Learning to Align Semantic Segmentation and 2.5D Maps for Geolocalization

Anil Armagan, Martin Hirzer, Peter M. Roth, Vincent Lepetit
Institute of Computer Graphics and Vision, Graz University of Technology
{armagan, hirzer, roth, lepetit}@icg.tugraz.at

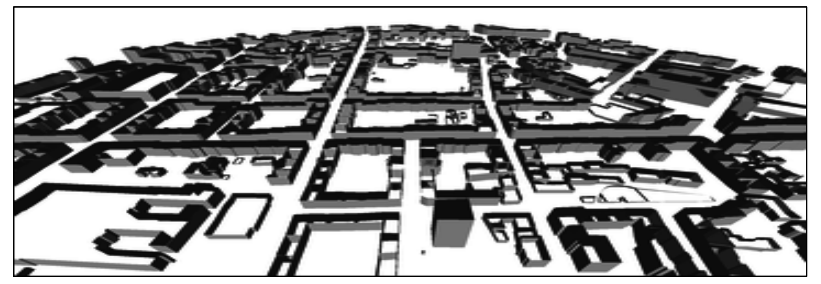
IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR 2017, July 21-26, Honolulu

Contribution

We present an efficient method for geolocalization in urban environments starting from a coarse pose estimate provided by GPS and compass information and using a simple **untextured 3D model** of the surrounding buildings. We train two deep networks to predict the best direction to improve the pose estimate, given a semantic segmentation of the input image and a rendering of the buildings from this estimate. We then iteratively apply these networks until converging to a good pose.

2.5D Maps

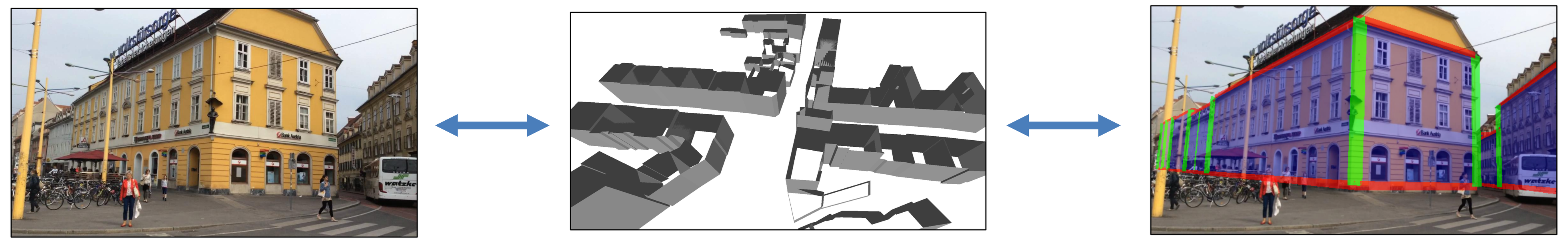
- Light weight 3D models
- Ground plane locations of buildings' corners
- Height of each building
- No texture or color
- Available on OpenStreetMap



2.5D Map

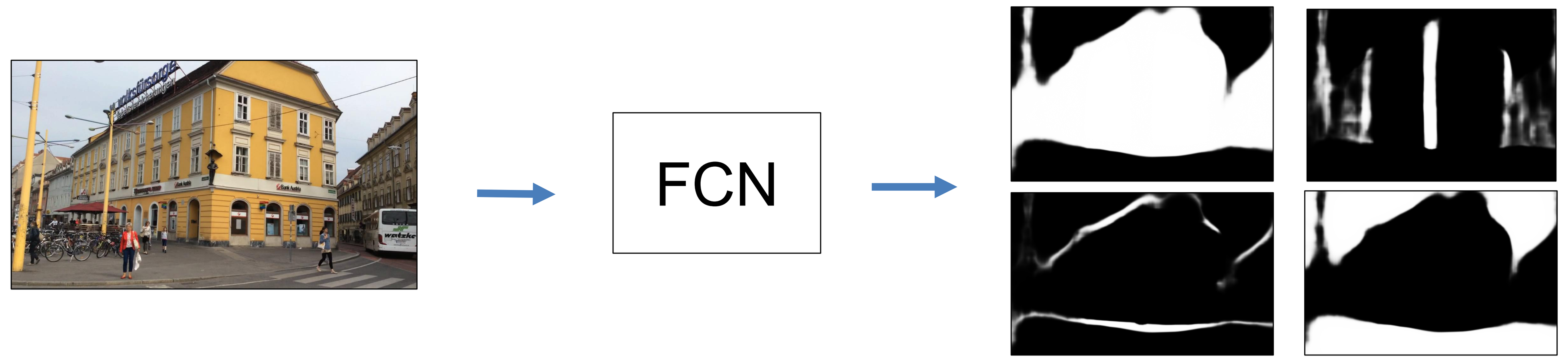
Relation Between Input Image and 2.5D Map

- How to construct the relation between the image and the model?

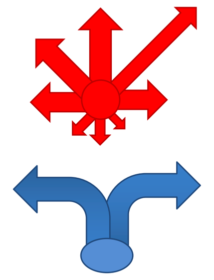


Input image | 2.5D map | Reprojection of the model under an accurate pose

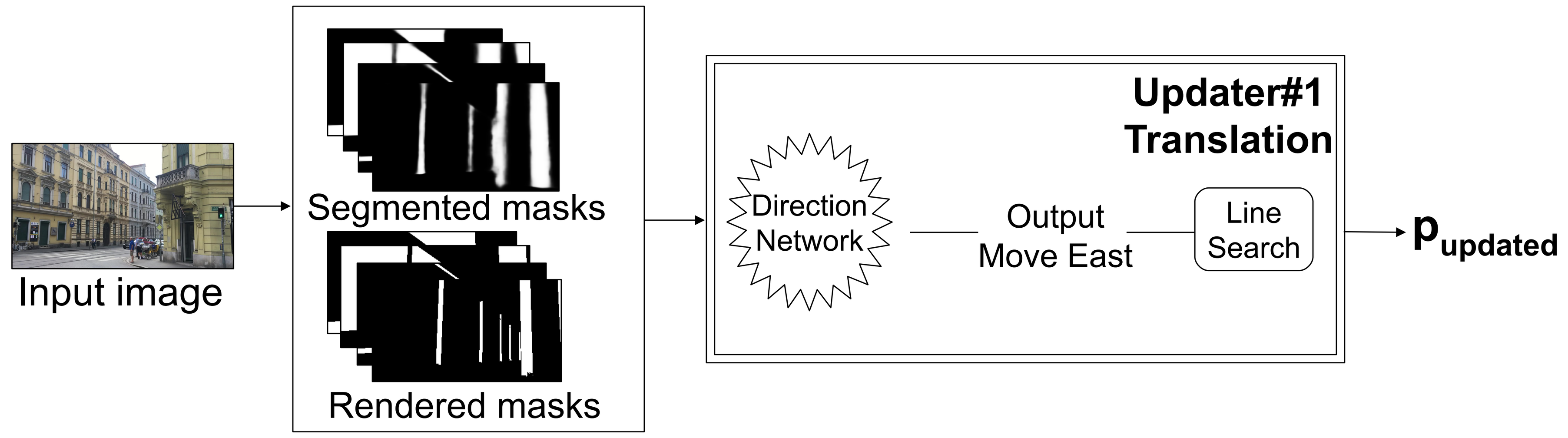
- We use **FCN** [1] to semantically segment the input image into related classes: *façade, vertical edge, horizontal edge and background*.



Learning to Predict a Direction for Pose Update



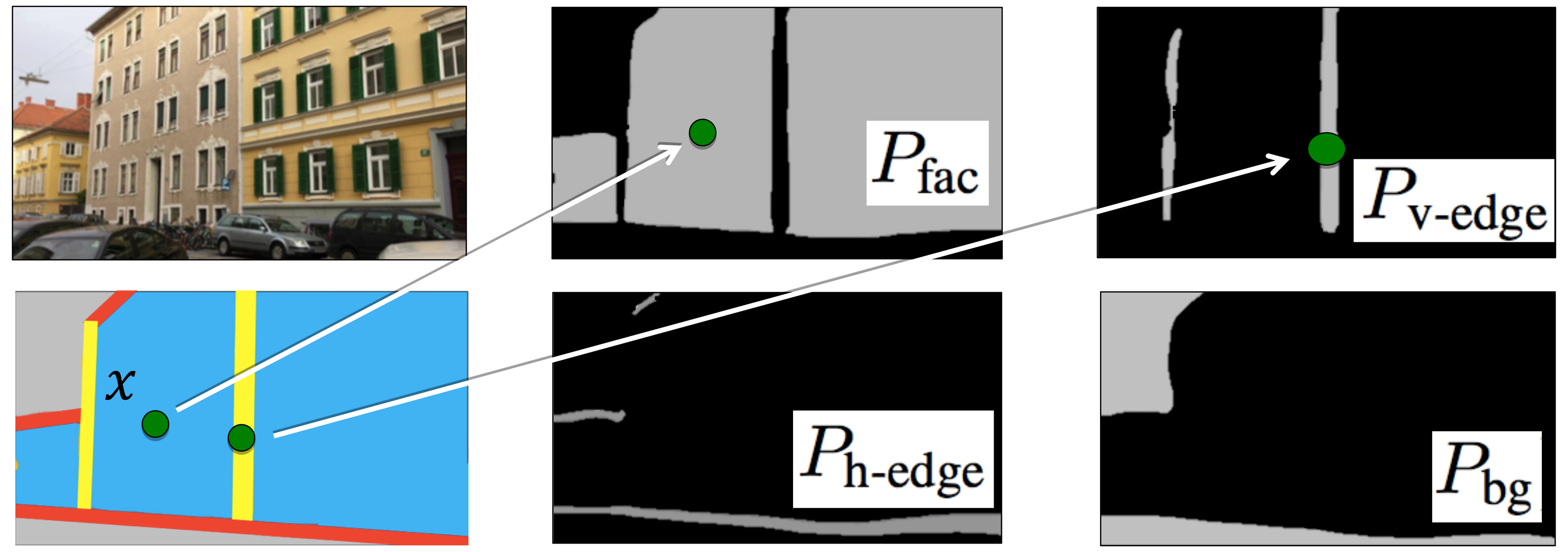
- Translation network: 8 directions or "don't move"
- Orientation network: 2 directions or "don't rotate"
- Example of a pose update by the Translation network:



- The networks predict only a direction for a pose update. The magnitude of the update is found by a **line search** algorithm: We perform a line search along the predicted direction, and keep the pose with the maximum likelihood score $\mathcal{L}$ (see below).

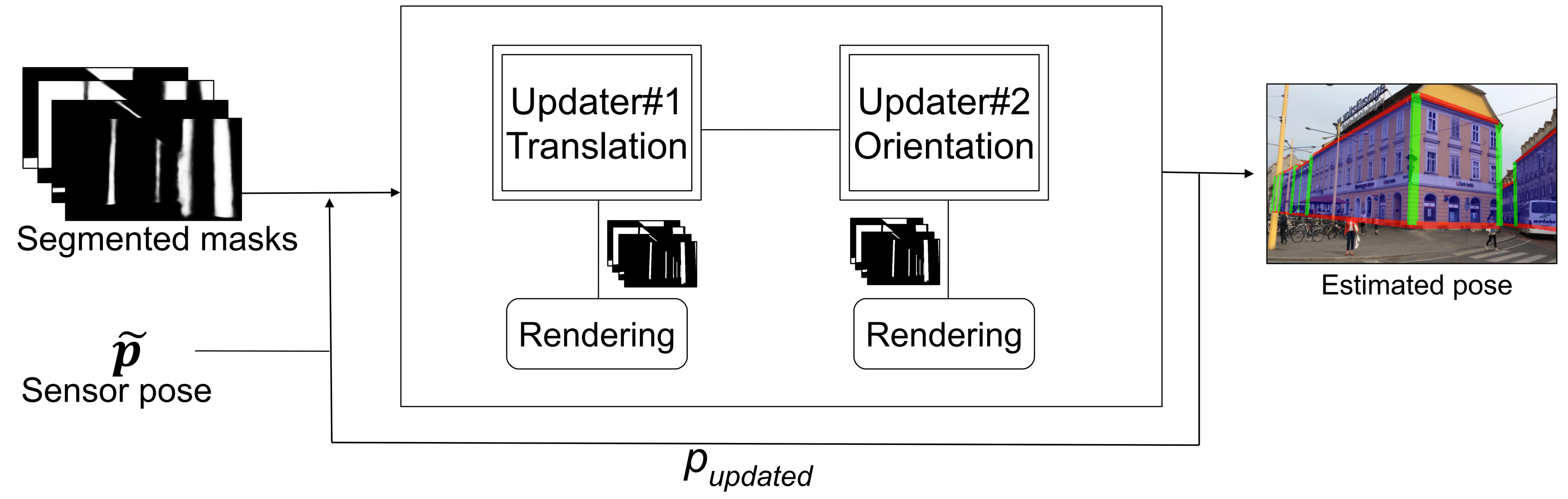
Pose Log-Likelihood $\mathcal{L}$

$$\mathcal{L}(pose) = \sum_{x} \log P_{class(Render(pose),x)}(x)$$



Render(pose) | Segmentation

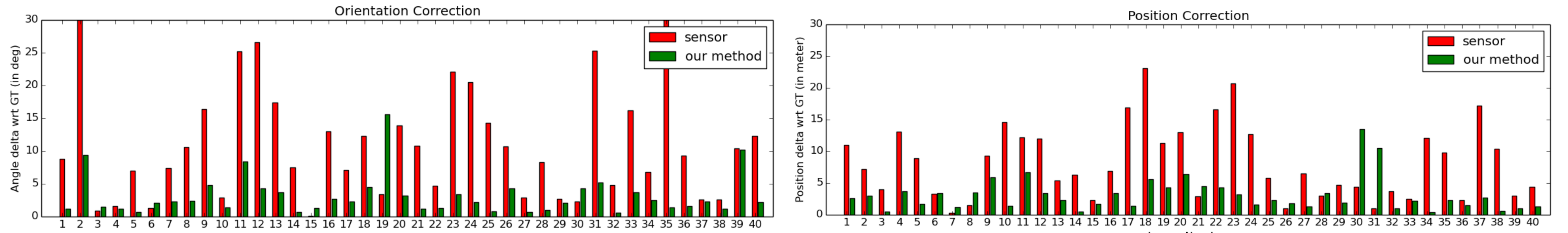
System Overview



- Both networks are iterated until their predictions are **"don't move"** and **"don't rotate"** at the same time.

Results

- We tested our approach on 40 test images and evaluated the position and orientation errors of $\tilde{p}$ and the pose found by our method w.r.t. the ground truth pose. Our method decreases the mean orientation error of $\tilde{p}$ from 11.3° to 3.2° and the mean position error from 13.4m to 3.1m.



Ground truth | Sensor pose | Our method | Ground truth | Sensor pose | Our method

References

[1] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

Acknowledgement

This work was funded by the **Christian Doppler Laboratory for Semantic 3D Computer Vision.**

Project web page: