University of Applied Sciences and Arts Northwestern Switzerland
School of Architecture, Civil Engineering and Geomatics

MASTER OF SCIENCE IN ENGINEERING

Monocular Depth Estimation with Generative Adversarial Networks

Depth maps are the most important element to gain the third dimension in various fields of application. They are usually collected with active sensors or through image-based estimation methods. In both cases, additional depth-capable sensor or camera poses are required to extrapolate the depth maps, which makes them less ubiquitous than normal RGB cameras. Aiming to find a more universal solution, this thesis researches the task of estimating the depth directly from a single RGB image. After the creation of a dataset, a generative adversarial networks structure is explored. To improve the estimations, diverse methods are implemented, which can eventually generate better depth maps than low-cost depth-sensors.

Data collection and post-processing

The dataset has been collected with different instruments ranging from smartphone with Time-of-Flight camera or LiDAR sensor, terrestrial laser scanning and multiple fixed cameras. The post-processing operations for the dataset preparation includes: image matching to extract the depth maps from the multiple fixed cameras' images, the upsampling of the low-resolution LiDAR data through a neural network with no quality loss and the equirectangular projection of the spherical data generated with the laser scanner (*Fig. 1*).

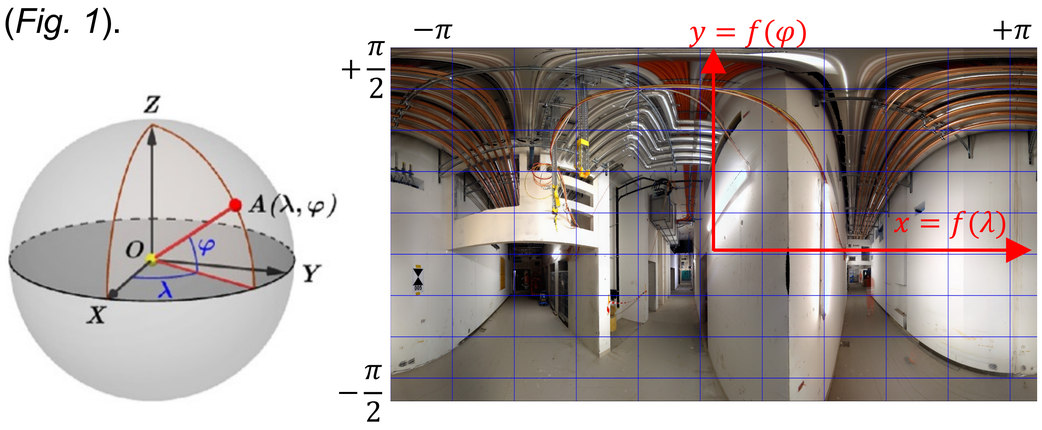

Fig. 1: Equirectangular projection of spherical data

Generative Adversarial Networks structure

The Generative Adversarial Networks' structure consists in a min-max game between a generator, which is trained to generate the depth maps conditioned bay the RGB-image distribution, and a discriminator, which is simultaneously trained with the latter to classify the presented depth maps as generated or true. The implemented networks based the generator on a *ResNeXt-101* backbone, which allowed the use of transfer learning to accelerate the training process, and the discriminator on a simple convolutional neural network.

Training and validation

As the collected dataset is small for such a complex task, during the iterative training process (*Fig. 2*), a data augmentation function was implemented to randomly flip and crop the samples, and additional open-source datasets has been used, to improve the generalization of the model. In addition, an early stopping function was integrated, which calculates the error on a separate validation set to avoid the overfitting of the model.

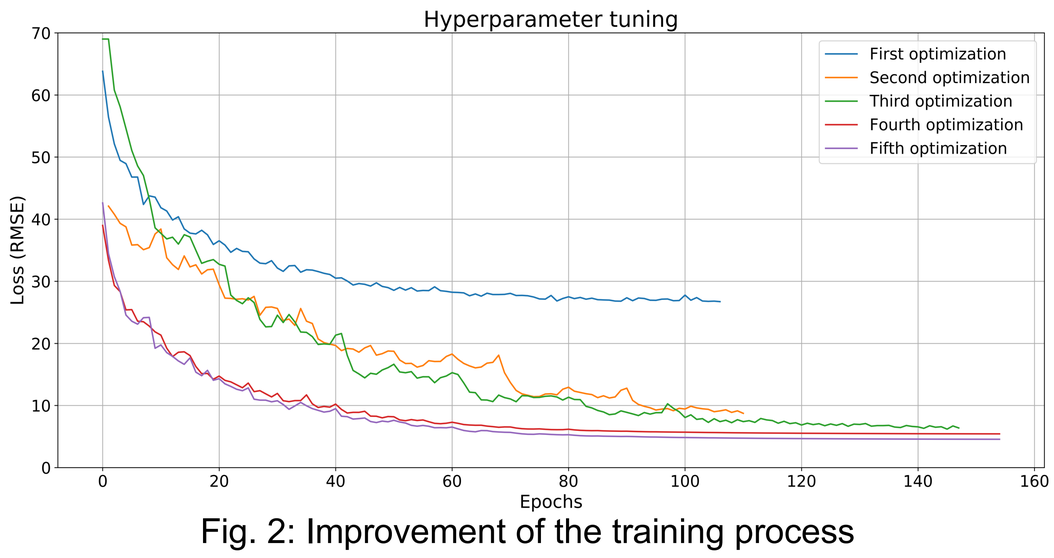

Fig. 2: Improvement of the training process

Testing result of the baseline

After the refinement, the created model has been tested on the 10% of the dataset, which was preserved before the training to test the model on samples never seen during that step. The model was able to reach a RMSE of 15.94 (6.25% on a intensity scale between 0 and 255) and an accuracy of 98%. The generated depth maps present a good structural consistency and minor blurry or incomplete zones in comparison to the ground truth (*Fig. 3*).

Fig. 3: Resulting depth maps from the GAN baseline

Depth map enhancement with multiple estimations

The limitation of the receptive field of the baseline network has been overcome through multiple estimations, which offer a good structure consistency through the low-resolution estimation and richness of details through the high-resolution estimation. With the help of an additional merging network, it was possible to combine the two estimations exploiting the advantages of both and reaching an RMSE of 14.66 (5.7% for intensity between 0 and 255) (*Fig. 4*).

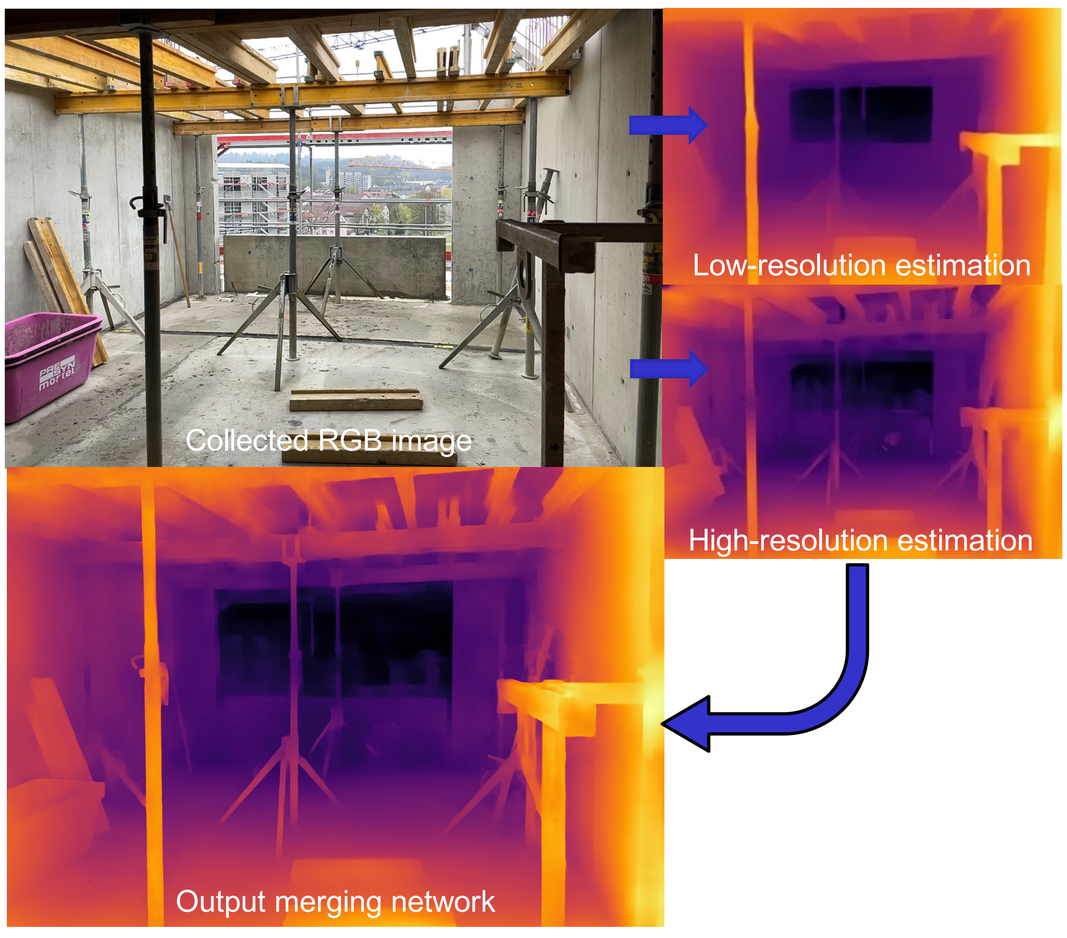

Fig. 4: Double estimation and merging network's results

References:

Boussias-Alexakis, E./Tsironisa, V./Petsa, E./Karras, G. 2016. Automatic Adjustment of Wide-base Google Street View Panoramas. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Science XLI-B1: 639-645*.

Miangoleh, S. Mahdi H./Dille, Sebastian/Mai, Long/Paris, Sylvain/Aksoy, Yagiz 2021. Boosting Monocular Depth Estimation Models to High-Resolution via Content-Adaptive Multi-Resolution Merging. *URL: http://arxiv.org/abs/2105.14021*

Author: Elia Ferrari
Examiner: Prof. Martin Christen
Expert: Bernhard Dräyer