Where in the World is this Image? Transformer-based Geo-localization in the Wild (Supplementary Material)

Shraman Pramanick[1], Ewa M. Nowara[1], Joshua Gleason[2],
Carlos D. Castillo[1], and Rama Chellappa[1]

[1] Johns Hopkins University
[2] University of Maryland, College Park
{spraman3,carlosdc,rchella4}@jhu.edu,
ewa.m.nowara@gmail.com,gleason@umd.edu

In this supplementary material, we provide additional details on geo-cell partitioning, data augmentation, hyper-parameter values, baselines, evaluation metrics and illustrate additional quantitative and qualitative results.

A Implementation Details & Hyper-parameter Values

A.1 Adaptive Geo-cell Partitioning

We utilize the *S*2 *geometry library*[3] to divide the earth's surface into a fixed number of non-overlapping geographic cells. To directly compare our results with baselines, we use the same partitioning approach like [6], where we subdivide the earth's surface into three resolutions containing 3298, 7202, and 12893 geo-cells referred to as coarse, middle, and fine cells, respectively. The partitioning ensures each cell contains at least 50 and at most 5000, 2000, and 1000 training images for the coarse, middle, and fine resolution. Limiting the number of training images into a minimum and maximum range per geo-cell gives two advantages. First, the training set does not suffer from class imbalance, which is pivotal for classifying many classes. Second, the geographic areas which are heavily photographed are subdivided into smaller cells, allowing more precise geo-localization of these regions (such as big cities and tourist attractions). However, one drawback of this approach is that many geographic areas have less than the required minimum number of images. Consequently, many locations (such as oceans, remote mountainous regions, deserts, poles) are discarded because of insufficient images. With the minimum range of 50, the partitioning covers almost 84% of the entire earth's surface.

In classification-based geo-localization, the number of geo-cells closely relates to the prediction accuracy. In other words, since the predicted GPS coordinates are always the mean location of all training images in the predicted geo-cell, coarse cells often can not produce good street-level accuracy. On the other hand, fine cells improve localization precision by generating smaller geo-cells in highly photographed areas. Figure A.1 shows the improvement of street-level (1 km) and city-level (25 km) geolocational accuracy by using the predictions from finer cells on four different datasets. Since the continent-level (2500 km) accuracy is not directly related to the size of cells, it remains almost unchanged with geo-cell resolution variation. Next, we use an ensemble of hierarchical classification

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

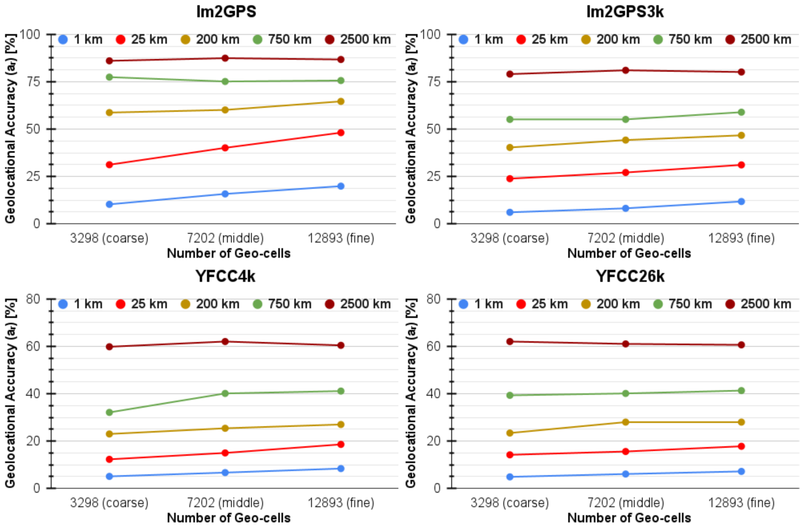

Fig. A.1: **Effect of the predictions from three different partitioning schemes on the performance of TransLocator.** The street-level (1 km) and city-level (25 km) geolocational accuracy significantly improves as we employ finer geo-cells. However, the continent-level (2500 km) accuracy remains almost unchanged.

using all three resolutions. However, in agreement with [10], this method does not achieve a consistent improvement than considering only fine partitioning. Moreover, the ensemble increases inference time by almost 9%. Following these observations, we use the predictions of the fine geo-cells in all our experiments. We believe that adding a retrieval network after classification would improve the performance by allowing the system to search within the predicted geo-cell. However, this paper does not consider any retrieval extensions for a fair comparison with the baselines.

A.2 Data Augmentation

Since the training set contains images in various orientations, resolutions and scales, we use extensive data augmentation. The augmentation policy includes: *RandomAffine* with degrees $(0, 15)$, *ColorJitter* containing $\{$brightness, contrast, saturation, hue$\}$ strength of $\{0.4, 0.4, 0.4, 0.1\}$ with a probability of 0.8, *RandomHorizontalFlip* with a probability of 0.5, *Resize* in $(256, 256)$, *Tencrop* with size $(224, 224)$ and standard *Normalization*. We apply *ColorJitter* only in the RGB channel. Table A.1 shows an empirical analysis of the effectiveness of different augmentation techniques on training TransLocator. We start with the standard *Flip*, *Resize* and *Normalization* operations. Adding *Affine* transformation and *ColorJitter* helps in improving the performance by a tiny margin. However, the *TenCrop* augmentation shows to have a significant effect, improving $0.5 - 1.1\%$ street-level accuracy in Im2GPS and Im2GPS3k datasets. Since the important visual cues for geo-localization often reside on the edges of the image, taking multiple crops from different positions and averaging the predictions helps in improving the performance.

Table A.1: **Role of different data augmentation techniques on training TransLocator.** RHF, R, N, RA, CJ and TC denotes *RandomHorizontalFlip*, *Resize*, *Normalization*, *RandomAffine*, *ColorJitter* and, *Tencrop*, respectively, using the parameters mentioned in Section A.2.

Dataset	**Method**	**Distance (a_r [%] @ km)**				
		Street	**City**	**Region**	**Country**	**Continent**
		1 **km**	25 **km**	200 **km**	750 **km**	2500 **km**
Im2GPS [2]	RHF + R + N	18.8	46.2	62.8	73.6	83.6
	+ RA	18.8	46.5	63.1	73.8	83.8
	+ RA + CJ	19.0	46.8	63.2	74.1	84.0
	+ RA + CJ + TC	**19.9**	**48.1**	**64.6**	**75.6**	**86.7**
Im2GPS 3k [2]	RHF + R + N	11.3	30.4	45.7	58.0	78.4
	+ RA	11.4	30.6	46.0	58.0	78.5
	+ RA + CJ	11.4	30.8	45.9	58.2	78.7
	+ RA + CJ + TC	**11.8**	**31.1**	**46.7**	**58.9**	**80.1**

Table A.2: **Hyper-parameters of TransLocator.**

Hyper-parameters	**Notation**	**Value**
#dim for dense layers in MFF	-	$[768, 8, 1]$
#dim for classification FC	coarse	$[768, 3298]$
	middle	$[768, 3000, 7202]$
	fine	$[768, 6000, 12893]$
	scene	$[768, 3/16/365]$
Training		
Batch-size	-	256
Epochs	N	40
Optimizer	-	AdamW
Loss	-	CE
Base learning rate	α	0.1
Momentum	-	0.9
Learning rate scheduler	-	Cosine
Warmup epochs	-	2
Weight decay	-	0.0001

A.3 Hyper-parameter Details

In Table A.2, we furnish the details of hyper-parameters used during training. Grid search is performed on batch size, learning rate, and the depth of classifier heads to find the best hyper-parameter configuration. The model is evaluated after every epoch on the validation set and the best model was taken to be evaluated on the test set. We use AdamW [5] optimizer with cosine learning rate scheduler and fixed number of warmup steps for optimization without gradient clipping.

B Baselines

In this section, we provide additional details about the baseline methods. Very few approaches in the literature have attempted to geo-locate images on a scale of an entire world without any restrictions. In the last 5 years, CNNs trained with large-scale datasets have significantly improved the planet-scale geo-localization performance. To the best of our knowledge, we are the first to introduce the effectiveness of fusion transformer architecture for this ill-posed problem. We compare our method with the following baselines:

- **Im2GPS** [1] is the first to attempt planet-scale geo-localization by using a simple retrieval approach to match a given query image based on a combination of different hand-crafted image descriptors to a reference dataset containing more than 6M GPS-tagged images. This approach has later been improved [2] by refining the search with multi-class support vector machines.
- **PlaNet** [12] is the first deep neural network trained for unconstrained planet-scale geo-localization. More specifically, PlaNet divides the earth in 26263 geo-cells and trains an inception [9] network with batch normalization [3] using 91M geo-tagged images. PlaNet outperforms both versions of Im2GPS [1, 2] by a substantial margin.
- **[L]kNN** [11] proposes a retrieval-based geo-localization system which combines the Im2GPS and PlaNet by using features extracted by CNNs for nearest neighbour search. Though this method uses a 5-times smaller training set than PlaNet, the retrieval-based approach requires a substantially larger inference time and disk space than classification approaches.
- **CPlaNet** [8] develops a combinatorial partitioning algorithm to generate a large number of fine-grained output classes by intersecting multiple coarse-grained partitionings of the earth. This technique allows creating small geo-cells while maintaining sufficient training examples per cell and hence improves the street- and city-level geolocational accuracy by a large margin.
- **MvMF** [4] introduces 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.
- **ISNs** [6] reduces the complexity of planet-scale geo-localization problem by leveraging contextual knowledge about environmental scenes. To deal with the huge diversity of images on earth's surface, this approach trains three different ResNet101 networks for *natural*, *urban*, and *indoor* scenes, and achieves the current state-of-the-art performance on Im2GPS and Im2GPS3k. However, training different networks is cost-prohibitive and can not be generalized to a larger number of scenes. Our work addresses the limitations of ISNs by training a unified dual-branch transformer network in a multi-task framework and improves the state-of-the-art results by a significant margin.

C Evaluation Metrics

In a classification setup, we train TransLocator using cross-entropy loss which is closely associated with the classification accuracy. However, we evaluate TransLocator using geolocational accuracy. Hence, we empirically verify the strong correlation between these two metrics. We consider the Top-N classification accuracy for 8 different N values (1, 5, 10, 50, 100, 200, 300, 500) and geolocational accuracy (a_r) for 8 different r values (1, 25, 100, 200, 400, 750, 1500, 2500) in similar intervals, and observe the correlation among them on different evaluation sets. The Pearson correlation coefficients between the two metrics for TransLocator are 0.978, 0.984, 0.985 and 0.982 on Im2GPS, Im2GPS3k, YFCC4k and YFCC26k, respectively. We also observe a similarly high correlation for the ViT-MT model. Figure 4 in the main paper and Figure C.1 in supplementary material illustrates the strong linear correlation between the two metrics for TransLocator and ViT-MT on all 4 evaluation sets.

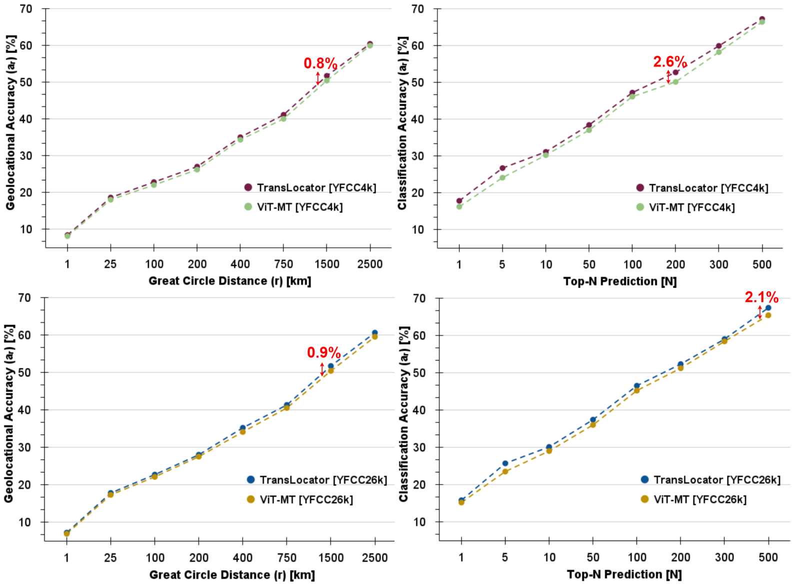

Fig. C.1: **Strong linear relationship between the geolocational and classification accuracy metrics.** The positive correlation enables us to treat geo-localization as a classification problem.

D Additional Quantitative Results

In this section, we investigate the contribution of different loss functions and provide a direct comparison of TransLocator with the ISNs [6].

D.1 Ablation Study on Training Objective

As discussed in section 3.4 of the main paper, our overall training objective contains three losses for geo-cell prediction and one for scene recognition. Table D.1 shows how each loss function contributes to the performance of TransLocator on Im2GPS and Im2GPS3k datasets. We begin with training TransLocator separately on coarse, middle, and fine geo-cells. As the size of the geo-cells reduces, the geolocational accuracy with a smaller distance threshold typically improves. With the fine geo-cells, TransLocator gains 5 - 8.3% street-level accuracy than using the coarse cells. Combining all three different geo-cells helps the system learn geographical features at different scales, leading to a more discriminative classifier and improving the street-level accuracy by 0.2 - 0.9%. The performance is further improved by another 0.7 - 0.9% after adding the scene information, which reduces the complexity of the data space by providing contextual knowledge about the surroundings.

D.2 Differences from ISNs

As the Individual Scene Networks (ISNs) [6] is the first method that utilizes scene information for geo-localization, we present clear differences of our method from ISNs.

Table D.1: **Ablation study on different losses of the training objective of TransLocator.** The finer geo-cells helps to improve the geolocational accuracy with smaller distance threshold.

Dataset	$\mathcal{L}_{geo}^{corase}$	$\mathcal{L}_{geo}^{middle}$	$\mathcal{L}_{geo}^{fine}$	$\mathcal{L}_{scene}$	**Distance** (a_r **[%] @ km)**				
					Street 1 **km**	**City** 25 **km**	**Region** 200 **km**	**Country** 750 **km**	**Continent** 2500 **km**
Im2GPS [2]	✓	✗	✗	✗	9.8	30.9	54.4	72.5	84.7
	✗	✓	✗	✗	14.0	38.4	58.5	72.9	**86.7**
	✗	✗	✓	✗	18.1	46.0	61.8	73.1	85.7
	✓	✓	✓	✗	19.0	47.2	62.7	73.5	85.7
	✓	✓	✓	✓	**19.9**	**48.1**	**64.6**	**75.6**	**86.7**
Im2GPS 3k [2]	✓	✗	✗	✗	5.9	22.7	40.7	54.9	77.0
	✗	✓	✗	✗	8.0	25.5	42.8	56.9	78.1
	✗	✗	✓	✗	10.9	29.8	45.0	56.8	78.1
	✓	✓	✓	✗	11.1	30.2	45.0	56.8	78.1
	✓	✓	✓	✓	**11.8**	**31.1**	**46.7**	**58.9**	**80.1**

Table D.2: **Comparison of unified and separate systems for different scene kinds.** Using separate systems is not only cost-prohibitive but also does not utilize semantic similarities across different scene kinds.

Dataset	**Method**	**Train Set {Scene}**	**Evaluation Set {Scene} (Distance** (a_r **[%] @ km)**								
			Natural			**Urban**			**Indoor**		
			1 **km**	200 **km**	2500 **km**	1 **km**	200 **km**	2500 **km**	1 **km**	200 **km**	2500 **km**
Im2GPS [1]	ISNs (M, f, S_3) [6]	Natural	2.5	48.8	71.3	–	–	–	–	–	–
		Urban	–	–	–	22.6	56.5	89.9	–	–	–
		Indoor	–	–	–	–	–	–	15.8	31.6	57.9
	TransLocator w/o scene	Natural	3.8	54.9	77.1	–	–	–	–	–	–
		Urban	–	–	–	24.8	65.0	88.2	–	–	–
		Indoor	–	–	–	–	–	–	20.4	55.2	79.0
	TransLocator	All	5.0	60.8	83.8	27.5	72.9	86.9	26.3	61.4	94.7
Im2GPS 3k [2]	ISNs (M, f, S_3) [6]	Natural	3.2	31.8	63.1	–	–	–	–	–	–
		Urban	–	–	–	14.1	44.8	72.7	–	–	–
		Indoor	–	–	–	–	–	–	9.2	17.8	48.4
	TransLocator w/o scene	Natural	4.0	36.5	70.8	–	–	–	–	–	–
		Urban	–	–	–	14.4	45.9	80.2	–	–	–
		Indoor	–	–	–	–	–	–	10.4	28.2	68.5
	TransLocator	All	4.7	42.6	75.2	15.0	45.9	83.3	13.0	32.7	78.2

First, ISNs train three separate ResNet101 networks for natural, urban and indoor scenes. In contrast, TransLocator uses a unified dual-branch transformer backbone for all scenes. Using three different networks is cost-prohibitive and restricts the system from sharing the learned features across different scene kinds that likely have higher-order semantic similarities. To directly comprehend the effectiveness of a unified network, we train TransLocator without its scene recognition head separately on natural, urban and indoor images. As shown in Table D.2, the single network achieves better performance than separate networks for all three scene kinds. Table D.2 also exhibits the effectiveness of dual-branch transformer backbone than ResNet for geo-localization. Moreover, unlike ISNs, the segmentation branch of TransLocator helps produce better qualitative performance under challenging real-world appearance variation.

E Additional Qualitative Results

In this section, we visualize a few example images from the Im2GPS and Im2GPS3k datasets localized within 1 km, 200 km, and 2500 km from ground truth locations by

TransLocator in Figure E.1. The corresponding Grad-CAM [7] activation maps highlight the necessary pixels used for the decision. Famous landmarks and tourist attractions like the Washington Monument, Eiffel Tower, Niagara Falls, and Trafalgar Square are correctly localized. Moreover, TransLocator often yields surprisingly accurate results for images with more subtle geographical cues, like a sea-beach in Venice or a uniquely-shaped building in Seoul. Minor errors like 100 – 200 meters occur due to fine cells' size and can further be removed by using a retrieval extension. More difficult samples, like a forest in Tanzania and a desert in Utah, are localized within a range of 10-20 km, which can be attributed to the bigger geo-cells in those sparsely-populated areas. TransLocator can learn to recognize famous streets, buildings, water-bodies, plants, animals and yields surprisingly good results even in remote locations.

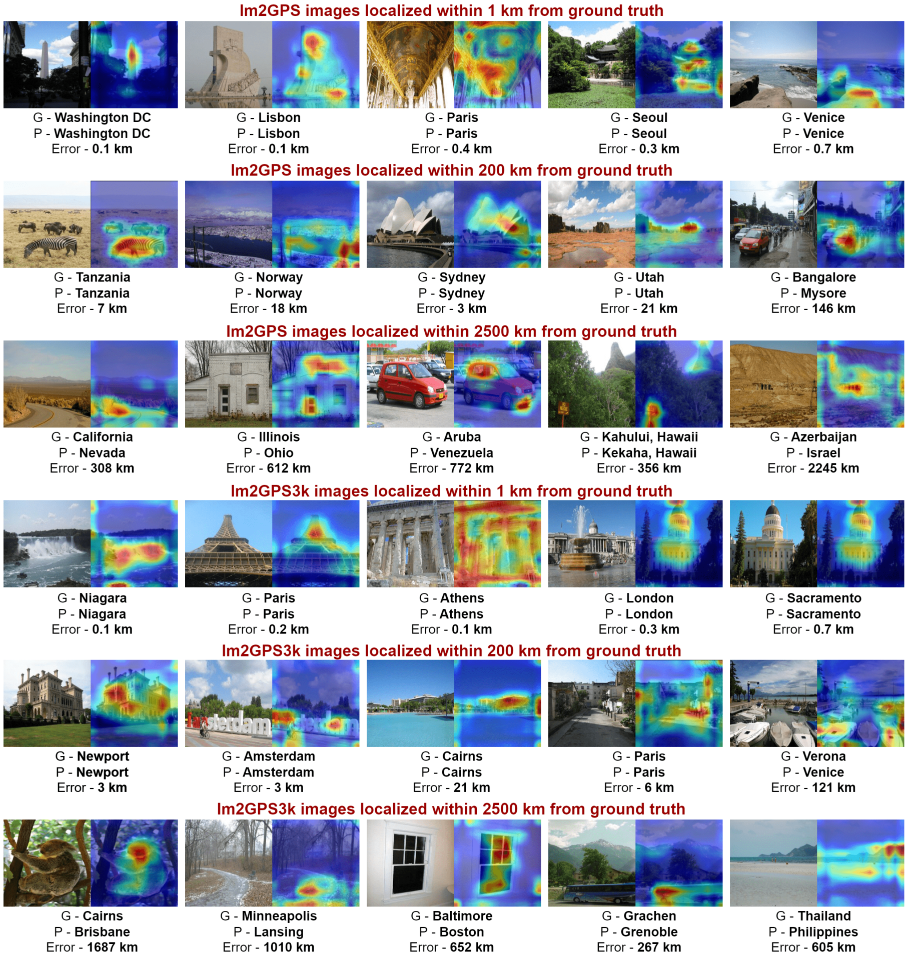

Fig. E.1: **Example images from Im2GPS and Im2GPS3k dataset localized within three different distance threshold by TransLocator.** The corresponding Grad-CAM [7] activation maps highlights the most important pixels used for prediction.

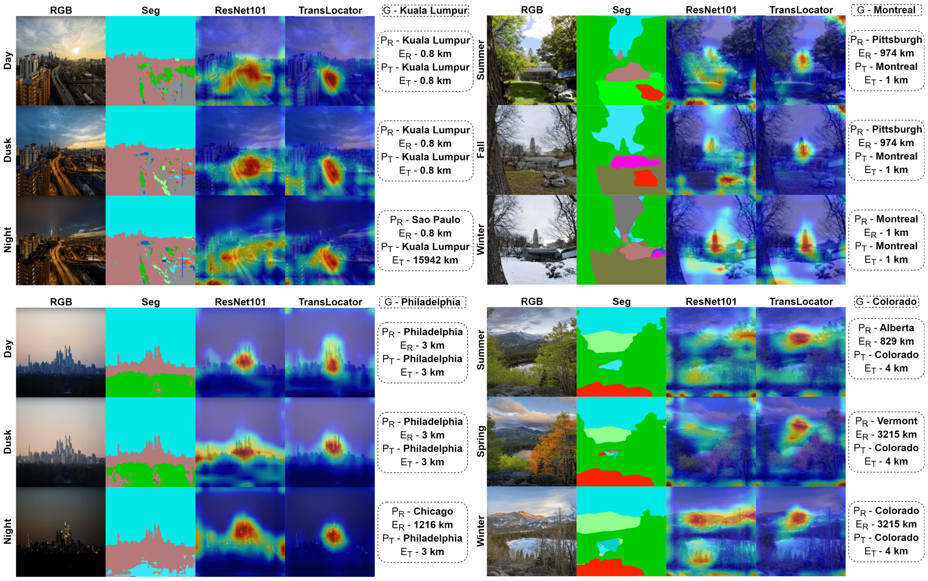

Fig. E.2: **Qualitative comparison of TransLocator and ResNet101 on images with same location but under challenging appearance variations.** Unlike ResNet101, TransLocator attends to similar regions in each image and locates them correctly. G denotes ground truth, P_R, E_R, P_T and E_T denotes predicted location and prediction error by ResNet101 and TransLocator, respectively. Best viewed when zoomed in and in color.

Next, we illustrate a few more cases[4] of drastic appearance variation in the same location depending on the time of the day, weather, or season in Figure E.2. Though the RGB images experience extreme variation, the corresponding semantic segmentation maps remain unchanged. Thus, TransLocator can learn robust features and produce consistent activation maps across such radical appearance changes. In contrast, ResNet101 fails to recognize such variation.

References

1. Hays, J., Efros, A.A.: Im2gps: estimating geographic information from a single image. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008)
2. Hays, J., Efros, A.A.: Large-scale image geolocalization. In: Multimodal Location Estimation of Videos and Images, pp. 41–62. Springer (2015)
3. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. pp. 448–456. PMLR (2015)
4. Izbicki, M., Papalexakis, E.E., Tsotras, V.J.: Exploiting the earth's spherical geometry to geolocate images. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 3–19. Springer (2019)

[4] Collected from the Internet under creative commons license.

5. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
6. Muller-Budack, E., Pustu-Iren, K., Ewerth, R.: Geolocation estimation of photos using a hierarchical model and scene classification. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 563–579 (2018)
7. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626 (2017)
8. Seo, P.H., Weyand, T., Sim, J., Han, B.: Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 536–551 (2018)
9. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9 (2015)
10. Theiner, J., Müller-Budack, E., Ewerth, R.: Interpretable semantic photo geolocation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 750–760 (2022)
11. Vo, N., Jacobs, N., Hays, J.: Revisiting im2gps in the deep learning era. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2621–2630 (2017)
12. Weyand, T., Kostrikov, I., Philbin, J.: Planet-photo geolocation with convolutional neural networks. In: European Conference on Computer Vision. pp. 37–55. Springer (2016)