Where in the World is this Image? Transformer-based Geo-localization in the Wild

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

[1]*Johns Hopkins University,* [2]*University of Maryland, College Park*

ECCV TEL AVIV 2022

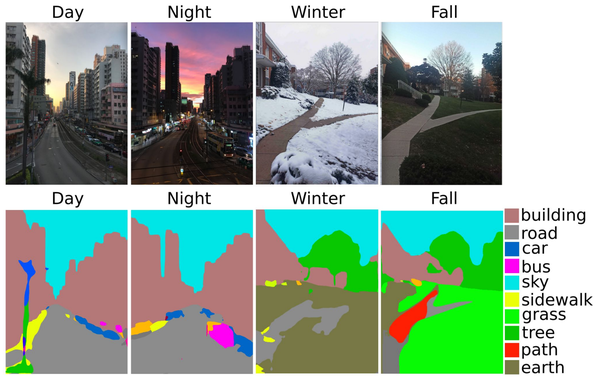
Problem Statement:

- Geo-localization is the task of predicting geographic location from images.
- The goal of this work is planet-scale geo-localization from a **single image.**
- Necessary to detect fine-grained cues present in small regions of the image.

Challenges:

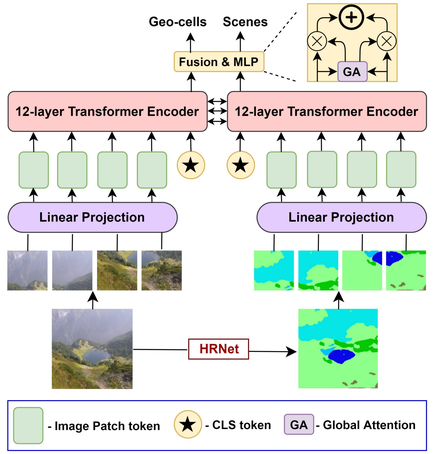
Our Approach (*TransLocator*):

- **Vision Transformer (ViT)** encodes global information and models long-range dependencies across different patches in an image.
- **Semantic Segmentation** is robust to drastic appearance variations of the same location in different daytime or weather.
- **Unified Multi-Task Learning** is able geo-locate and predict the scene type (i.e., natural, urban, indoor) to better learn scene-specific features.



Our Model (*TransLocator*):

- **We treat geo-localization as a classification task.**
- Vision transformer splits images into an ordered sequence of patches, which are then projected and fed into the network.
- Semantic segmentation maps (obtained with HRNet) are fed along with RGB images to a **dual-branch** vision transformer. The two branches interact after every transformer layer.
- Multi-task learning objective simultaneously predicts the **geo-cell and the scene in the image**.



Main Experimental Results:

- SOTA results on public geo-localization datasets with the following significant performance improvements - **Im2GPS: 5.5%, Im2GPS3k: 14.1%, YFCC4k: 4.9%, YFCC26k: 9.9%.**

Results on Im2GPS3k dataset:

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
[L]kNN	7.2	19.4	26.9	38.9	55.9
PlaNet	8.5	24.8	34.3	48.4	64.6
CPlaNet	10.2	26.5	34.6	48.6	64.6
INSs	10.1	27.2	36.2	49.3	65.6
INSs	10.5	28.	36.6	49.7	66.0
ViT-MT	11.0	29.0	42.6	54.8	71.6
TransLocator	**11.8**	**31.1**	**46.7**	**58.9**	**80.1**

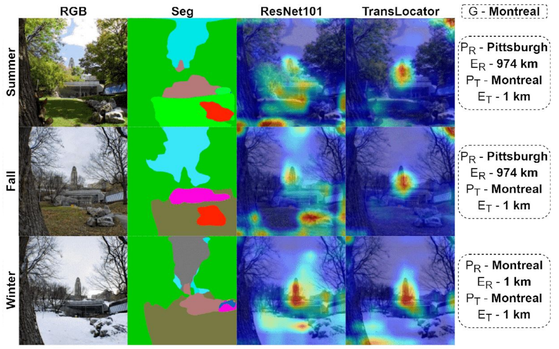
Ablation Experiments:

- **Segmentation maps, multi-modal feature fusion** and **multi-task learning** significantly improves performance.

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
ResNet101	9.0	25.1	32.8	46.1	63.5
EfficientnET-b4	9.2	26.8	32.7	47.0	63.9
ViT Base	9.9	28.0	37.8	54.2	70.7
+ Seg	10.5	29.1	42.5	55.8	73.6
+ Seg + MFF	11.1	30.2	45.0	56.8	78.1
+ Seg + MFF + Scene	**11.8**	**31.1**	**46.7**	**58.9**	**80.1**

Qualitative Results:

- We compare *TransLocator* with *ResNet101* on images from the same location under challenging appearance variations.



References:

- Raghu, M. et al.; Do vision transformers see like convolutional neural networks?
- Muller-Budack, E. et al.; Geolocation estimation of photos using a hierarchical model and scene classification.