Object-Level Explanations for Image Geolocation Models: a GeoGuessr use-case

Emilie Durrieu
ENAC, University of Toulouse
emilie.durrieu@utoulouse.fr

Christophe Hurter
ENAC, University of Toulouse
christohe.hurter@enac.fr

Philippe Muller
IRIT, University of Toulouse, ANITI
philippe.muller@irit.fr

Victor Boutin
CNRS
victor.boutin@cnrs.fr

Abstract

When humans play geolocation games such as GeoGuessr, they rely on concrete visual cues, such as road markings, vegetation, or architectural details, to infer where an image was captured. Whether image geolocation models rely on similar object-level evidence remains difficult to determine, as attribution methods like Grad-CAM typically highlight diffuse regions rather than coherent visual entities, making it difficult to link model predictions to specific objects or perceptible patterns. In this work, we propose an object-centric analysis pipeline to investigate the visual evidence used by geolocation models. Starting from attribution maps, we extract salient regions and segment them into object-like elements. We evaluate their predictive relevance through deletion and insertion tests, comparing attribution-guided crops to randomly selected regions with similar coverage. Experiments on a three-country benchmark show that attribution-guided crops consistently retain more information for the model's prediction than random crops. These results suggest that attribution maps can be decomposed into interpretable, perceptible elements, providing a step toward object-level analysis of geolocation models.

1. Introduction

Image-based geolocation aims to predict where an image was captured using only its visual content. Beyond practical applications such as location verification or geographic indexing, this task has gained public attention through games such as *Geoguessr* [1] and *WorldGuessr*[2]. When humans solve these tasks, they typically rely on recognizable objects and structures, such as road signs, vegetation, vehicles, and architectural elements.

[1] https://www.geoguessr.com
[2] https://www.worldguessr.com

Modern Convolutional Neural Networks (CNNs) achieve strong performance by framing geolocation as a large-scale classification problem [20]. However, understanding which visual evidence drives these predictions remains difficult, limiting our ability to interpret, trust, and analyze geolocation models at a meaningful, concept-level granularity. Moreover, attribution maps such as those obtained with GradCAM [14] often highlight broad or overlapping regions that do not correspond to discrete, perceptible objects, making it difficult to link model decisions to interpretable visual patterns.

To address this, we propose an object-centric analysis pipeline for geolocation interpretability. Attribution maps produced from a trained classifier are first thresholded to identify salient regions. These regions are then segmented into object-like elements using a segmentation model, producing discrete visual units that can be analyzed individually. Finally, we evaluate whether these extracted elements preserve predictive information through deletion and insertion-based faithfulness tests (see Figure 1).

Our experiments on a three-country benchmark show that attribution-guided object-like elements retain more predictive information than randomly selected regions of similar coverage. These results suggest that attribution maps contain localized evidence that can be meaningfully decomposed into perceptible visual regions, opening new opportunities for object-level interpretability and concept-based explanations in image geolocation models.

2. Related Work

2.1. Geolocation Models

Automatic image geolocation aims to predict the location where an image was taken, using either global descriptors or deep learning features. Early methods such as IM2GPS [7] relied on hand-crafted features and nearest-neighbor retrieval over large image databases. More recent approaches

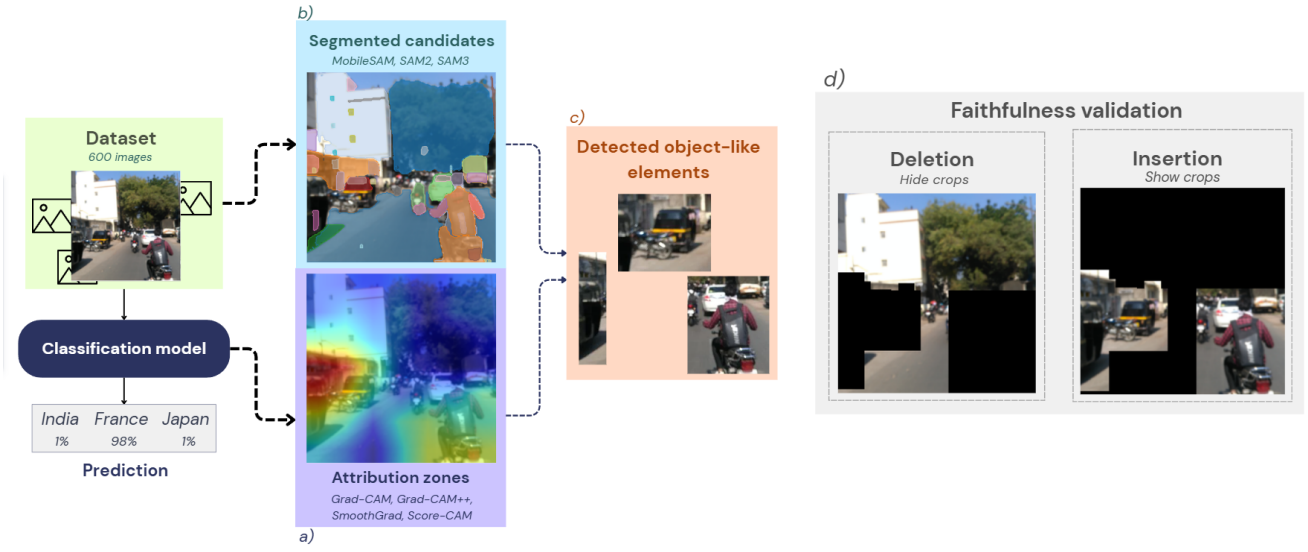

Figure 1. Overview of the proposed object-centric pipeline: (a) Saliency maps are extracted from a trained classifier using GradCAM++. (b) Images are segmented into candidate regions. (c) Each segment is scored based on overlap with attribution maps to identify relevant object-like elements. (d) Faithfulness validation is performed with a insertion/deletion test to assess their predictive relevance.

employ deep convolutional neural networks to formulate geolocation as a classification problem: PlaNet [20] divides the world into discrete cells and predicts the most likely location, while IM2GPS Revisited [18] and PIGEON [6] improve accuracy using CNN features and hierarchical modeling. These methods achieve impressive performance on both world-scale and street-level datasets, but they generally focus on prediction accuracy rather than understanding the visual cues that drive their decisions.

2.2. Explainable AI in vision

Attribution and visualization methods in XAI aim to reveal which parts of an image influence a model's prediction. Gradient-based methods such as GradCAM [14] and Integrated Gradients [17] generate class-specific attention maps from convolutional features, while perturbation-based methods like LIME [13] and RISE [11] estimate importance by observing output changes under input modifications. Gradient-based approaches are particularly convenient for our study as they provide smooth and spatially coherent maps that can be efficiently computed and combined with segmentation to isolate object-like regions. Applying XAI to geolocation models is relatively unexplored at the object level. Shi et al. [15] visualize city-level CNN predictions to highlight relevant regions, showing that urban structures and landmarks drive the model's decisions. Our work builds on this idea but focuses on object-level interpretability, extracting object-like elements from attribution-guided regions to study their relevance.

2.3. Concept-based XAI

Concept-based explanations provide a more structured view of model reasoning by linking predictions to human-interpretable concepts. Concept Bottleneck Models [9] train networks to first predict predefined concepts before outputting a final label, ensuring interpretability by design. Automated Concept-based Explanation (ACE) [5] extracts visual concepts from images using segmentation and clustering, then evaluates their relevance to model predictions. In contrast, our approach focuses on the geolocation context and extracts object-like elements directly from attribution maps, without relying on predefined concepts or modifying the model. This enables a post-hoc, object-level analysis of the visual cues driving geolocation predictions.

3. Methodology

We investigate whether visual regions highlighted by attribution methods capture predictive information in geolocation models. Our approach consists of three main steps illustrated in Figure 1: (1) Train a CNN classifier for geolocation. (2) Extract attribution-guided regions from model predictions. (3) Decompose these regions into object-like elements and evaluate their predictive relevance.

3.1. Problem Setup

Let f be a Convolutional Neural Network (CNN) trained to predict the country of origin among N possible countries. Given an input image x representing a ground-level geolocation image, the model outputs a logit vector $f(x)$ in $\mathbb{R}^N$,

which is then converted into a probability distribution via softmax. Our goal is to identify the visual cues present in x that contribute to the model's predictions.

3.2. Obtaining visual explanations

To localize the most predictive image regions, we apply feature attribution methods to our model f. For a given input image and predicted class, these methods produce an attention map highlighting pixels that strongly influence the output. We threshold the map to retain the most salient pixels by selecting the top-p percentile of pixels based on attribution values, with p chosen empirically by evaluating the resulting coverage and faithfulness on a validation set. This strategy balances keeping highly informative pixels while avoiding overly large or diffuse regions. We refer to the obtained salient maps as *attribution-guided regions*.

3.3. Extracting object-like elements

Within each attribution-guided region, we extract discrete visual elements using a segmentation model. We define an *object-like element* as a perceptible visual unit, such as a car or a street sign. These elements are not assigned semantic labels, and their shapes depend on the segmentation output. Each candidate segment is assigned a relevance score based on three factors:

1. **Overlap with high-saliency regions:** the fraction of the segment that overlaps pixels above a threshold in the attribution map.
2. **Average importance:** the mean attribution value within the segment.
3. **Central importance:** the attribution value at the segment's geometric center.

These factors are then combined using the geometric mean.

Segments below threshold $s_{\min}$ are discarded, and the remaining segments are ranked by score. To avoid redundancy, overlapping segments are filtered using a containment-based IoU criterion, which ensures small details are preserved while near-duplicate segments are removed. Finally, each segment is converted into a rectangular bounding box and slightly padded to preserve contextual information. These rectangular zones constitute the final set of object-like elements used for evaluation.

4. Our extraction method

Dataset: We use images from the dataset OSV-5M [1] containing over 5 million street-level images from 225 different countries collected via the Mapillary API. For our study, we select images from three countries: France, India, and Japan, and retrieve the original high-resolution images through the API. The data is split into 100k images (33,333 per country) for training and 600 images (200 per country) for pipeline evaluation. All images are resized to 224×224 pixels to match the input resolution of the classifier. Pixel values are normalized using ImageNet statistics. During training, standard data augmentation is applied, including random horizontal flipping and color jitter, while evaluation images are processed without augmentation.

Classification Model: We finetune a ResNet50 [8] pretrained on ImageNet [4] for country classification. The network is trained with Cross-Entropy loss with label smoothing of 0.1, using the AdamW optimizer. The learning rate was set to 3e-4, with a weight decay of 0.02. The model was trained for 300 epochs with early stopping based on validation loss (patience = 30). The model reached an accuracy of 88% on the training set and 87.42% on the test set, as well as 86.3% on the 600 pipeline images.

Attention regions: To identify regions that influence the model's predictions, we apply several attribution methods: Grad-CAM, Grad-CAM++, SmoothGrad, and Score-CAM. GradCAM [14] produces class-specific localization maps by weighting convolutional feature maps using gradients, while GradCAM++ [3] refines this process by using pixel-wise and higher-order gradient weighting, thereby improving multi-instance localization. SmoothGrad [16] reduces noise by averaging gradients over multiple perturbed inputs, and Score-CAM [19] generates class activation maps in a gradient-free manner using forward-pass class scores.

Object extraction: Within attribution-guided regions, we extract perceptible object-like elements using segmentation models. We experiment with three Segment Anything Model (SAM) variants: lightweight *MobileSAM* [10], *SAM2* [12], and the recent *SAM3* [2] which extends SAM2 with concept-guided prompting. For SAM3, the employed concepts are extracted from a Geoguessr guidebook using an LLM, resulting in roughly 200 candidate concepts associated with geolocation.

5. Evaluation

The goal of this evaluation is to determine whether the object-like elements extracted by our pipeline correspond to visual regions that are genuinely important for the geolocation model's prediction. In particular, we investigate whether these extracted regions retain predictive information by performing two faithfulness tests commonly used in explainable AI:

1. *Deletion*: Extracted crops are occluded in the image, and we measure the drop in classification accuracy. A larger drop indicates that these removed crops contained important information for the model's prediction.
2. *Insertion*: Only the extracted crops are preserved while the rest of the image is masked. The model predicts from this reduced input, and the retained classification accuracy quantifies how much predictive information is contained within the object-like elements.

For comparison, we generate random crops with similar coverage and size constraints. The procedure is repeated

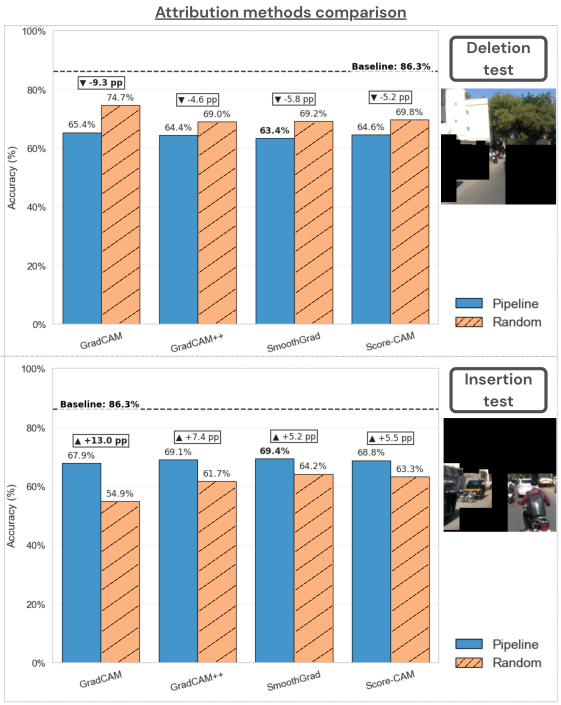

Figure 2. Comparison of attribution methods (GradCAM, GradCAM++, SmoothGrad, Score-CAM) using MobileSAM as the segmentation model.

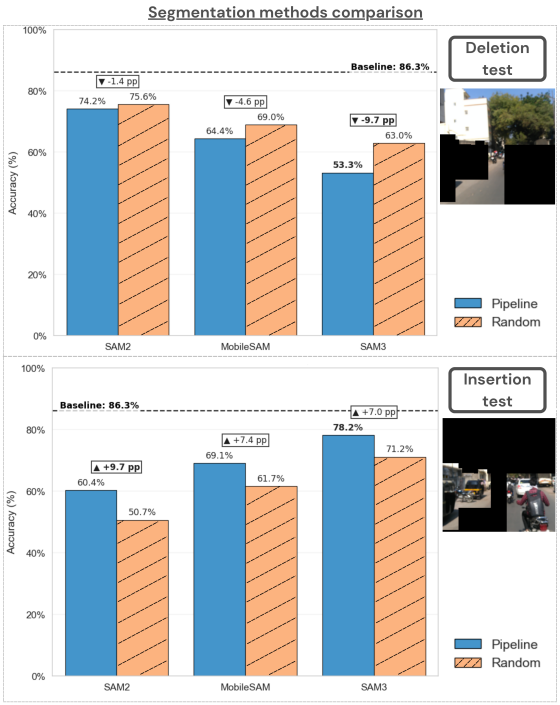

Figure 3. Comparison of segmentation methods (MobileSAM, SAM2, SAM3) using GradCAM++ as the attribution method.

ten times per image, and the results are averaged to reduce variance. This baseline allows us to determine whether attribution-guided extraction identifies more informative regions than random selection.

We further analyze how different attribution methods (Figure 2) and segmentation strategies (Figure 3) influence the effectiveness of the extracted regions.

Across all tested configurations, attribution-guided crops consistently outperform random crops. In the deletion test, occluding these regions causes a larger drop in accuracy, indicating that they correspond to areas that are important for the model's predictions. In the insertion test, retaining only these regions preserves significantly more predictive information than random crops of similar coverage. These results suggest that attribution maps contain structured predictive signals that can be decomposed into localized visual elements without losing task-relevant information.

Interestingly, segmentation quality appears to influence the results. As shown in Figure 3, SAM2 performs poorly in the Deletion test, which may be explained by the relatively small number of segments it generates. This results in lower spatial coverage of the image, limiting the amount of information removed when the crops are ablated. In contrast, SAM3 produces a significantly larger number of segments than the other methods and achieves stronger results in both ablation settings. This suggests that higher segmentation coverage enables the pipeline to capture a broader set of visual cues contributing to the model's predictions.

Qualitative inspection of the extracted object-like elements reveals a diverse set of visual patterns, including vehicles, walls, street signs, and road markings (see Figure 4). This diversity suggests that the pipeline captures a range of perceptible cues that can contribute to geolocation predictions. At the same time, some elements overlap or correspond only to parts of larger structures, while others remain amorphous, reflecting limitations of the segmentation step.

6. Conclusion

We present a post-hoc, object-centric framework for interpreting image geolocation models by transforming attribution maps into discrete, object-like elements that can be evaluated directly. By combining existing attribution and segmentation methods as well as deletion/insertion faithfulness tests, our approach bridges diffuse pixel-level explanations and localized visual evidence without requiring predefined concepts or modifications to the underlying classifier. On a three-country benchmark, attribution-guided crops consistently outperform random regions with comparable coverage, indicating that the extracted elements retain meaningful predictive information and that predictions can be traced to perceptible visual cues. Our results also suggest that segmentation quality plays an important role in this

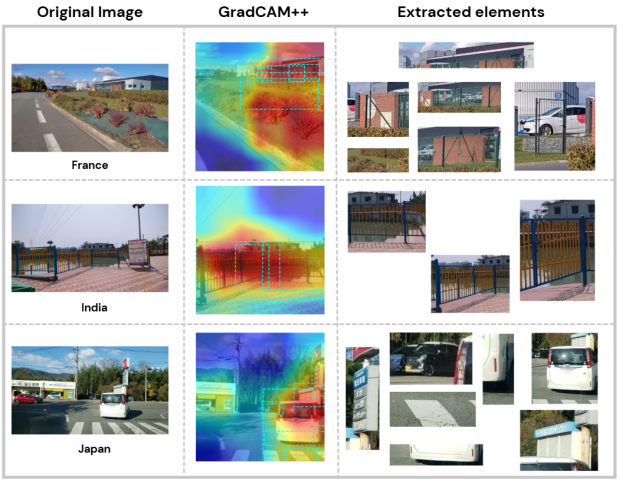

Figure 4. Attribution-guided object-like elements extracted from geolocation images. The method identifies diverse visual cues, including vehicles, road markings, and barriers, although some regions correspond to partial or overlapping structures.

analysis, as broader and more appropriate region coverage leads to stronger faithfulness outcomes.

At the same time, the extracted elements remain approximate and may be over-segmented, and our evaluation is limited to a three-country setting without additional baselines. Future work includes extending to larger, fine-grained benchmarks, explore alternative attribution and region-selection strategies, and incorporate human evaluation to assess whether the visual evidence identified by the model aligns with human reasoning in geolocation tasks such as GeoGuessr.

References

[1] Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, Constantin Aronssohn, Nacim Bouia, Stephanie Fu, Romain Loiseau, Van Nguyen Nguyen, Charles Raude, Elliot Vincent, Lintao Xu, Hongyu Zhou, and Loic Landrieu. OpenStreetView-5M: The Many Roads to Global Visual Geolocation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21967–21977, 2024. 3

[2] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with Concepts, 2025. 3

[3] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. 3

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[5] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards Automatic Concept-based Explanations. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 2

[6] Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. PIGEON: Predicting Image Geolocations. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12893–12902, 2024. 2

[7] James Hays and Alexei A. Efros. IM2GPS: Estimating geographic information from a single image. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 1

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[9] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept Bottleneck Models. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020. 2

[10] Yehui Liu, Yuliang Zhao, Xinyue Zhang, Xiaoai Wang, Chao Lian, Jian Li, Peng Shan, Changzeng Fu, Xiaoyong Lyu, Lianjiang Li, Qiang Fu, and Wen Jung Li. MobileSAM-Track: Lightweight One-Shot Tracking and Segmentation of Small Objects on Edge Devices. *Remote Sensing*, 15(24), 2023. 3

[11] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models, 2018. 2

[12] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos, 2024. 3

[13] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California, 2016. Association for Computational Linguistics. 2

[14] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks

via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 1, 2, 3

[15] Xiangwei Shi, Seyran Khademi, and Jan van Gemert. Deep Visual City Recognition Visualization, 2019. 2

[16] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: Removing noise by adding noise, 2017. 3

[17] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017. 2

[18] Nam Vo, Nathan Jacobs, and James Hays. Revisiting IM2GPS in the Deep Learning Era. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2640–2649, 2017. 2

[19] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 24–25, 2020. 3

[20] Tobias Weyand, Ilya Kostrikov, and James Philbin. PlaNet - Photo Geolocation with Convolutional Neural Networks. In *Computer Vision – ECCV 2016*, pages 37–55, Cham, 2016. Springer International Publishing. 1, 2