) [4, 5, 24, 31, 41, 42, 43, 53, 57], predicting a coordinate at planet-scale without any restrictions is more complex. Landmarks (usually tourist attractions) can partly be verified by humans, whereas many photos give little indication of the actual place or region. Since the performance of all non-restricted approaches [13, 18, 30, 45, 54, 55] exceeds by far that of an human, the question arises which features have been learned and which image features are relevant for a certain prediction. State-of-the-art approaches define the task at planet-scale as a classification problem, where the earth is divided into many small cells heavily based on the dataset distribution [30, 45, 55] (called *par-*

*titioning*). The performance of these models is compared by measuring the classification accuracy at a certain error level (e.g., at city-level with a prediction error of less than 25 km). Despite their impressive results, the geolocation error is not perfect which implies that geolocalization systems should be treated with careful consideration with respect to trustworthiness. Furthermore, existing approaches do not assess whether or how well the geolocation can be actually predicted from a given photo. The quadratic boundaries of the  $s2$  partitioning [55] are arbitrary (see Figure 1) and the cells of *CPlaNet* [45] are initialized randomly which is counter-intuitive with respect to comprehensibility. Following these considerations, a CNN-based approach for geolocation estimation should therefore also be assessed with regard to interpretability of its results.

In this paper, we address this issue and introduce a novel semantic partitioning method where the cells are not rectangular or arbitrarily shaped as in previous approaches [30, 45, 55], but rather refer to real, interpretable locations (e.g., streets, districts, cities, or even countries) derived from the Open Street Map (OSM) [34] database (Figure 1). As a result, training and output of a model are more comprehensible to humans by default. Further, we suggest a *concept influence* metric to investigate the post-hoc interpretability by measuring the influence of semantic visual concepts on individual predictions. An example output is visualized in Figure 1. Experimental results on two large-scale public test sets show that our model using semantic partitioning achieves state-of-the-art results.

The rest of the paper is organized as follows. Related work for photo geolocation estimation and explainability of CNNs is reviewed in Section 2. The partitioning method and *concept influence* score are described in Section 3, while experimental results are reported in Section 4, both for accuracy and interpretability of results. Section 5 summarizes the paper and outlines areas of future work.

## 2. Related Work

In Section 2.1, related work for image geolocalization is reviewed. A brief overview for explainability of CNNs and relevant concepts is given in Section 2.2.

### 2.1. Geolocation Estimation of Photos

Whereas only few approaches [18, 30, 45, 54, 55], work without limitations at planet-scale, the majority simplifies the task of geolocalization, for example, by predicting landmarks and cities [4, 31, 41, 57], natural areas [5, 24, 42, 43, 53], or geo-related attributes [23]. Previous work uses either image retrieval approaches or models the task as a classification problem. The task of geolocation estimation at planet-scale has an overlap with methods from instance-level image retrieval [4, 11, 31, 37, 52] where benchmark datasets consist of popular places, land-

marks [19, 35, 36], and tourist attractions which can be verified by humans. Common to all is the usage of triplet ranking or contrastive embeddings to learn discriminative image representations, whereas Liu et al. [26] introduce an alternative loss function. These representations are used to retrieve the most similar images in a reference database in order to determine the geolocation as proposed by *Im2GPS* [13, 14].

Weyand et al. [55] introduce the classification approach *PlaNet*, where a GPS coordinate is mapped to a discrete class label using a quad-tree approach that divides the surface of the earth into distinct regions by utilizing the  $s2$  geometry library [40]. This  $s2$  partitioning is used at multiple spatial scales in one network to exploit hierarchical knowledge [30, 54]. A pre-classification step that assigns a photo to one of three scene types (natural, urban, indoor) also leads to improvements [30]. Seo et al. [45] propose a combinatorial partitioning where the overlaps of multiple coarse-grained partitionings create one fine-grained partitioning. Izbicki et al. [18] introduce the *Mixture of von-Mises Fisher* (MvMF) loss function for the classification layer that exploits the earth’s spherical geometry and refines the geographical cell shapes in the partitioning.

All aforementioned approaches rely on CNNs and their predictions are therefore difficult to interpret.

### 2.2. Explainability of Neural Networks

The size of deep learning architectures and the complexity of their training turn them into black-box systems, making it unclear to humans what information is encoded or on which reasons a decision is based upon [29]. A comprehensive overview of the terminology and developed methods in the area of explainable artificial intelligence (XAI) is provided by Gilpin et al. [10] and Guidotti et al. [12]. Approaches to explainability fall into three categories [12], those that focus on (1) the explanation or interpretation of system output, (2) the representation inside a network, or (3) the creation of “explaining” systems that are capable of providing explanations that are more comprehensible by humans. Within the first category mainly the question has to be answered, why a certain input leads to a particular output (i.e., explanations at instance level). Proxy models are based on the idea to reduce the model’s complexity in order to make their decisions more comprehensible [3, 44, 60]. *Salience mapping* aims to highlight regions in the input that have influence on the network output. A comparison of methods can be found in [2, 12], whereas a detailed analysis is conducted in [1, 7]. To inspect the representation inside the network [56], the role of individual units (neurons, filter, layer) is investigated (e.g., [6, 33, 46, 58]). In general, explanation generating systems are designed to provide any kind of explanation understandable to humans. For instance, one special architecture are attention-based networks that can be utilized for this purpose (e.g., [28, 38]).

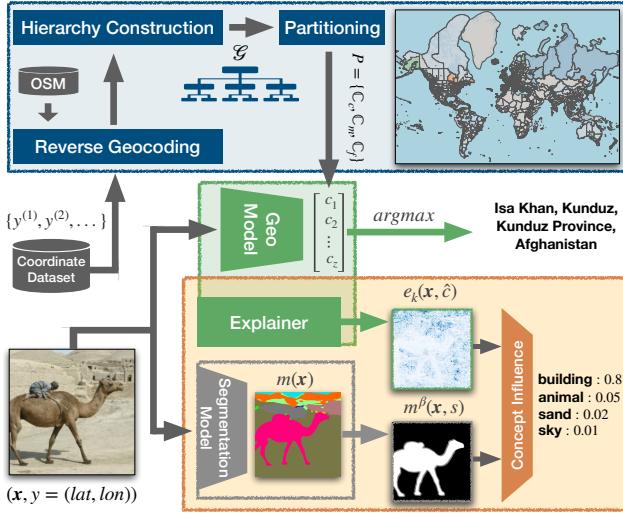


Figure 2. Overview of the overall system: Top (blue): Workflow to create a Semantic Partitioning (*SemP*), bottom (orange): Components to measure the *concept influence* for individual samples.

### 3. Interpretable Semantic Photo Geolocation

The following two subsections address two different issues with regard to interpretability. First, a novel partitioning method is proposed that relies on data that are derived from a geo database where metadata about many regions and places such as their size or exact boundaries is provided. Second, a method is presented to automatically assess image features that are relevant for a model’s decision based on semantic visual concepts like *sky* or *person*. Their workflows and connections are outlined in Figure 2.

#### 3.1. Semantic Partitioning

State-of-the-art methods for photo geolocalization rely on classification approaches [30, 45], where the design of the classification layer is crucial for the model’s output with respect to prediction accuracy, but also regarding the information that is provided to users. The main idea is to divide the earth surface into a discrete set of classes  $\mathbb{C}$  based on the dataset distribution to then train a classification network [55]. We follow a similar idea, but after all our cells cover territorial borders (e.g., countries, cities), natural geological boundaries (e.g., rivers, mountains) or man-made barriers (e.g., roads or railways that separate districts). In addition to an improved understanding of the created cells, the assumption is that a CNN will learn better image representations, since the classes are more distinguishable. The following steps describe formally how that Semantic Partitioning (*SemP*) is constructed (Section 3.1.2), initially starting with two fundamental requirements, the reverse geocoding and the hierarchy construction (Section 3.1.1).

#### 3.1.1 Reverse Geocoding & Hierarchy Construction

Following the idea of classification, a mapping is needed from the continuous GPS coordinate space to a discrete set of existing locations which is called reverse geocoding. Frameworks for reverse geocoding generate an address vector, e.g., (*Long Beach, Los Angeles County, California, USA*) with a coordinate as input. We choose *Nominatim* [32] since it is open source software and relies on OSM. Formally, a reverse geocoder maps each coordinate  $y^{(i)} = (\text{latitude}, \text{longitude})$  in a dataset  $\mathbb{D} = \{y^{(1)}, y^{(2)}, \dots\}$  to an address vector  $\mathbf{l}^{(i)} = (l_1, \dots, l_u)$  of arbitrary length  $u$  and is ordered from fine to coarse. This mapping is denoted as  $\mathbb{D}^l = \{\mathbf{l}^{(1)}, \mathbf{l}^{(2)}, \dots\}$ .

**Hierarchy Construction:** Since hierarchical knowledge is valuable with respect to performance [30] and all necessary information is already provided by the reverse geocoder, we transform the multi-label problem into a single-label problem by constructing a hierarchy similar to the *s2* library but with semantically meaningful nodes and edges. In order to create a partitioning from the obtained addresses  $\mathbb{D}^l$ , it is required to build a hierarchy where each discrete location  $l$  (e.g., *Long Beach*) can be assigned to its next coarser distinct location (e.g., *Los Angeles County*). A directed (multi-) graph  $\mathcal{G} = (V, E)$  can be constructed using all edges that occur in the mapping  $\mathbb{D}^l$ . The total number of nodes corresponds to the number of locations; an edge exists between two adjacent nodes encoded in the mapping  $\mathbb{D}^l$ , i.e.,  $(l_i, l_{i+1}) \forall i \in [1, \dots, u - 1]$  for every  $\mathbf{l} = (l_1, \dots, l_u) \in \mathbb{D}^l$ . Nodes without incoming edges (roots) usually correspond to countries. Ideally,  $\mathcal{G}$  consists only of trees, where exactly one parent node is assigned to each node with the exception of the root nodes. Otherwise,  $\mathcal{G}$  must be transformed into a hierarchy. For each location only the most frequent incoming edge is kept. Reasons for multiple parents are possible incorrect assignments for some instances or missing assignments that cause shortcuts. Therefore, the mapping  $\mathbb{D}^l$  is subsequently replaced with the locations from the shortest path of the finest location  $l_1$  in  $(l_1, \dots, l_u) = \mathbf{l} \quad \forall \mathbf{l} \in \mathbb{D}^l$  to its root node in the hierarchy  $\mathcal{G}$  and referred to as  $\mathbb{D}^{l^*}$ .

#### 3.1.2 Partitioning Construction & Cell Assignment

In order to create the *SemP*, the set of coordinates  $\mathbb{D}$  is first transformed to a (hierarchical) multi-label dataset as described in Section 3.1.1. To consider only the finest location  $l_1^{(i)}$  for each  $\mathbf{l}^{(i)}$  in the transformed dataset  $\mathbb{D}^l$  is a valid partitioning. However, the number of classes is generally not manageable and previous work (e.g., [55]) controls the granularity of a partitioning, i.e., the number of classes. Although this is an essential step it leads to a trade-off prob-

lem. On the one hand, fewer but geographical larger cells are easier to distinguish, but they also reduce the spatial resolution of the outputs and consequently result in less accurate predictions. On the other hand, more but smaller cells provide a good spatial resolution, but are more difficult to distinguish and also decrease the number of training examples per cell, making the model more prone to overfitting. Hence, locations  $l$  that occur sporadically, i.e., whose count is below a threshold  $\tau_{\mathbb{D}}$ , are removed directly as they are assumed to be not relevant. To construct a partitioning  $\mathbb{C}$  at a certain spatial level, the mapping  $\mathbb{D}^{l^*}$  and the derived hierarchy  $\mathcal{G}$  (as explained in Section 3.1.1) is needed. The finest locations in  $\mathbb{D}^{l^*}$  form a temporal (initial) partitioning, i.e., all  $l_1$  from all  $l \in \mathbb{D}^{l^*}$ . To assign a dataset  $\mathbb{D}_{\text{new}}$  to classes from a created partitioning  $\mathbb{C}$ , two steps are necessary. First, the same reverse geocoder has to create an initial assignment  $\mathbb{D}_{\text{new}}^{l^*}$  and these discrete locations are filtered by the available locations (now classes) from the partitioning  $\mathbb{C}$ . Given the  $i$ -th sample  $l^{(i)} \in \mathbb{D}_{\text{new}}^{l^*}$ , the location  $l_0^{(i)}$  from  $l^{(i)}$  corresponds to the finest available one according to the partitioning  $\mathbb{C}$ .

### 3.1.3 Learning & Inference

With the classes  $\mathbb{C}$  obtained from the presented partitioning method, a CNN can be trained directly on the classification task using the cross-entropy loss ( $\mathcal{L}_{\text{CE}}$ ) where the number of classes corresponds to the number of cells of the partitioning. Initially, only a dataset consisting of image-coordinate pairs is necessary where the coordinates are transformed to classes according to Section 3.1.2. As in [30], multiple partitionings can be combined to force the model to learn some kind of hierarchical knowledge. Given a tuple of partitionings  $\mathcal{P} = (\mathbb{C}_1, \dots, \mathbb{C}_n)$  which differ only in  $\tau_{\min}$  and are ordered from fine to coarse, each cell in  $\mathbb{C}_i$  can be assigned to its corresponding cell in  $\mathbb{C}_{i+1}$  by exploiting the hierarchy  $\mathcal{G}$ . One fully-connected layer per partitioning  $\mathbb{C}_i$  is added on top of an appropriate CNN architecture. During training, the multi-partitioning classification loss is defined as the sum of all individual losses per partitioning  $\mathcal{L}_{\text{CE}}^{\text{multi}} = \sum_{i=1}^{|\mathcal{P}|} \mathcal{L}_{\text{CE}}^i$ . During inference, the class at the finest partitioning with the highest probability after applying softmax function corresponds to the predicted cell  $\hat{c}$ . To predict a GPS coordinate, the assigned coordinate for the respective class is taken, selected by the average positions of the assigned samples from  $\mathbb{D}$  during the partitioning process.

## 3.2 Measuring the Input Feature Importance

In the task of photo geolocalization, we do not know which image regions are crucial for the model’s prediction and cannot validate the decisions. While methods for visualizing feature attributions have been researched in recent years, the main focus was on object recognition where the

highlighted areas are comprehensible at least to humans [2]. Inspired by these approaches, we propose a method to measure the influence of specific objects (e.g., *vehicle* or *person*) and semantic image regions (e.g., *sky* or *ground*) regarding the model’s prediction. The goal is not to identify a concrete concept that is responsible for the prediction, which would be counter intuitive, since a decision should not be reduced to image regions exclusively. Rather it can be helpful to estimate the overall impact of a given semantic concept, to identify misleading concepts, or to provide explanatory information in form of a more comprehensible (text and quantitative values) and summarized (reduced to relevant concepts) explanation map to users. Attribution maps also provide an importance value per pixel, but also a lot of noise [47] and allow the observer a lot of freedom for interpretation.

**Required Components:** For the formal setup an input image  $x \in \mathbb{R}^{w \times h \times d}$  and a trained CNN for geolocation estimation  $\Psi_g$  are required. Furthermore, only two components are needed to calculate the influence of concepts on the prediction. An *explanation map*  $e$  assigns an importance score to each input pixel of  $x$  for a certain prediction [46, 48, 50], e.g., a target class  $\hat{c}$  in case of a classification model. The maximum over the (color) channel dimension  $d$  is taken as only image regions are of interest, hence we define  $e : \mathbb{R}^{w \times h \times d} \mapsto \mathbb{R}^{w \times h}$ . Note, that usually only the gradients of the model  $\Psi_g$  have to be accessible for the calculation. A *segmentation map*  $m$  divides image areas into semantic groups (e.g., a region, object, or texture). The segmentation mask for one concept  $s \in \mathbb{S}$  is the indicator function where the presence of  $s$  on a pixel is indicated by  $m : \mathbb{R}^{w \times h \times d} \mapsto \{0, 1\}^{w \times h}$  and denoted as  $m(x, s)$ .

Assuming that the area of the segmentation boundaries, i.e., the border between two concepts is of interest for the geo model resulting in activations in the explanation map, the active area of the binary mask  $m(x, s)$  for concept  $s$  can be enlarged using a morphological dilation, i.e.,  $\beta$  pixels around its surface and is denoted as  $m^\beta(x, s)$  as seen in Figure 2 (gray colorized area around the camel’s surface).

**Concept Influence:** The aim is to measure the influence of a specific concept  $s$  using the explanation map  $e(x, \hat{c})$  and the segmentation map  $m(x, s)$  for a specific concept  $s$ . As stated by Ghorbani et al. [9], in many settings, only the most important features are of explanatory interest. They compute the pixel-wise intersection of the  $k$  most important features from  $e(x, \hat{c})$  to measure the difference between two explanation maps (top- $k$  intersection). However, this is done for a slightly different purpose, that is generating and evaluating manipulated explanation maps. Inspired by this, we adapt this measure to define the influence of a concept  $s$  visible in image  $x$  with respect to a geoprediction  $\hat{c}$  of the

model. We define the pixel-wise intersection  $tki$  between the binary segmentation mask  $m(\mathbf{x}, s)$  and the binary mask of the top- $k$  features  $e_k(\mathbf{x}, \hat{c})$  as

$$tki = \frac{1}{k} \sum_{i=1}^w \sum_{j=1}^h m(\mathbf{x}, s)_{i,j} \wedge e_k(\mathbf{x}, \hat{c})_{i,j} \quad (1)$$

where  $m(\mathbf{x}, s)$  and  $e_k(\mathbf{x}, \hat{c})$  are both in  $\{0, 1\}^{w \times h}$  and  $\wedge$  is the pixel-wise boolean *and* operation. For instance, if all top- $k$  pixels are within the shape of the concept  $s$  then  $tki = 1$ . In our experiments we set the parameter  $k$  to 1,000 as proposed by Ghorbani et al. [9]. As large objects or areas are preferred, a normalization step is crucial for application. The defined top- $k$  intersection ( $tki$ ) is hence normalized by the relative size of the concept that is defined as:

$$\bar{s} := \frac{1}{wh} \sum_{i=1}^w \sum_{j=1}^h (m(\mathbf{x}, s))_{i,j} \quad (2)$$

The resulting score is the definition of the *concept influence* ( $ci$ ) metric:

$$ci(m(\mathbf{x}, s), e_k(\mathbf{x}, \hat{c})) = \frac{tki}{\bar{s}} \quad (3)$$

A  $ci$  score of less than or equal to one means that the top- $k$  pixels of the explanation map are more likely to be in other regions of the image, i.e. class  $s$  has little or no influence on the final prediction. The  $ci$  score indicates whether a concept  $s$  contains relatively large number of activations of the explainer  $e$ . When fixing the minimum required relative concept size to  $0 < \bar{s}_{\min} < 1$ ,  $ci \in [0, \frac{1}{\bar{s}_{\min}}]$  is well defined and only those concepts are considered for the calculation that cover at least this area. Additionally, we assume that smallest concepts that cover only a minimal area in the image assumed to be irrelevant or noisy and set  $\bar{s}_{\min} = 0.1$  in our experiments.

Finally, given a segmentation map  $m$  and an explanation map  $e$  for model  $\Psi_g$ , we can automatically determine and evaluate which image regions with respect to all visible concepts in an image are relevant for a geo prediction with the introduced metric  $ci$ .

## 4. Experimental Results

In this section, the proposed partitioning method is evaluated with respect to geolocation accuracy and its capability of providing an improved interpretability (Section 4.1). Afterwards, the *concept influence* metric introduced in Section 3.2 is evaluated (Section 4.2).

### 4.1. Semantic Partitioning

We demonstrate the capability of our approach through a comparison with a model that also exploits the hierarchical knowledge from multiple partitionings [30].

**Datasets:** We utilize the MediaEval Placing Task 2016 (*MP-16*) [21] dataset that is a subset from the Yahoo Flickr Creative Commons 100 Million (*YFCC100M*) [51] both for partitioning construction and training. Its only restriction is that an image contains a GPS coordinate, thus it contains images of landmarks, landscape images, but also images with little to no geographical cues. Like Vo et al. [54] images are excluded from the same authors as in the *Im2GPS3k* test set and duplicates are removed resulting in a dataset size of 4,723,695 image-coordinate pairs. For validation, a random sampled subset comprising 25,600 images from *YFCC100M* without overlap to the training images is created and denoted as *YFCC-Val26k*. For testing, we focus on three popular benchmark datasets (*Im2GPS* [13], *Im2GPS3k* [54] and *YFCC4K* [54]). *YFCC4K* comes from the same image domain as the training dataset but is designed for general computer vision tasks making the test set more challenging. In contrast, *Im2GPS* and *Im2GPS3k* contain some landmarks, but the majority of images is recognizable only in a generic sense like landscapes.

**Partitioning Parameters:** First, the coordinates from the *MP-16* are transformed to a multi-label dataset containing 2,191,616 unique locations. To initially reduce the number, we apply a filtering step with  $\tau_D = 50$  resulting in manageable 46,240 unique locations. Due to the proven importance of a multi-partitioning, we directly evaluate this setting. For a fair comparison we construct a multi-partitioning that consists of three individual partitionings (coarse, middle, fine) with a similar total number of unique classes compared to Müller-Budack et al. [30] and follow their notation. For the partitioning construction several thresholds  $\tau_{\min}$  can be applied to get a similar number of classes as shown in Table 1. For this reason, we select the model that performs best on the validation set for the comparison and evaluation on the test sets. Furthermore, we investigate three additional settings: (1) To keep the parameters fixed, but applying one filter, i.e., utilization of locations with an area (polygon or multi-polygon) only, denoted as *SemP<sub>a</sub>*; (2) testing the hierarchical prediction variant ( $f$  vs.  $f^*$ ); and (3) testing the scalability by doubling the number of classes.

**Evaluation Metric:** For evaluation it is common to calculate the geolocation accuracy at multiple error levels, i.e., the tolerable error in terms of distance to the ground-truth location is varied. Formally, the geolocation accuracy  $a_r$  at scale  $r$  (in km) is defined as follows for a set of  $N$  samples:

$$a_r \equiv \frac{1}{N} \sum_{i=1}^N u \left( d(l_{\text{gt}}^{(i)}, l_{\text{pred}}^{(i)}) < r \right), \quad (4)$$

where the distance function is the Great Circle Distance (*GCD*) and  $u(\cdot)$  is the indicator function whether the dis-

Configuration	$ C $	$a_r [\%]$ @ km				
		1	25	200	750	2500
$SemP(\{100, 125, 150\}, f^*)$	14877	4.8	11.0	18.5	33.6	53.9
$SemP(\{100, 125, 150\}, f)$	14877	7.5	15.8	23.8	38.0	56.6
$SemP_a(\{100, 150, 250\}, f)$	12886	6.2	16.1	24.4	38.0	55.3
$SemP(\{100, 150, 250\}, f)$	15127	6.6	16.4	24.0	37.6	55.4
$SemP(\{100, 125, 250\}, f)$	15016	7.5	15.9	24.1	38.3	56.6
$SemP_a(\{75, 100, 150\}, f)$	16808	6.6	16.4	24.0	37.6	55.4
$SemP(\{50, 75, 100\}, f)$	34049	8.9	16.6	24.1	37.9	56.3
$s_2(M, f^*)$	15606	6.8	16.4	24.6	38.4	56.8

Table 1. Experimental results on the validation set of *YFCC-Val26k* for several multi-partitonings where  $|C|$  is the total number of unique classes.

tance is smaller than the tolerated radius  $r$ . The suggested values for  $r$  according to [54, 55] are used.

**Network Training & Inference:** We choose the commonly applied ResNet50 [15, 16] As our network architecture since it provides a good trade-off in terms of performance and training time. The classification layers are added on top of the global pooling layer. Instead of initializing the parameters of all models with *ImageNet* weights, the weights from a model trained for ten epochs on countries extracted from the *MP-16* is taken to derive features related to the problem of geolocation estimation (as described in Appendix A). Stochastic Gradient Descent (SGD) with an initial learning rate of 0.01, a momentum of 0.9, and weight decay of  $10^{-4}$  is used to optimize the weights for 15 epochs. The learning rate is exponentially decreased by a factor of 0.5, initially after every three epochs and later from epoch 12 on every epoch. Training is performed with a batch size of 200 and validation is done after 512,000 images. During training, images are augmented as follows to prevent overfitting: Images are initially resized to a minimum edge size of 320 pixels. After that, we crop a region of random size (between 0.66 and 1.0 of original size) and random aspect ratio (3/4 to 4/3) of the original aspect ratio. This crop is finally resized to  $224 \times 224$  pixels and randomly flipped horizontally. For validation, the image is resized to a minimum edge size of 256 pixels and then a center crop of size 224 pixels is made. The model on the validation set with lowest loss  $\mathcal{L}_{CE}^{\text{multi}}$  is selected. For testing, five square crops are made and the mean prediction after applying softmax is taken.

**Results on the Validation Set:** The geolocation accuracies on the *YFCC-Val26k* validation set are reported in Table 1. Surprisingly, the hierarchical prediction ( $f^*$ ) [30] is, in contrast to the assumptions, even worse than considering only the finest partitioning ( $f$ ). One technical reason might be the fundamental different underlying structure of the hi-

Approach		$a_r [\%]$ @ km				
		1	25	200	750	2500
Im2GPS3k [54] (2,997 images)						
[L]7011C [54]		4.0	14.8	21.4	32.6	52.4
[L]kNN, $\sigma = 4$ [54]		7.2	19.4	26.9	38.9	55.9
PlaNet [55] repro. in [45]		8.5	24.8	34.3	48.4	64.6
CPlaNet[1-5] [45]		9.9	26.3	34.5	48.8	64.6
CPlaNet[1-5, PlaNet] [45]		10.2	26.5	34.6	48.6	64.6
$s_2(M, f^*)$ [30]		9.7	27.0	35.6	49.2	66.0
$ISN(M, f^*, S_3)$ [30]		10.5	28.0	36.6	49.7	66.0
$s_2(M, f^*)$ (repro.)		10.0	27.0	36.5	<b>50.9</b>	<b>67.2</b>
$SemP(\{100, 125, 250\}, f)$		11.1	<b>27.1</b>	<b>36.7</b>	50.4	66.1
$SemP_a(\{75, 100, 150\}, f)$		10.0	27.0	37.2	50.1	65.0
$SemP(\{50, 75, 100\}, f)$		<b>11.5</b>	27.0	36.3	49.3	65.9
YFCC4k [54] (4,536 images)						
[L]kNN, $\sigma = 4$ [54]		2.3	5.7	11.0	23.5	42.0
PlaNet [55] repro. in [45]		5.6	14.3	22.2	36.4	55.8
CPlaNet[1-5] [45]		7.3	14.7	21.7	36.2	<b>55.6</b>
CPlaNet[1-5, PlaNet] [45]		7.9	14.8	21.9	36.4	55.5
$s_2(M, f^*)$ (repro.)		6.6	16.4	<b>24.1</b>	36.8	55.1
$SemP(\{100, 125, 250\}, f)$		7.3	15.3	23.9	<b>37.2</b>	54.3
$SemP_a(\{75, 100, 150\}, f)$		6.6	15.9	23.6	36.6	52.7
$SemP(\{50, 75, 100\}, f)$		<b>9.3</b>	<b>17.1</b>	<b>24.1</b>	36.9	54.3
Im2GPS [13] (237 images)						
Human [54]		-	-	3.8	13.9	39.3
Im2GPS [13]		-	12.0	15.0	23.0	47.0
[L]kNN, $\sigma = 4, 28M$ [54]		14.4	33.3	47.7	61.6	73.4
[L]7011C [54]		6.8	21.9	34.6	49.4	63.7
PlaNet [55]		8.4	24.5	37.6	53.6	71.3
CPlaNet[1-5] [45]		16.0	36.7	46.4	62.4	78.5
CPlaNet[1-5, PlaNet] [45]		<b>16.5</b>	37.1	46.4	62.0	78.5
$s_2(M, f^*)$ [30]		15.2	40.9	51.5	65.4	78.5
$ISN(M, f^*, S_3)$ [30]		16.9	43.0	51.9	66.7	80.2
MvMF ( $c = 2^{17}$ ) [10]		8.4	32.6	39.4	57.2	80.2
$s_2(M, f^*)$ (repro.)		15.2	<b>40.9</b>	<b>51.9</b>	<b>65.8</b>	<b>80.6</b>
$SemP(\{100, 125, 250\}, f)$		15.2	36.7	48.1	64.1	78.1
$SemP_a(\{75, 100, 150\}, f)$		12.2	35.4	47.3	63.3	78.1
$SemP(\{50, 75, 100\}, f)$		16.0	38.4	49.4	63.7	78.5

Table 2. Performance comparison of our *SemP* method to several approaches measuring the geolocation accuracy ( $a_r$ ) on common benchmark datasets.

erarchy  $\mathcal{G}$  in contrast to the quad-tree [55], resulting in a significantly lower depth and more variable number of child nodes. Humans may perceive locations hierarchically, but these coarse regions are not the ones with visually discriminative features. All configurations with similar number of classes perform similarly well which shows that the exact choice of hyperparameters is not essential. From all models, we select the model that have the best geolocation accuracy ( $SemP(\{100, 125, 250\}, f)$ ), particularly for the error levels 1 km and 2,500 km. When fixing the configu-

ration to  $\tau_{\min} \in \{100, 150, 250\}$  but considering only locations where geodata is available, the performance drops at finer scale but remains similar for the other scales. This is caused by the fact that for coarser regions, shape boundaries are more often available in OSM. This filtering is intended to show that, depending on the filters applied, the performance can vary. Doubling the classes improves the accuracy slightly at street level, but converges when further doubling as already observed by Izbicki et al. [18].

**Comparison to State of the Art:** Quantitative results are reported for three test sets in Table 2. The best performing model ( $SemP(\{100, 125, 250\}, f)$ ) and the best performing model containing only locations with geodata on the validation set are selected. In addition, we report the results for the model, where the number of classes is doubled. The reproduced results from  $s2(M, f^*)$  [30] are slightly better than the original results despite the usage of the smaller ResNet50. On all three test sets and at all scales we obtain results similar to the state-of-the-art models [30, 45] for  $SemP(\{100, 125, 250\}, f)$ .

Now, we discuss the findings from some qualitative results in detail with a focus on the interpretability of individual predictions. The primary goal is to develop a partitioning that is intuitively comprehensible to humans but the quantitative results have shown  $SemP$  provides competitive results regarding geolocation accuracy. In Figure 3 (last two rows) four examples from *Im2GPS3k* are visualized where both  $SemP_a(\{75, 100, 150\}, f)$  and  $s2(M, f)$  share the same range of geolocation accuracy. For each partitioning the cells with the top probabilities (max. 25) are colored in the zoomed region of the world map. The predicted label is depicted below of the maps.

Both models are trained on two different types of partitionings and achieve similar geolocation accuracies. However, there are two main advantages of the proposed partitioning method over the  $s2$  method during inference. First, not only a coordinate is provided but also the human-readable class label (e.g., "la Sagrada Familia, Barcelona, Spain") where its level of detail is ordered according to the semantic hierarchy and provided metadata. Second, the visualization on the relevant part of the world map is much more clearly structured (Figure 3 third row), since the boundaries of the cells are not arbitrary selected but rather follow geographical borders which finally leads to a better understanding of a prediction. In line with this, the procedure of constructing smaller or more detailed cells is more natural in semantic partitioning  $SemP$ , since it follows a real hierarchical structure (e.g., from city to district), in contrast to the  $s2$  algorithm with a hierarchy fixed to exactly four finer cells due to its underlying quad-tree.

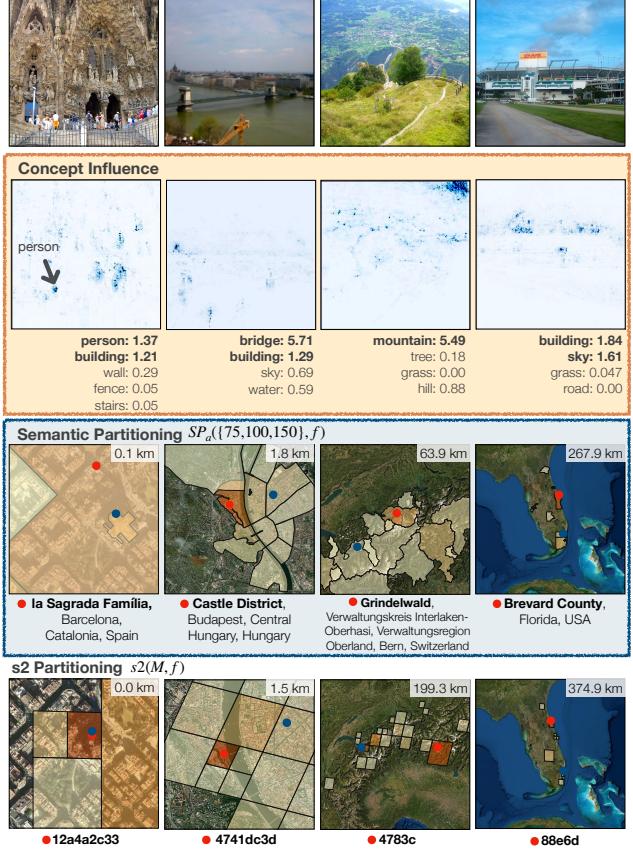


Figure 3. Output of the proposed *concept influence* metric in addition to the explanation map [50] and qualitative comparison of predictions from the  $SemP_a$  model and  $s2(M, f)$  [30] (last two rows). The top-25 classes and probabilities are visualized on the zoomed world map. Below is the predicted class label whereas the red marker is the predicted coordinate (blue is ground-truth).

## 4.2. Understanding the Input Feature Importance

It is assumed that concepts such as *road* or *person* provide little visual cues to predict the actual location, and if these concepts influence the prediction, it may lead to false predictions. Thus, the aim is not only to analyze these concepts but also to identify those that could have a positive or misleading effect on the prediction.

### 4.2.1 Initial Setup

We focus on the reproduced  $s2(M, f^*)$  model in this experiment as it provides the best geolocation performance without the usage of ensemble models. Since current segmentation models achieve high-quality results, we apply the *HRNetV2* [49] which is trained on the *ADE20k* [59] dataset that contains 150 object classes (e.g., *person*, *car*, *bottle*) and concepts for scene parsing (e.g., *sky*, *ground*, *mountain*) and is therefore well suited. According to the results

Concept	GCD $\leq 750 \text{ km}$			GCD $> 750 \text{ km}$		
	Images	$\bar{ci}_{\text{median}}$	$ci_{\mu}$	Images	$\bar{ci}_{\text{median}}$	$ci_{\mu}$
sky	4053	1.22	1.60	4052	1.00	1.45
animal	141	1.22	1.76	458	1.51	1.90
building	3038	1.06	1.35	2521	1.05	1.34
mountain	787	1.03	1.39	672	0.98	1.30
flower	101	1.02	1.29	416	1.18	1.40
person	2802	0.92	1.14	6193	1.14	1.35
wall	2461	0.79	0.91	7107	0.80	0.94
tree	2262	0.78	1.12	2730	0.87	1.19
painting	122	0.71	0.88	354	0.84	1.00
ceiling	362	0.63	1.00	974	0.67	0.96
signboard	237	0.59	1.07	497	0.61	0.97
rock	304	0.55	0.91	681	0.68	1.02
car	286	0.50	0.85	399	0.58	0.83
sea	408	0.29	0.66	495	0.45	0.76
fence	166	0.26	0.56	213	0.31	0.65
boat	102	0.24	0.58	110	0.51	0.86
floor	629	0.24	0.53	1732	0.35	0.59
plant	460	0.23	0.48	1117	0.48	0.66
earth	1146	0.18	0.43	1931	0.30	0.55
water	753	0.18	0.45	1399	0.44	0.65
field	165	0.15	0.39	156	0.19	0.46
grass	839	0.12	0.36	1025	0.19	0.44
sand	144	0.10	0.44	188	0.16	0.39
road	803	0.08	0.32	681	0.10	0.35
sidewalk	340	0.05	0.29	295	0.11	0.45

Table 3. The *concept influence* ( $ci$ ) is summarized for each concept in YFCC-Val26k using median and mean ( $\mu$ ) and separated in correct and incorrect *GCD* prediction. Only concepts are reported with at least 100 samples.

of a benchmark study [17], the method of *Integrated Gradients* [50] is chosen as the basis explanation method, extended by a variant of *SmoothGrad* [47] which seeks to alleviate noise in explanation maps. Inserting random Gaussian noise in  $n$  copies of the input image and then averaging the produced explanation maps cleans up artifacts. The noise parameter is set to  $\sigma = 0.15$  as suggested by the authors, but the sample size is reduced from 20 to 5 due to computing complexity without major result changes. Please note, that the variant as in [17] is used which squares each value before averaging.

**Evaluation Metrics:** To automatically identify concepts that have influence on the prediction, we compute the median ( $ci_{\text{median}}$ ) and mean ( $ci_{\mu}$ ) for the proposed  $ci$  measure on a full dataset individually for each concept  $s \in \mathbb{S}$ . Concepts with geo-information clues should have  $ci$  scores greater or equal to one according to its definition (Section 3.2) and concepts without any hints close to zero. To identify which concepts can be misleading for a prediction, a definition is required for a correct and incorrect one. The model did a correct prediction if the *GCD* error to the actual location is smaller than a user-defined radius  $r$ . We set this radius to country level ( $r = 750 \text{ km}$ ) since it provides a good balance between correct and incorrect predicted samples. In addition to the instance level matches, global features learned by the model are reflected.

## 4.2.2 Influence of Individual Concepts

The following observations can be made in Table 3 where the results are reported for the YFCC-Val26k dataset. Please note that only a subset of concepts is used where at least 100 images per concept are available. Concepts like *sky*, *building*, *mountain* or *animal* have a relatively high influence ( $ci \approx 1$ ) on a correct prediction measured both on  $ci_{\text{median}}$  and  $ci_{\mu}$ , which suggests that these concepts contain useful visual cues for prediction. On the contrary, concepts like *road*, *grass*, *water*, *earth*, or *sand* have low values on all measurements, which seems reasonable since these concepts are rather general concepts that are visually similar all over the world. The concept of *sky* has an initially surprisingly large influence on the prediction. The two examples in Figure 3 (last two images) indicate that architectural details of buildings or peaks of mountain ranges can be relevant, i.e., the sky-touching concepts. With the introduction of the morphological dilation ( $m^{\beta}(\mathbf{x}, s)$ ) this area is covered. A repetition of this experiment with the enlarged area ( $\beta = 3$ ) confirmed the assumption (Appendix B) that this issue is one reason for the relatively high  $ci$  score of concept *sky*.

In the case of an incorrect prediction, the ratios change. Most interesting, the measurements increase drastically for the concept *person*. This concept has a greater influence on wrong predictions than on correct ones. In addition, the higher number of images where people are visible is significantly higher when false predictions are made. The same observation can be made for other concepts that are not fixed to one specific place (e.g., *animal*, *car* or *floor*). This supports the assumption that some concepts can lead to confusions (Figure 3 first image) or are not useful for geoestimation, and can result in incorrect predictions (indicated by the pure number of incorrect predictions (Table 3)).

The examples in Figure 3 show an additional property of the presented metric. It does not determine the concrete concept that is crucial for a prediction but rather which concepts are influential. Unfortunately, the magnitude for  $ci > 1$  can hardly be assessed due to the unavoidable normalization with the concept size. In summary, we have identified concrete semantic concepts that contain geo-information features and have shown that some of them can be misleading in case of false predictions.

## 5. Conclusions

In this paper, we have presented a novel semantic photo geolocalization system that allows for the interpretation of results. To achieve this, we have proposed a semantic partitioning method which naturally leads to an improved comprehensibility of predictions, while at the same time state-of-the-art results are achieved with respect to geolocation accuracy on three common benchmark test sets. In addition, we have suggested a novel metric to assess the impor-

tance of semantic visual concepts for a certain prediction to provide additional explanatory information, and to allow for a large-scale analysis of already trained models. In the future, we plan to incorporate visual similarities between classes based on geographical features during optimization, e.g., derived from a knowledge base since currently visual and spatial proximate classes are equally penalized as visual and spatial dissimilar classes.

## A. Model Pre-Training

It was mentioned that the initial weights from all trained models are not selected randomly or taken from a model that was trained on *ImageNet* [39] as commonly chosen, but with weights of a pre-trained model which already learned geo-related features to reduce the training time. This sections provides missing details about the weight initialization.

**Task and Data:** To construct a simple geolocation classification problem, each image is associated with the country (including maritime borders) the image was taken. Note, that an image is associated to its country by checking whether its respective coordinate is within the area that the country covers and no function of the method presented in Section 3 like *Reverse GeoCoding* is applied. As data basis serves the MP-16 [21] dataset for training and *YFCC-Val26k* for validation. Due to class imbalance, the USA is divided into individual states and a country is only considered if there are at least 1,000 images for training which results in 173 classes in total.

**Network Training and Results:** One ResNet-50 is initialized with *ImageNet* weights and optimized using the SGDR [27] method with weight decay of  $10^{-4}$ , a momentum of 0.9, a maximum learning rate of  $5 \times 10^{-3}$  and one cycle is performed three times per epoch. The model is trained for ten epochs and is validated every 4,000 steps for a batch size of 128 (512,000 images). The same image pre-processing and data augmentation methods are applied as described in the paper. The standard classification loss (cross-entropy) is used as loss function, i.e. one-hot-encoded vectors as ground-truth class) and the best model on the validation set is selected.

After training, this trained model achieves a top-1 accuracy of 0.24 percent (0.49 top-5).

## B. Influence of Individual Concepts

The Table 4 allows a direct comparison to the results reported in the paper. The *concept influence* ( $ci$ ) with  $\beta = 3$  is summarized in contrast to  $\beta = 0$  as in the paper for each concept in *YFCC-Val26k* using median and mean ( $\mu$ ) and

Concept	GCD $\leq 750$ km			GCD $> 750$ km		
	Images	$ci_{\text{median}}$	$ci_{\mu}$	Images	$ci_{\text{median}}$	$ci_{\mu}$
windowpane	115	1.67	2.09	283	1.67	2.08
mountain	897	1.42	1.91	786	1.21	1.67
sky	4252	1.39	1.78	4291	1.12	1.58
animal	174	1.36	1.73	541	1.45	1.85
house	101	1.31	1.49			
airplane	103	1.24	1.50	175	1.32	1.61
building	3296	1.18	1.47	2818	1.12	1.42
flower	120	1.01	1.19	481	1.11	1.30
tree	2598	0.99	1.34	3078	1.00	1.36
person	3039	0.90	1.10	6594	1.12	1.31
bridge	106	0.86	1.32			
wall	2726	0.86	0.95	7713	0.89	1.00
ceiling	417	0.70	1.08	1137	0.77	1.03
painting	147	0.66	0.84	423	0.87	1.00
sea	460	0.58	0.95	559	0.74	1.05
signboard	280	0.57	0.99	612	0.63	0.97
door	104	0.53	0.74	259	0.59	0.88
rock	382	0.53	0.95	833	0.78	1.10
car	332	0.49	0.83	465	0.63	0.86
floor	749	0.32	0.59	2126	0.41	0.65
fence	245	0.32	0.58	308	0.39	0.71
boat	141	0.30	0.67	161	0.50	0.88
water	840	0.27	0.55	1562	0.55	0.71
earth	1336	0.27	0.53	2184	0.41	0.63
plant	561	0.26	0.51	1294	0.52	0.70
field	186	0.26	0.56	192	0.33	0.59
grass	984	0.24	0.45	1170	0.32	0.56
sand	160	0.18	0.48	208	0.30	0.55
table	111	0.17	0.51	567	0.41	0.65
road	917	0.12	0.35	799	0.17	0.42
sidewalk	468	0.09	0.34	392	0.19	0.52

Table 4.

separated in correct and incorrect *GCD* prediction. Again, only concepts are reported with at least 100 samples.

## References

- [1] J. Adebayo, J. Gilmer, M. Muelly, I. J. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 9525–9536, 2018. URL <http://papers.nips.cc/paper/8160-sanity-checks-for-saliency-maps.2>
- [2] M. Ancona, E. Ceolini, A. C. Öztireli, and M. H. Gross. A unified view of gradient-based attribution methods for deep neural networks. *CoRR*, abs/1711.06104, 2017. URL <http://arxiv.org/abs/1711.06104.2>
- [3] R. Andrews, J. Diederich, and A. B. Tickle. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowl. Based Syst.*, 8(6):373–389, 1995. doi: 10.1016/0950-7051(96)81920-4. 2
- [4] R. Arandjelovic, P. Gronát, A. Torii, T. Pajdla, and J. Sivic. Netvlad: CNN architecture for weakly supervised place recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1437–1451, 2018. doi: 10.1109/TPAMI.2017.2711011. 1, 2
- [5] G. Baatz, O. Saurer, K. Köser, and M. Pollefeys. Large scale visual geo-localization of images in mountainous terrain. In A. W. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision - ECCV 2012 - 12th Eu-*

- ropean Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II*, volume 7573 of *Lecture Notes in Computer Science*, pages 517–530. Springer, 2012. doi: 10.1007/978-3-642-33709-3\\_\\_37. URL [https://doi.org/10.1007/978-3-642-33709-3\\_37](https://doi.org/10.1007/978-3-642-33709-3_37). 1, 2
- [6] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3319–3327. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.354. 2
- [7] A. Dombrowski, M. Alber, C. J. Anders, M. Ackermann, K. Müller, and P. Kessel. Explanations can be manipulated and geometry is to blame. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 13567–13578, 2019. URL <http://papers.nips.cc/paper/9511-explanations-can-be-manipulated-and-geometry-is-to-blame>. 2
- [8] M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.*, 111(1):98–136, 2015. doi: 10.1007/s11263-014-0733-5. 1
- [9] A. Ghorbani, A. Abid, and J. Y. Zou. Interpretation of neural networks is fragile. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3681–3688. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.33013681. 4, 5
- [10] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. In F. Bonchi, F. J. Provost, T. Eliassi-Rad, W. Wang, C. Cattuto, and R. Ghani, editors, *5th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2018, Turin, Italy, October 1-3, 2018*, pages 80–89. IEEE, 2018. doi: 10.1109/DSAA.2018.00018. 1, 2, 6
- [11] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. Deep image retrieval: Learning global representations for image search. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, volume 9910 of *Lecture Notes in Computer Science*, pages 241–257. Springer, 2016. doi: 10.1007/978-3-319-46466-4\\_\\_15. URL [https://doi.org/10.1007/978-3-319-46466-4\\_15](https://doi.org/10.1007/978-3-319-46466-4_15). 2
- [12] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5):93:1–93:42, 2019. doi: 10.1145/3236009. 1, 2
- [13] J. Hays and A. A. Efros. IM2GPS: estimating geographic information from a single image. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*. IEEE Computer Society, 2008. doi: 10.1109/CVPR.2008.4587784. 1, 2, 5, 6
- [14] J. Hays and A. A. Efros. Large-scale image geolocation. In J. Choi and G. Friedland, editors, *Multimodal Location Estimation of Videos and Images*, pages 41–62. Springer, 2015. doi: 10.1007/978-3-319-09861-6\\_\\_3. URL [https://doi.org/10.1007/978-3-319-09861-6\\_3](https://doi.org/10.1007/978-3-319-09861-6_3). 2
- [15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. 1, 6
- [16] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in Computer Science*, pages 630–645. Springer, 2016. doi: 10.1007/978-3-319-46493-0\\_\\_38. URL [https://doi.org/10.1007/978-3-319-46493-0\\_38](https://doi.org/10.1007/978-3-319-46493-0_38). 6
- [17] S. Hooker, D. Erhan, P. Kindermans, and B. Kim. A benchmark for interpretability methods in deep neural networks. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 9734–9745, 2019. URL <http://papers.nips.cc/paper/9167-a-benchmark-for-interpretability-methods-in-deep-neural-networks>. 8
- [18] M. Izbicki, E. E. Papalexakis, and V. J. Tsotras. Exploiting the earth’s spherical geometry to geolocate images. In U. Brefeld, É. Fromont, A. Hotho, A. J. Knobbe, M. H. Maathuis, and C. Robardet, editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2019, Würzburg, Germany, September 16-20, 2019, Proceedings, Part II*, volume 11907 of *Lecture Notes in Computer Science*, pages 3–19. Springer, 2019. doi: 10.1007/978-3-030-46147-8\\_\\_1. URL [https://doi.org/10.1007/978-3-030-46147-8\\_1](https://doi.org/10.1007/978-3-030-46147-8_1). 1, 2, 7
- [19] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In D. A. Forsyth, P. H. S. Torr, and A. Zisserman, editors, *Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I*, volume 5302 of *Lecture Notes in Computer Science*, pages 304–317. Springer, 2008. doi: 10.1007/978-3-540-88682-2\\_\\_24. URL [https://doi.org/10.1007/978-3-540-88682-2\\_24](https://doi.org/10.1007/978-3-540-88682-2_24). 2
- [20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, 2017. doi: 10.1145/3065386. 1

- [21] M. A. Larson, M. Soleymani, G. Gravier, B. Ionescu, and G. J. F. Jones. The benchmarking initiative for multimedia evaluation: Mediaeval 2016. *IEEE Multim.*, 24(1):93–96, 2017. doi: 10.1109/MMUL.2017.9. 5, 9
- [22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. Publisher: Ieee. 1
- [23] S. Lee, H. Zhang, and D. J. Crandall. Predicting geo-informative attributes in large-scale image collections using convolutional neural networks. In *2015 IEEE Winter Conference on Applications of Computer Vision, WACV 2015, Waikoloa, HI, USA, January 5-9, 2015*, pages 550–557. IEEE Computer Society, 2015. doi: 10.1109/WACV.2015.79. 2
- [24] T. Lin, S. J. Belongie, and J. Hays. Cross-view image geolocation. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 891–898. IEEE Computer Society, 2013. doi: 10.1109/CVPR.2013.120. 1, 2
- [25] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In D. J. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1\\_48. URL [https://doi.org/10.1007/978-3-319-10602-1\\_48](https://doi.org/10.1007/978-3-319-10602-1_48). 1
- [26] L. Liu, H. Li, and Y. Dai. Stochastic attraction-repulsion embedding for large scale image localization. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2570–2579. IEEE, 2019. doi: 10.1109/ICCV.2019.00266. 2
- [27] I. Loshchilov and F. Hutter. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL <https://openreview.net/forum?id=Skq89Scxx>. 9
- [28] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 289–297, 2016. URL <http://papers.nips.cc/paper/6202-hierarchical-question-image-co-attention-for-visual-question-answering>. 2
- [29] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38, 2019. doi: 10.1016/j.artint.2018.07.007. 2
- [30] E. Müller-Budack, K. Pustu-Iren, and R. Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII*, volume 11216 of *Lecture Notes in Computer Science*, pages 575–592. Springer, 2018. doi: 10.1007/978-3-030-01258-8\\_35. URL [https://doi.org/10.1007/978-3-030-01258-8\\_35](https://doi.org/10.1007/978-3-030-01258-8_35). 1, 2, 3, 4, 5, 6, 7
- [31] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han. Large-scale image retrieval with attentive deep local features. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3476–3485. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.374. 1, 2
- [32] Nominatim. Nominatim. <https://https://nominatim.org/>. Accessed: 2029-11-14. 3
- [33] C. Olah, A. Mordvintsev, and L. Schubert. Feature visualization. *Distill*, 2(11):e7, 2017. 2
- [34] OpenStreetMap contributors. Planet dump retrieved from <https://planet.osm.org>. <https://www.openstreetmap.org>, 2017. 2
- [35] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*. IEEE Computer Society, 2007. doi: 10.1109/CVPR.2007.383172. 2
- [36] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*. IEEE Computer Society, 2008. doi: 10.1109/CVPR.2008.4587635. 2
- [37] F. Radenovic, G. Tolias, and O. Chum. CNN image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, volume 9905 of *Lecture Notes in Computer Science*, pages 3–20. Springer, 2016. doi: 10.1007/978-3-319-46448-0\\_1. URL [https://doi.org/10.1007/978-3-319-46448-0\\_1](https://doi.org/10.1007/978-3-319-46448-0_1). 2
- [38] A. S. Ross, M. C. Hughes, and F. Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In C. Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 2662–2670. ijcai.org, 2017. doi: 10.24963/ijcai.2017/371. 2
- [39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. 1, 9
- [40] s2library. S2library. <https://s2geometry.io/>. Accessed: 2029-11-14. 2
- [41] T. Sattler, M. Havlena, K. Schindler, and M. Pollefeys. Large-scale location recognition and the geometric burstiness problem. In *2016 IEEE Conference on Computer Vi-*

- sion and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1582–1590. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.175. 1, 2
- [42] O. Saurer, G. Baatz, K. Köser, L. Ladicky, and M. Pollefeys. Image based geo-localization in the alps. *Int. J. Comput. Vis.*, 116(3):213–225, 2016. doi: 10.1007/s11263-015-0830-0. 1, 2
- [43] G. Schindler, M. A. Brown, and R. Szeliski. City-scale location recognition. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*. IEEE Computer Society, 2007. doi: 10.1109/CVPR.2007.383150. 1, 2
- [44] G. P. Schmitz, C. Aldrich, and F. S. Gouws. ANN-DT: an algorithm for extraction of decision trees from artificial neural networks. *IEEE Transactions on Neural Networks*, 10(6):1392–1401, 1999. Publisher: IEEE. 2
- [45] P. H. Seo, T. Weyand, J. Sim, and B. Han. Cplanet: Enhancing image geolocation by combinatorial partitioning of maps. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*, volume 11214 of *Lecture Notes in Computer Science*, pages 544–560. Springer, 2018. doi: 10.1007/978-3-030-01249-6\_33. URL [https://doi.org/10.1007/978-3-030-01249-6\\_33](https://doi.org/10.1007/978-3-030-01249-6_33). 1, 2, 3, 6, 7
- [46] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Y. Bengio and Y. LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014. URL <http://arxiv.org/abs/1312.6034>. 2, 4
- [47] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017. URL <http://arxiv.org/abs/1706.03825>. 4, 8
- [48] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller. Striving for simplicity: The all convolutional net. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. URL <http://arxiv.org/abs/1412.6806>. 4
- [49] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang. High-resolution representations for labeling pixels and regions. *CoRR*, abs/1904.04514, 2019. URL <http://arxiv.org/abs/1904.04514>. 7
- [50] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 2017. URL <http://proceedings.mlr.press/v70/sundararajan17a.html>. 4, 7, 8
- [51] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. Li. YFCC100M: the new data in multimedia research. *Commun. ACM*, 59(2):64–73, 2016. doi: 10.1145/2812802. 5
- [52] G. Tolias, R. Sicre, and H. Jégou. Particular object retrieval with integral max-pooling of CNN activations. In Y. Bengio and Y. LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL <http://arxiv.org/abs/1511.05879>. 2
- [53] E. Tzeng, A. Zhai, M. Clements, R. Townshend, and A. Zakhor. User-driven geolocation of untagged desert imagery using digital elevation models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2013, Portland, OR, USA, June 23-28, 2013*, pages 237–244. IEEE Computer Society, 2013. doi: 10.1109/CVPRW.2013.42. 1, 2
- [54] N. N. Vo, N. Jacobs, and J. Hays. Revisiting IM2GPS in the deep learning era. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2640–2649. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.286. 1, 2, 5, 6
- [55] T. Weyand, I. Kostrikov, and J. Philbin. Planet - photo geolocation with convolutional neural networks. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, volume 9912 of *Lecture Notes in Computer Science*, pages 37–55. Springer, 2016. doi: 10.1007/978-3-319-46484-8\_3. URL [https://doi.org/10.1007/978-3-319-46484-8\\_3](https://doi.org/10.1007/978-3-319-46484-8_3). 1, 2, 3, 6
- [56] Q. Zhang and S. Zhu. Visual interpretability for deep learning: a survey. *Frontiers Inf. Technol. Electron. Eng.*, 19(1):27–39, 2018. doi: 10.1631/FITEE.1700808. 2
- [57] Y. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T. Chua, and H. Neven. Tour the world: Building a web-scale landmark recognition engine. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 1085–1092. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206749. 1, 2
- [58] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL <http://arxiv.org/abs/1412.6856>. 2
- [59] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ADE20K dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5122–5130. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.544. 7
- [60] J. R. Zilke, E. L. Mencía, and F. Janssen. Deepred - rule extraction from deep neural networks. In T. Calders, M. Ceci, and D. Malerba, editors, *Discovery Science - 19th International Conference, DS 2016, Bari, Italy, October 19-21,*

2016, *Proceedings*, volume 9956 of *Lecture Notes in Computer Science*, pages 457–473, 2016. doi: 10.1007/978-3-319-46307-0\_29. URL [https://doi.org/10.1007/978-3-319-46307-0\\_29](https://doi.org/10.1007/978-3-319-46307-0_29). 2