Artificial Intelligence Research and Development
A. Cortés et al. (Eds.)

doi:10.3233/FAIA220349

Automatic Outdoor Image Geolocation with Focal Modulation Networks

Fabio MURGESE [a,1], Gerard ALCAINA [a,2], Mehmet Oğuz MÜLÂYİM [a], Jesus CERQUIDES [a], and Jose Luis FERNANDEZ-MARQUEZ [b]

[a] *Artificial Intelligence Research Institute (IIIA), CSIC, Cerdanyola del Vallès, Spain*
[b] *Citizen Cyberlab, University of Geneva, Geneva, Switzerland*

Abstract.
We address the problem of estimating a photo's geographical location. Success in this estimation enables many impactful applications, like facilitating Disaster Management circumstances. However, this is also a very challenging task. Due to the complexity of the problem, we restrict the area of geolocation to a single city, treating geolocation as a classification problem where the districts of a city are the classes to be distinguished. In this paper, we exploit the *Focal Modulation Network* that is proven to perform effectively and efficiently in visual modeling for real-world applications. Experimental results on two diverse datasets, crawled from online sources, show the effectiveness of our approach. We can geolocate correctly more than two-thirds of test images from the larger dataset and about one-third from an experimental training dataset of a ten-times smaller size.

Keywords. Image Classification; Focal Modulation Networks; Outdoor Image Geolocation; Barcelona Geolocation

1. Introduction

Image geolocation (i.e., identifying the geographical location of an image) is a highly challenging task since photos taken even from the same location exhibit immense variations due to different camera settings, seasons, daylight conditions, and present objects. Also, images are often ambiguous and could provide very few cues about their location. In the absence of discriminative landmarks, humans can leverage their world knowledge to infer the location of a photo, using hints like the language of street signs or the driving direction of cars. Most previous work on image geolocation focused on identifying and geolocating landmark buildings (e.g., [3, 14]) whereas very few approaches tried to geolocate images just by using pixels (e.g., [7, 16]).

Our primary motivation for precise outdoor image geolocation is its application to the Disaster Management field, where it can have a critical impact because a fast response is of paramount importance to help emergency aid. Social media data has demonstrated to be extremely relevant to evaluate damage and improve the understanding after a natural disaster occurs [4, 5], especially, in the first 24/48h which are crucial to allow emergency

[1] Corresponding Author: Fabio Murgese, IIIA, CSIC, Cerdanyola 08193. E-mail:fabiomur95@gmail.com
[2] Corresponding Author: Gerard Alcaina, IIIA, CSIC, Cerdanyola 08193. E-mail:galcaina98@gmail.com

responders to coordinate their actions. Finding the location of the social media content is a major challenge to make social media information ready to be used by emergency responders. For example, it is relevant to observe a Twitter photo containing a school which has been damaged after an earthquake, however that information will be hardly usable if we do not know the location where the photo was taken, i.e., where the damaged school is located.

Online image databases and social media provide us with invaluable data for training our Machine Learning models and later quickly geolocating the images shared from disaster areas. In this work, we propose to exploit the *Focal Modulation Network* (FocalNet) [19] which is demonstrated by its authors to outperform the state-of-the-art for effective and efficient visual modeling in real-world applications. The main contribution of this paper is an open-source image crawler software that is able to retrieve all available data from Flickr[3] and Mapillary[4] for a chosen city and attach their district information in order to fit with FocalNets requirements for the training of the geolocator. Consequently, we can discriminate between the areas of a city at different granularities, a non-trivial task within the geolocation field. Specifically, we can choose to distinguish districts or neighborhoods by using alternative data sources, in our case, stored as GeoJSON[5] files that contain the coordinates of the geographical borders of cities and their sub-regions.

Additionally, we publish two datasets that comprise the coordinates and the districts of images that were taken in Barcelona, Spain: 1) The Flickr dataset is made of about 18k images that were crawled by our pipeline posing the districts of the city of Barcelona as query keywords; 2) The Mapillary dataset consists of more than 182k images, with attached geographic coordinates and district information. The photos in the latter dataset were crawled by a recursive algorithm that enabled us to query the Mapillary Application Programming Interface (API) iteratively without worrying about the limit of data that can be retrieved in one shot from the platform. More details will be disclosed in Section 3. Then, we trained different FocalNet models on both datasets to understand which benefits come with the usage of two datasets of very different nature: one large robotic and one smaller, non-robotic.

We introduce related work in Section 2. In Section 3, we preface the process followed to gather data from multiple sources. Then, in Section 4, we give the set-up for experiments and reflect on the performances of trained models and the results. Section 5 summarizes the findings of this work and gives the outlook for the next research steps.

2. Background

In the field of emergency management, social media images have been geolocated by humanitarian communities such as GISCorp[6], VOST Europe[7] and Standby Task Force (SBTF)[8]. Humanitarian networks involved hundred of volunteers contributing remotely.

[3] https://www.flickr.com/
[4] https://www.mapillary.com/
[5] https://geojson.org/
[6] https://www.giscorps.org/
[7] https://vosteurope.org/
[8] https://standbytaskforce.org/

Crowdsourcing participatory platforms have been designed to simplify the manual geolocation of social media images [15]. However, automatic geolocation would help the crowd effort scale, ease the work of the communities by allowing them to focus on only the refinement of the automatically calculated geolocation.

In recent years, there are two main approaches that tackle the problem of geolocating images: by comparison and by classification [12]. The former is based on an information-retrieval approach that matches a query photo against millions of geotagged images, whereas the latter tries to predict the right class using a single trained model.

Initially, Im2GPS [6] attempted to solve the problem by retrieving the neighbors of a query photo in a database of 6 million geotagged Flickr images and geolocating the query by assigning it the location of the nearest match. Nevertheless, due to the difficulty to classify non-generic scenes using this technique, multiple alternatives appeared addressing the problem with image classification.

Later, PlaNet [18] approached the problem in a different way: it is a worldwide image geolocation classifier that divides the Earth surface into multi-scale geographic cells. The main drawback of this technique was the difficulty to cover places where photos are very unlikely to be taken. New techniques inspired by this approach, such as [9, 13], were proven to get better contextual information of the images adding a hierarchical model and scene classifiers.

The novelty introduced by this paper is to use FocalNets [19] within the geolocation context, adopting a classification approach, as they have been proven to outperform the state-of-the-art Self-Attention counterparts [8, 17]. Concisely, focal modulation first aggregates contexts around each query, in order to be able to modulate the query with the combined context. In this way, it simplifies the process by enabling input-dependent token (i.e., the query) interaction. Also, with this model it is possible to generate summarized tokens at distinct levels of granularity applying query-agnostic aggregations. In the end, these contexts are fused into the query vector, after being selectively aggregated in conformity with the query content.

3. Data gathering

When dealing with the outdoor image geolocation problem, first step is to understand the nature of the data that can help solve the task at hand. More specifically, the sources of datasets for geolocation can be categorized as robotic or non-robotic. Robotic ones are

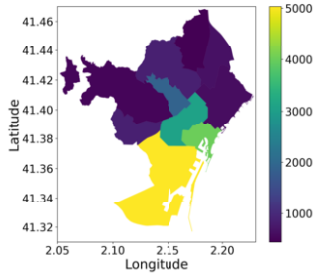
(a) Extracted from Flickr.

(b) Extracted from Mapillary.

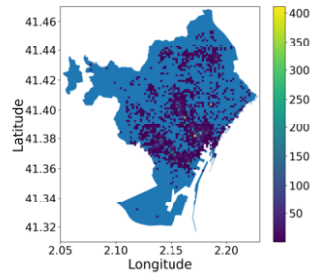
Figure 1. Sample images of non-robotic (left) and robotic (right) styles.

District	Images
Sants-Montjuïc	5,070
Ciutat Vella	4,017
Eixample	3,170
Gràcia	1,888
Les Corts	918
Horta-Guinardó	864
Sant Andreu	651
Sant Martí	603
Sarrià-Sant Gervasi	525
Nou Barris	463

(a) # of images.



(b) Distribution of images.



(c) Exact image locations.

Figure 2. Barcelona Flickr crawled dataset.

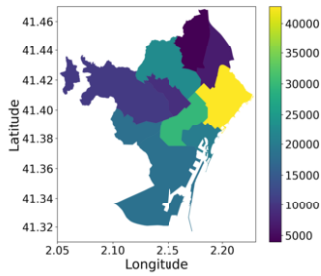
typically created by recording videos from static cameras mounted on a vehicle (e.g., a car, a bicycle), resulting in images available in sequences temporally consistent (assuming the motion of the vehicle) and from a steady point of view (usually the street), with no big changes in the vertical direction, as explained in [12]. Non-robotic datasets, on the other hand, are created by collections of online images, usually taken nearby points of interest, making the geolocation task easy for landmarks, but difficult for not widely-known places. And, here comes the potential value of exploiting robotic datasets, which, by nature, have better coverage of these less considered areas. Figure 1 gives a flavour of the diversity of both styles.

In our case, we use Mapillary to assemble robotic data and Flickr for non-robotic data. Public APIs of these online imagery platforms help us retrieve the maximum amount of images of rich variety for a specific region. Once we retrieve data for a specific city from the two sources, we attach the district information for every image leveraging a GeoJSON file for the city of interest, thus facilitating the replicability of the pipeline for other cities.

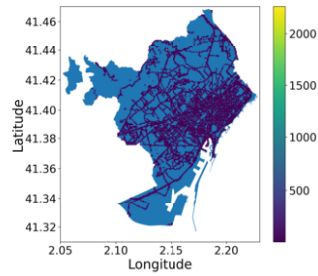
The first dataset we curated is non-robotic and consists of images crawled from Flickr. To query the Flickr API, we used the words representing the ten districts of Barcelona and the word "barcelona" itself. Due to its API limit of maximum 4,000 images per query word, the resulting Flickr dataset is much smaller than the Mapillary one,

District	Images
Sant Martí	42,735
Eixample	29,720
Horta-Guinardó	22,520
Ciutat Vella	20,298
Sants-Montjuïc	18,345
Les Corts	18,053
Sarrià-Sant Gervasi	10,545
Gràcia	9,106
Sant Andreu	6,864
Nou Barris	3,944

(a) # of images.



(b) Distribution of images.



(c) Exact image locations.

Figure 3. Barcelona Mapillary crawled dataset.

and contains 17,939 images. Figure 2 gives the distribution of Flickr images in our area of interest.

Then, we created our second dataset by crawling Mapillary data posing as queries coordinate boxes within the geographical borders of the city of Barcelona. Specifically, we used as a starting point the centroid of the city, computed using the GeoJSON coordinates, and crawled the photos available within a box around the centroid. Having a limitation of 2,000 images imposed by the Mapillary API, we needed to create an algorithm that splits our original box recursively into multiple smaller ones to manage to retrieve all existing images within the municipality. In the end, we created a robotic dataset of 182,130 images, representing the city of Barcelona. The distribution of the images in this dataset are given in Figure 3.

From Figures 2 and 3, we can clearly see the differences between two datasets regarding both the distribution of images, and, most importantly, their densities. Figure 3b displays more uniformly-distributed images throughout the surface of the whole municipality of Barcelona, with a peak in Sant Martí due to a denser road infrastructure that can be seen in Figure 3c, where the image spots highlight the city roads. This contrasts with Figure 2b where we have a higher number of images concentrated into what corresponds to the most well-known places of the city as it's shown on the hot spots of Figure 2c. Both datasets and the code for generating them are publicly available at [2] and [1], respectively.

4. Experiments

Within the scope of this paper, we first ran our experiments on the smaller Flickr dataset, then we compared the results with the larger robotic Mapillary dataset, using FocalNets to discriminate the different districts of Barcelona. First, using the city's GeoJSON data, we assigned each image in the dataset to the class representing the district in which the image's GPS coordinates fall. Then, we split both datasets into training and validation sets of 90% and 10% sizes respectively.

Yang et al. [19] made available three different models: FocalNet-T (tiny), FocalNet-S (small) and FocalNet-B (base). These models differ in depth layouts and hidden dimensions. For our first experiment, we used the FocalNet-S model and we trained this network from scratch on the 17,939 Flickr images that were previously cropped to 224×224 pixels. We trained[9] the model for 50 epochs with a batch size of 32 using the default hyperparameters configuration. This includes the AdamW optimization [10] and a cosine learning rate scheduler [11], with initial learning rate $5e{-}7$ for the first 20 warm-up epochs and $5e{-}4$ for the following ones. Gradient clipping norm is set to 5.0 and the weight decay is set to 0.05.

We carried out a second experiment with the same dataset exploiting the FocalNet-T model using the same configuration as the previous one in order to compare how the two models behave with the geolocation task at hand.[10] The code for the experiments is publicly available at [1].

[9] Experiments are run on an AMD EPYC 7313P 16-Core processor with 128GB RAM and an NVIDIA GeForce RTX 3090 graphic card.

[10] We note that we left experiments with the larger FocalNet-B model as future work due to time constraints and the limited availability of the shared resources to run the experiments.

Table 1 summarizes the classification accuracy we achieved with both models. We evaluated the models on the validation set and we reached a mean accuracy of 32.8% with the FocalNet-S model, while we achieved a slightly lower value with the FocalNet-T model. If we consider, instead, the accuracy of top-5 predictions, the score goes up to about 83% for both models, a much higher score that shows the uncertainty of the model predicting exactly the best class. We also observe that the training time drops considerably for the FocalNet-T model due the model layout and less number of parameters to learn.

Considering our Flickr dataset, having a standard guess in favor of the most represented class (Sants-Montjuïc) we would have an accuracy of $\sim 28\%$. Now, comparing this score with the mean accuracy of our model (over all classes), we can argue that the model is learning, even by using this small and unbalanced dataset. These preliminary results with the small dataset encouraged us for the next experiments with the larger Mapillary dataset for Barcelona.

We carried out two different experiments using Mapillary data training both the FocalNet-T and FocalNet-S models for 50 epochs, using a batch size of 32 and the same hyperparameters as before. And as expected, we reached a much higher accuracy score using the larger dataset, as depicted in Table 1. As we saw with the smaller Flickr dataset, the smaller FocalNet-T model is as accurate as the FocalNet-S model for this task, but having the advantage of considerably smaller training time. We also give the validation cross-entropy loss and the accuracy of the top-1 predictions of all trained models in our experiments in Figure 4.

	Dataset	Accuracy top-1	Accuracy top-5	Training time
FocalNet-T	Flickr	32.1%	83.5%	38mins
FocalNet-S	Flickr	32.8%	83.1%	1h 33mins
FocalNet-T	Mapillary	68.2%	94.6%	9h 08mins
FocalNet-S	Mapillary	68.5%	94.6%	15h 27mins

Table 1. Barcelona district classification accuracies with FocalNets.

One thing to notice is that the Mapillary dataset results are much better than the Flickr ones. We argue that this could not be just because of the different sizes of the two datasets. We think that the much higher scores in the Mapillary based experiments could be due to the nature of the data itself: robotic datasets are the result of sequences

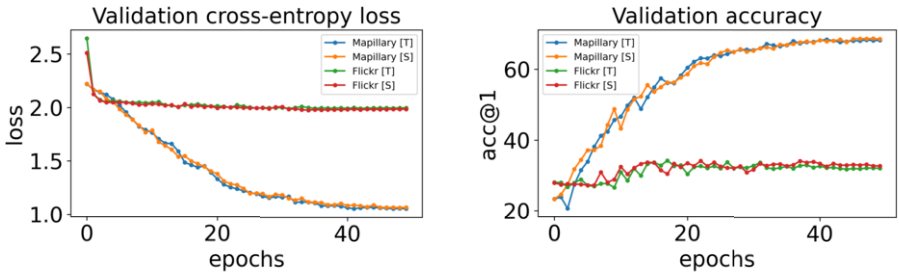
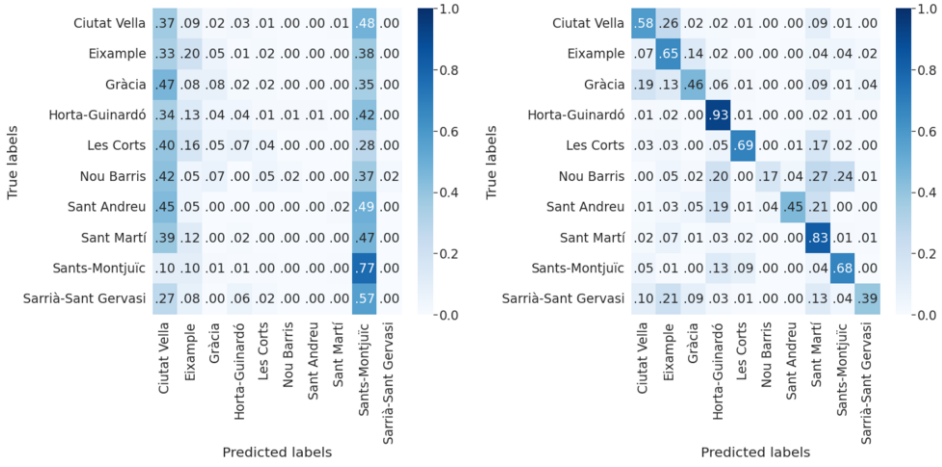


Figure 4. Validation curves on both datasets with FocalNet-T and FocalNet-S models.

of images from a consistent point of view, taken at different time spans. Therefore, some of the validation images can be relatively similar to the training ones: this dependency is inherited by how the photos have been shot in the first place. So, it could be possible that our models were trained on some images that are the previous sequences of frames used in our validation set to evaluate the models. In spite of everything, as we can see in Figure 4, the Mapillary validation losses and accuracies are likely to improve: with a higher number of epochs we could reach better scores.

Additionally, we generated confusion matrices over the validation data in order to compare the strengths and weaknesses of the geolocator obtained on both datasets. From Figure 5a, we can conclude that the geolocator for Flickr is prone to classify the images in a skewed distribution in favor of the two most represented classes. This contrasts with Figure 5b, where we observe that the classification process with the Mapillary dataset is quite robust with the exception of a minority of the classes.



(a) Flickr dataset (b) Mapillary dataset

Figure 5. Confusion matrices on validation data with FocalNet-T models.

Our experiments show that, from a practical point of view, the classification using the Flickr dataset is not reliable enough; this could be because of the nature of the images that make up the dataset or as a result of the small dataset size and unbalancedness of the classes. We regard these first results as a motivation for further experimentation with non-robotic datasets.

5. Conclusions and future work

The motivation for our research regarding this paper is the automatic geolocation of images taken from disaster areas. In Disaster Management, timely spotting of the places affected by a calamity has a significant role in redirecting emergency help. To this end, correctly geolocating images taken from the affected area using an automated software pipeline could be of great aid to first responders and could facilitate the rescue of people subject to natural disasters. The replicability of such a pipeline in new disaster areas

could allow these emergencies to be managed with the least effort. For this purpose, by developing an automatic geolocator for a given area of interest, (e.g., a city, a region) we are providing a tool that could add impactful help even into dramatic situations.

As a first step to building a geolocation pipeline, we implemented a crawler for two major online image sources for cities, namely Flickr and Mapillary. As the second step, we chose a novel classification model, Focal Modulation Networks (FocalNets) [19], that outperforms the state-of-the-art and requires relatively short training time. We used FocalNets within the context, first, of a small non-robotic dataset and, second, of a larger robotic dataset.

We regard the uncertainty in predicting the best class, especially in the case of non-robotic images, as a symptom of the need for further investigation. The first thing we have in mind is to experiment with equally-sized non-robotic and robotic datasets to scrutinize the discrepancy between the first results with these two datasets. Second step will be to incorporate a k-fold cross validation in our pipeline. This stage would give us a better insight into the generalization capabilities of our models. A later step will be to use both our datasets in a hybrid fashion to train our geolocators in order to see their classification performance with mixed image types. Having success with such a hybrid dataset will give great value to the abundant robotic imagery available online ($\sim$1.5 billion street-level images available from all around the world within Mapillary) and would bring about great opportunities to leverage everyday-growing pieces of information that social media provide us with. This would mean that, with our pipeline, scraping the Mapillary and Flickr APIs for a new city and training the model with this new data will provide us with the opportunity to discriminate between the districts of a city of interest in relatively short amount of time. Moreover, the usage of data gathered from other social media platforms (e.g., Twitter) could be very insightful to test how the geolocator responds to disparate variety of images extracted from real-world emergencies. Thinking about Disaster Management, when people constantly post new images showing a different perspective of the after-effects, nearly real-time geolocation could be of great aid, even for cities never analyzed before.

The code for our pipeline and the experiments, and the datasets are publicly available at [1] and [2], respectively.

Acknowledgements This work was partially funded by the EU H2020 project Crowd4SDG "Citizen Science for Monitoring Climate Impacts and Achieving Climate Resilience", #872944. Fabio Murgese is an MSc student of the Computer Science department at the Università di Pisa and Gerard Alcaina is an MSc student of the Mathematics department at the Universitat Autònoma de Barcelona.

References

[1] Artificial Intelligence Research Institute (IIIA), CSIC. Code for image geolocalization. https://github.com/IIIA-ML/geoloc, 2022. [Last accessed 30-May-2022].

[2] Artificial Intelligence Research Institute (IIIA), CSIC. Datasets for image geolocalization. https://github.com/IIIA-ML/geoloc-data, 2022. [Last accessed 30-May-2022].

[3] Yannis Avrithis, Yannis Kalantidis, Giorgos Tolias, and Evaggelos Spyrou. Retrieving landmark and non-landmark images from community photo collections. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 153–162, 2010.

[4] Sara Barozzi, Jose Luis Fernandez Marquez, Amudha Ravi Shankar, Barbara Pernici, et al. Filtering images extracted from social media in the response phase of emergency events. In *16th Conference on Information Systems for Crisis Response and Management*, pages 1–12, 2019.

[5] Clemens Havas, Bernd Resch, Chiara Francalanci, Barbara Pernici, Gabriele Scalia, Jose Luis Fernandez-Marquez, Tim Van Achte, Gunter Zeug, Maria Rosa (Rosy) Mondardini, Domenico Grandoni, Birgit Kirsch, Milan Kalas, Valerio Lorini, and Stefan Rüping. E2mc: Improving emergency management service practice through social media and crowdsourcing analysis in near real time. *Sensors*, 17(12):2766, 2017.

[6] 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. IEEE, 2008.

[7] Yunpeng Li, Noah Snavely, Dan Huttenlocher, and Pascal Fua. Worldwide pose estimation using 3d point clouds. In *European conference on computer vision*, pages 15–29. Springer, 2012.

[8] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[9] Luca Loria. Hierarchical classification model for content-based geolocation of outdoor images with visual explanations. Master's thesis, Politecnico di Milano, 2021.

[10] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[12] Carlo Masone and Barbara Caputo. A survey on deep visual place recognition. *IEEE Access*, 9:19516–19547, 2021.

[13] Eric Müller-Budack, Kader Pustu-Iren, and Ralph Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. In *Computer Vision – ECCV 2018*, pages 575–592, Cham, 2018. Springer International Publishing.

[14] Till Quack, Bastian Leibe, and Luc Van Gool. World-scale mining of objects and events from community photo collections. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 47–56, 2008.

[15] Amudha Ravi Shankar, Jose Luis Fernandez-Marquez, Barbara Pernici, Gabriele Scalia, Maria Rosa Mondardini, and Giovanna Di Marzo Serugendo. Crowd4ems: A crowdsourcing platform for gathering and geolocating social media content in disaster response. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:331–340, 2019.

[16] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Improving image-based localization by active correspondence search. In *European conference on computer vision*, pages 752–765. Springer, 2012.

[17] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[18] Tobias Weyand, Ilya Kostrikov, and James Philbin. PlaNet - photo geolocation with convolutional neural networks. http://arxiv.org/abs/1602.05314.

[19] Jianwei Yang, Chunyuan Li, and Jianfeng Gao. Focal modulation networks. https://arxiv.org/abs/2203.11926, 2022.