A Multi-Stage Multi-Task Neural Network for Aerial Scene Interpretation and Geolocalization

Alina Marcu[1,2] Dragoş Costea[2,3] Emil Sluşanschi[3] Marius Leordeanu[1,2,3]

[1] Institute of Mathematics of the Romanian Academy [2] Autonomous Systems [3] University Politehnica of Bucharest

Introduction

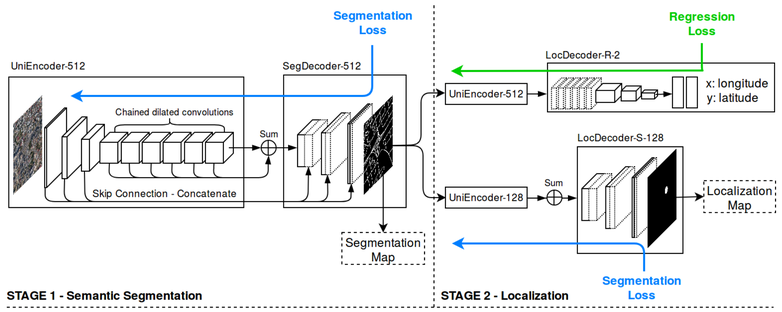
- The ability to interpret a scene and pinpoint its location is of growing interest in the domain of aerial images.
- We propose a novel multi-stage multi-task neural network that is able to handle segmentation and localization at the same time, in a single forward pass.

Multi-Stage Multi-Task Architecture

- Our proposed architecture follows a modular stage-wide strategy:

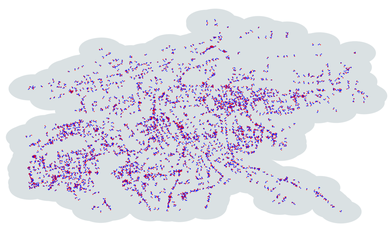
  Stage 1 is designed for semantic segmentation. Our network predicts pixelwise class labels. We argue that roads can be used as a unique footprint of an urban area, therefore we train MSMT-Stage-1 for road detection.

  Stage 2 provides a precise location using two branches. One branch uses a regression network, while the other is used to predict a location map trained as a segmentation task.



- Our architecture uses encoder-decoder modules at each stage, having the same encoder structure re-used.
- LocDecoder-R-2 predicts location as two real valued numbers for longitude and latitude. LocDecoder-S-128 predicts a localization map of size 128x128 on the whole area of possible locations. White pixels denote probable locations of the input image.

Our Localization Dataset



We collected 9531 512 x 512 pixels^2 images randomly chosen within a 100x100 m^2 square area around any intersection, covering in total an European urban area of around 70 km^2.

- The figure portraits the data distribution of our aerial image localization dataset. Each grey disk depicts a region of 500 meters radius around the training (blue centers) and testing (red centers) data.

More details about our work: https://sites.google.com/site/aerialimageunderstanding/

Semantic Segmentation on INRIA dataset

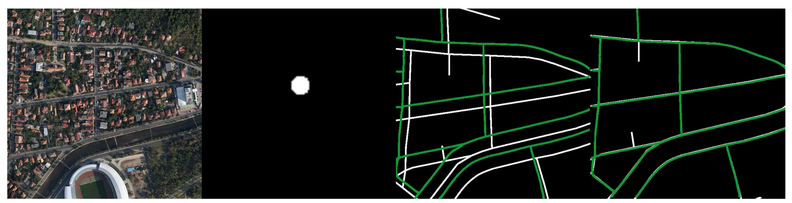
- For the task of semantic segmentation in aerial images, we report state-of-the-art results on the publicly available Inria dataset [3].
- The training set contains 180 color image tiles of size 5000×5000, covering a surface of 1500×1500 square meters each (at a 30 cm spatial resolution).
- Figure below presents from left to right, in order, the RGB input image, the prediction of our MSMT-Stage-1 model and the ground truth.

- The dataset covers various regions such as Austin (1), Chicago (2), Kitsap County (3), West Tyrol (4), Vienna (5). We report pixelwise results on each region set as well as Overall (6) Accuracy and IoU.

Method		(1)	(2)	(3)	(4)	(5)	(6)
MLP [3]	IoU	61.20	61.30	51.50	57.95	72.13	64.67
	Acc.	94.20	90.43	98.92	96.66	91.87	94.42
Mask R-CNN [2]	IoU	65.63	48.07	54.38	70.84	64.40	59.53
	Acc.	94.09	85.56	97.32	**98.14**	87.40	92.49
SegNet MTL [1]	IoU	**76.76**	67.06	**73.30**	66.91	76.68	73.00
	Acc.	93.21	**99.25**	97.84	91.71	**96.61**	95.73
MSMT-Stage-1	IoU	75.39	**67.93**	66.35	**74.07**	**77.12**	**73.31**
	Acc.	**95.99**	92.02	**99.24**	97.78	92.49	**96.06**

Localization and Alignment

- We apply a refinement step after geolocalization.
- The original error was among the highest using the segmentation method (40.38m), down to 0.32m after alignment.
- For MSMT with LocDecoder-S-128 averages are computed only for the 92.7% of cases when it does not leave the localization map blank.
- Figure depicts the RGB input image, dot segmentation generated by MSMT LocDecoder-S-128, segmented roads (green) at the predicted location on top of ground truth roads (white), before and after alignment.

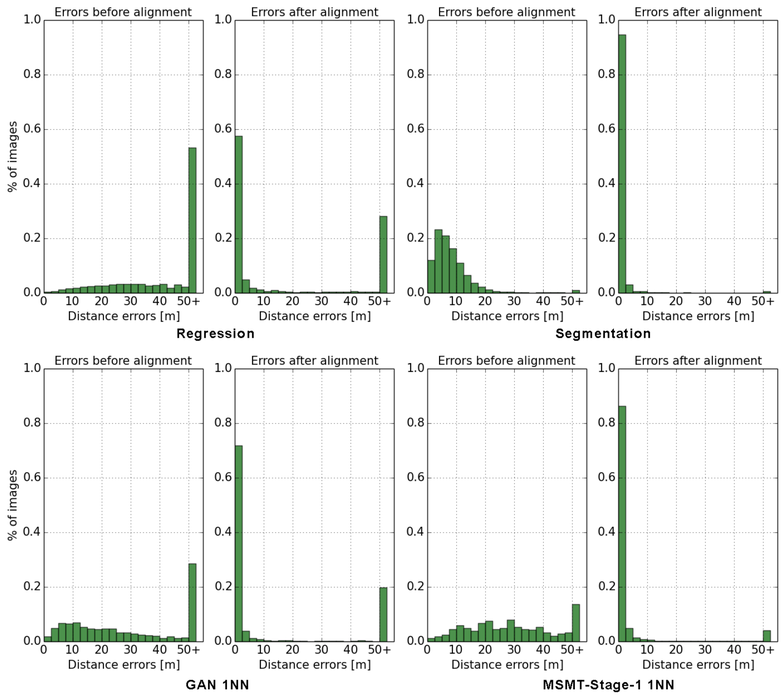


Localization error analysis

- For MSMT with LocCombined, we combined the two localization branches: while the segmentation net is on average much more precise, the regression head is used if the segmentation produces a blank map.

Method	Before		After	
	Mean	Median	Mean	Median
MSMT with LocDecoder-R-2	88.92	53.90	57.97	1.60
MSMT with LocDecoder-S-128	**9.03**	**6.85**	**1.89**	**0.75**
MSMT with LocCombined	26.93	7.27	18.42	0.78

- Using our segmentation method, 96.84% of test locations have an error of less than 20m without alignment.
- After alignment, 94.56% of the test locations are within 2.5m of the ground truth location and 97.58% are within 5 meters, which matches an approximate figure for a commercial GPS.



References

[1] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel. Multi-task learning for segmentation of building footprints with deep neural networks. *arXiv preprint arXiv:1709.05932*, 2017.

[2] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

[3] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *IEEE International Symposium on Geoscience and Remote Sensing (IGARSS)*, 2017.