

A Deep Learning Approach to Camera Pose Estimation

Federico Izzo

DISI, University of Trento

Trento, Italy

federico.izzo@studenti.unitn.it

Francesco Bozzo

DISI, University of Trento

Trento, Italy

francesco.bozzo@studenti.unitn.it

Abstract—The task of camera pose estimation aims to find the position of the camera that captured an image within a given environment. While different geometric approaches have already been studied in the literature, the aim of this project is to analyze and improve the performances of deep learning models for the camera pose estimation problem. In this work, we analyze models for both relative camera pose estimation (MeNet) and absolute camera pose estimation (PoseNet, MapNet). Moreover, we propose a pipeline for the generation of a ground truth dataset based on structure from motion techniques (COLMAP). Finally, we (1) show how the proposed framework has been used to build a dataset of the second floor of the Povo 1 building in the University of Trento, (2) train an absolute pose estimation model with PyTorch, and (3) deploy it through a web dashboard using FastAPI. The deep learning approach could give interesting results in combination with geometric methods, especially for: relocation after lost tracking, closed-loop detection, better dealing with moving objects in the scene. Even if the SOTA deep learning approach for this field is less accurate than geometric ones, it ensures more generalization capabilities at a reduced computational cost.

Index Terms—camera pose estimation, PoseNet, MapNet, deep learning, computer vision

I. INTRODUCTION

The *camera pose*, known also as *camera extrinsics*, can be expressed as a combination of two components:

- 1) a tuple of three elements that identifies the absolute coordinates x, y and z in a reference space:

$$x_c = (x, y, z) \quad x, y, z \in \mathbb{R} \quad (1)$$

- 2) a quaternion of four elements that identifies the rotation of the camera:

$$q_c = (qw, qx, qy, qz) \quad qw, qx, qy, qz \in \mathbb{R} \quad (2)$$

Consequentially, the pose is referred as $p_c = (x_c, q_c)$.

It is important to notice that this is not the only available representation for poses: other methods are based also on rotation matrices and Euler angles. It is worth specifying that even if Euler angles are the most straightforward and efficient in terms of memory consumption, they suffer from of the Gimbal lock problem. Also, although rotation matrices guarantee a good representation, they are more memory expensive (9 values) than quaternions (only 4 values): for this reason the latter form is preferred in this work.

Given an image I_c captured by a camera C , an absolute pose estimator E tries to predict the 3D pose orientation and location of C in world coordinates, defined for some arbitrary reference system in the 3D model. Formally speaking, the *absolute pose estimation (APE)* problem can be defined as the problem of estimating a function E taking as input an image I_c captured by a camera C and as output its respective pose:

$$E(I_c) = (x_c, q_c) \quad (3)$$

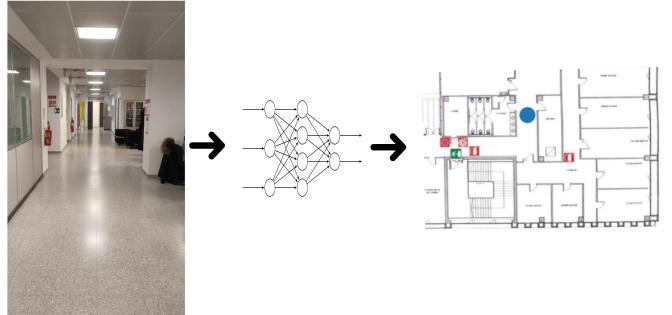


Fig. 1. APE model developed

Apart from APE, a popular task is also *relative pose estimation (RPE)*. In this kind of approach the estimator takes two images I_c^1 and I_c^2 captured by C and aims to predict the relative pose between them. In this case, the formulation of the function E described in Equation (3) is a little different, since it receives in input two images:

$$E(I_c^1, I_c^2) = (x_c^{rel}, q_c^{rel}) \quad (4)$$

where x_c^{rel} is defined as the absolute pose with *coordinates reference system* in I_c^1 or, in an equivalent way, as the translation vector from I_c^1 to I_c^2 .

With this work, we show how it is possible to build a deep learning model which is able to learn the function E using a data-driven approach.

In both scenarios, the estimator can be viewed as a model who describes an environment that can be questioned about the pose of an element within it.

II. RELATED WORKS

In the literature there are many deep learning approaches used to perform RPE and APE: here we focus on MeNet for the first and PoseNet [1] and MapNet [2] for the latter, since they can get good results without requiring huge computational costs.

APE deep learning models rely mostly on *transfer learning*: the idea is to use SOTA pre-trained vision models to extract features from images and use them to estimate camera extrinsics. The PoseNet model has been the first to be developed following this idea. The starting network for the knowledge transfer is a GoogLeNet [3], where the softmax classification layer is replaced with a sequence of fully connected layers. Even if the obtained results are decent, the model (1) lacks of generalization when applied to unseen scenes and (2) fails in learning complex environments.

In order to solve these problems, other techniques have been developed, which can be classified in:

- *end-to-end* approaches;
- *hybrid* approaches.

Most of the end-to-end proposed models are based on the PoseNet architecture, with the addition of some components, such as *encoder/decoder blocks*, *linear layers*, and *LSTM blocks*. The most successful model on this category is MapNet, with its related variants MapNet+ and MapNet+PGO [2].

On the other hand, hybrid approaches instead try to focus on different support tasks with the goal of helping the final pose prediction. Those techniques rely on unsupervised learning, 3D objects reconstruction and other data extracted with external tools: for these reasons, such methods are not considered in this work.

III. DATASET GENERATION

A. Tested approaches

The deep learning approaches applied in this document are *supervised learning* techniques that require a labeled dataset. Many attempts were made in order to generate a high-quality dataset for the deep learning models:

- *IMU sensors*: usage of gyroscope and accelerometer sensors of a smartphone to estimate the position of the camera during a video given a fixed origin point. The process of combining multiple noisy information sources to obtain a good estimation can be achieved through Kalman filters;
- *digital video*: usage of free online 3 dimensional datasets in which video can be recorder in a digital way;
- *motion capture system*: usage of a motion capture system that estimates the camera position following some tracking objects attached to the subject;
- *structure from motion techniques*: techniques that digitally reconstruct environments from a sequence of images with point clouds.

The main problem encountered with IMU sensors is the high presence of noise during data acquisition: this caused the processed signal to be very dirty, so much that the precision

is not acceptable for our intentions. A possible solution could have been to use some well calibrated hardware used in a controlled environment.

Most of the online-available 3D acquisitions are acquired with *depth sensors* or *LIDAR sensors*. For this reason, although the camera pose estimation would have been very precise, we could not reproduce this acquisition system in the University of Trento. In addition to that, only few datasets are freely available, and they usually lack of good documentation regarding how they were captured.

The motion capture system is able to follow the position of the tracked objects with extreme precision. In this case the main problems are related to the association of poses to video captured from the camera held by the tracked subject and difficulties involved in the system calibration.

The techniques of structure from motion are used to generate 3D models in case many photos are available. The overall idea is to feed the algorithm with these images in order to extract features and build a recomposition of the environment through a point cloud. During the reconstruction process, structure from motion tools also computes camera poses in an arbitrary reference system: this intermediate requirement have been exploited by us to generate a labeled dataset, as presented in Figure 2.

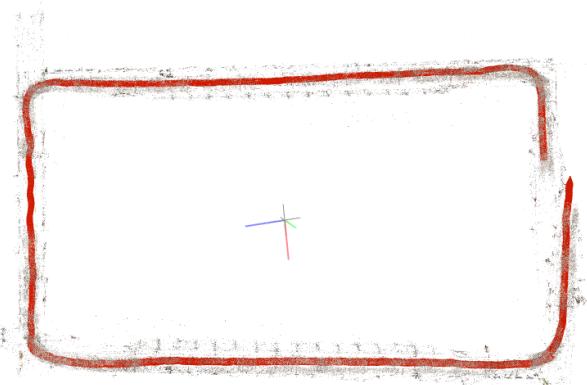


Fig. 2. Trajectory computed by COLMAP in the second floor of the Povo 1 building in the University of Trento. Note that structure from motion techniques may not detect closed-loops.

B. Pipeline

In this section we present an automatic pipeline which can be used to generate a camera pose estimation dataset using structure from motion techniques. The pipeline requires as input a video captured by any camera, even without sensor calibration. It is composed by several steps:

- 1) video split: the captured video is split into many frames using `ffmpeg`;
- 2) structure from motion: frames obtained from the previous step are fed into a structure motion tool. In our case we are using COLMAP [4];

- 3) dataset splitting: positions obtained during the camera estimation of the reconstruction process are split into three subsets: train, validation, and test.

In Figure 2 is presented the trajectory obtained with the structure from motion technique through COLMAP. The process involves a feature extraction phase.

C. COLMAP reconstruction

COLMAP [4] is a tool that allows to build a 3D point cloud reconstruction model of an environment by using photos of it. It is possible to generate point clouds with two different levels of precision: *sparse* and *dense*, and both of them are composed by:

- a set of points P which represent the features extracted from the photos (points in Figure 3);
- a set of poses \mathcal{V} associated with the images used for the reconstruction (red line in Figure 3);
- a coordinate reference system (CRS).

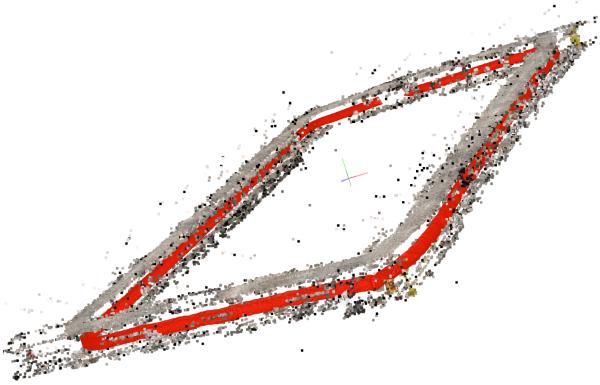


Fig. 3. Features extracted as a point cloud by COLMAP in the second floor of the Povo 1 building in the University of Trento. The red line represents the camera path.

The set of poses is represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and for each node i there is an association with a subset of environment features $\mathcal{F}_i : \mathcal{F}_i \subset P$. Edges in \mathcal{E} connect poses from images that use common features.

The sparse reconstruction algorithm firstly extracts features from each given image and builds the set P called *bag of words*. Once this is done, thanks to a convolution on each image, it creates spacial association between nodes in the graph comparing images with the bag of words. The convolution enables to map only subsections of the whole image: this allows to estimate the movement based on the position of the matrix used for the convolution within the image grid. Once associations are done, the features and the poses are composed in order to create the reconstructed model.

Also, the algorithm for the dense model is very similar to the one used for the sparse. The main difference is about how the association are created: in this case the algorithm uses point-wise associations and not convolutions. Even if it requires more computational resources, this approach allows to be more accurate and to create a more definite point cloud.

Both algorithms can work on a sequence of unrelated photos of the same environment or a sorted sequence of frames: note that the reconstruction process is way more efficient in the second case.

D. Coordinate reference system alignment

The dataset generated by COLMAP has a *coordinate reference system (CRS)* chosen arbitrarily during the reconstruction. Origin and axes of the COLMAP system may not coincide (actually this is always the case) with the real-world CRS. In order to align them, three steps are required:

- 1) scale: it is necessary to scale the relative positions of the points with respect to the real-world CRS. This can be achieved measuring the distance between two known points in both COLMAP and real life;
- 2) translation: align the origins;
- 3) rotation: align the axes.

The last two steps can be grouped together by using the euclidean or rigid transformation. It involves a rotation R , a translation t and at least three points for both the CRSs that represent the same locations. Equation (5) describes the rigid transformation.

$$\begin{aligned} A &= \{(x_1^A, y_1^A), (x_2^A, y_2^A), (x_3^A, y_3^A)\} \\ B &= \{(x_1^B, y_1^B), (x_2^B, y_2^B), (x_3^B, y_3^B)\} \\ B &= R \times A + t \end{aligned} \quad (5)$$

Matrix R and vector t are obtained using *Singular Value Decomposition (SVD)*, it takes a matrix E and return 3 other matrices, such that:

$$\begin{aligned} [U, S, V] &= \text{SVD}(E) \\ E &= USV^T \end{aligned} \quad (6)$$

To obtain the matrix R , the first step consists in aligning on the same origin the two datasets centroids. This is done by subtracting to each coordinate of each point the centroid of the respective dataset. After this, it is possible to ignore the translation component t and compute the rotation R :

$$\begin{aligned} H &= (A - \text{centroid}_A)(B - \text{centroid}_B)^T \\ [U, S, V] &= \text{SVD}(H) \\ R &= VU^T \end{aligned} \quad (7)$$

Finally, it is possible to use the Equation (5) to obtain the translation vector t :

$$\begin{aligned} B &= R \times A + T \\ \text{centroid}_B &= R \times \text{centroid}_A + T \\ t &= \text{centroid}_B - R \times \text{centroid}_A \end{aligned} \quad (8)$$

IV. MODELS

In this work we take in consideration some SOTA models for camera pose estimation, also adding small modifications to make them fit better to our use case scenario. In particular, we present:

- MeNet (add citation) for RPE;
- PoseNet [1] and MapNet [2] for APE.

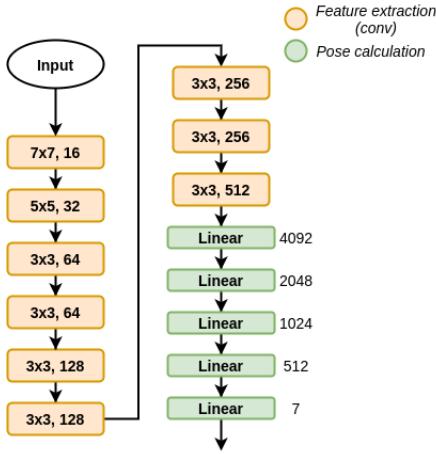


Fig. 4. The architecture of the MeNet model.

A. MeNet

The first model we would like to analyze is the MeNet model (Figure 4), which is specifically targeted for RPE. The input of the network consists in a stack of two images: the goal is to estimate the relative pose of the second image with respect to the first one. As presented in Figure 4, the MeNet is composed by nine deep convolutional layers followed by a pair of linear layers. While the first part of the network works as a feature extraction mechanism, the last layers are used to combine the extracted features.

In order to train the model, we use a loss function that consists in the weighted composition of two Mean Square Errors (MSE) computed separately on positions and quaternions:

$$Loss(w) = \frac{1}{N} \sum_{i=1}^N \|P^i - \hat{P}^i\|_2^2 + \alpha \|Q^i - \hat{Q}^i\|_2^2 \quad (9)$$

where P , \hat{P} , Q , \hat{Q} , and α are the ground truth position vector, the estimated position vector, the ground truth quaternions, the estimated quaternions, and the weight for balancing the displacement error and the rotation angle error.

B. PoseNet

Since the results given by RPE deep learning solutions are not very promising, from now on we are going to consider models strictly developed for APE. In this sense, the first model we present is the PoseNet model [1]. Just like the MeNet model, the PoseNet is made up by two components:

- feature extraction through a sequence of convolutional layers. This component has been named internally also as *backend*;
- pose regression on the extracted features using linear layers.

This model architecture is actually pretty convenient, since it can use pre-trained deep convolutional models, through the transfer learning approach. In most of the cases, this kind of models are trained on the ImageNet dataset [5], which

counts almost 3.2 million real-world images. Some examples of SOTA backend models that have been considered are: GoogLeNet [3], ResNet [6], and EfficientNet-B7 [7]. Table I shows the accuracy over the k=(1, 5) top predictions for the used backends on the ImageNet dataset.

TABLE I
BACKENDS PERFORMANCE IN IMAGENET

Model	Acc@1	Acc@5
GoogLeNet	69.778	89.530
ResNet-18	69.758	89.078
ResNet-34	73.314	91.420
ResNet-50	76.130	92.862
ResNet-152	78.312	94.046
EfficientNet-B7	84.122	96.908

Since ImageNet pre-trained models are used for classification purposes, we can just remove the last layers, and use only the feature extraction-related ones. In this way we are introducing in the PoseNet a feature extraction mechanism on real-world images with minimum effort: training by scratch these models would require a huge amount of computational power.

Also in this model we adopt the weighted loss described in Equation (9), also used in the MeNet model.

C. MapNet

The MapNet model for APE represents an evolution of the PoseNet model: in fact, the model architecture remains actually the same. On the contrary, the main difference between the PoseNet is the loss function used to train the model. In this case, the errors in the prediction of absolute poses are not the only ones which are penalized: also errors in the relative poses are taken in consideration.

The size of the last linear block depends on the dimension of the map that we would like to introduce.

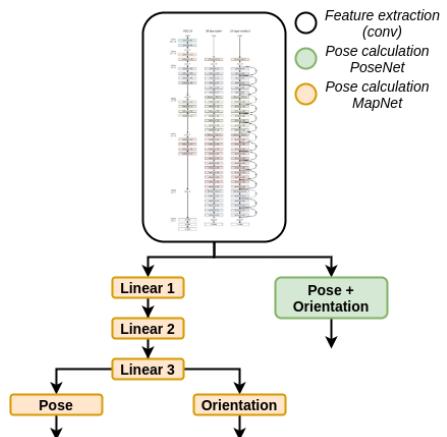


Fig. 5. The architecture of the PoseNet and MapNet models.

V. RESULTS

The settings used for sparse dataset generation of Povo 1 used a sequence of images with Poisson mesher.

A. PoseNet

Several pretrained models can be used as features extractor in the PoseNet structure. In Table IV are presented the most powerful ones for features extraction tested on the same final linear encoder. The overall trend is similar, this highlights that the extracted features are enough for the task independently of the backbone used.

TABLE II
POSENET BACKEND COMPARISON

Model	Position error	Rotation error	Total parameters	Trainable parameters
GoogLeNet	0.781	0.119	-	-
ResNet-18	0.635	0.288	11,180,103	3,591
ResNet-34	0.632	0.223	21,288,263	3,591
ResNet-50	0.707	0.191	23,522,375	14,343
ResNet-152	0.594	0.139	58,158,151	14,343
EfficientNet-B7	0.817	0.132	63,804,887	17,927

TABLE III
POSENET LOSSES COMPARISON

Loss	Position Error	Rotation Error
SmoothedL1Loss	0.594	0.139
L1Loss	0.906	0.226
MSE	NaN	NaN
weighted_custom	NaN	NaN

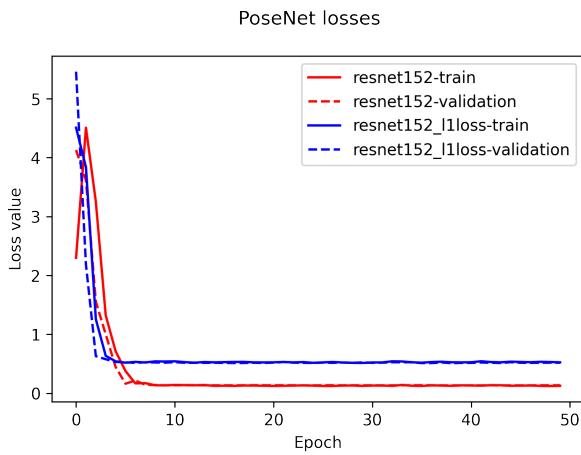


Fig. 6. PoseNet (ResNet-152) losses

B. MapNet

Several pretrained models can be used as features extractor in the MapNet structure. In Table IV are presented the most powerful ones for features extraction tested on the same final

linear encoder. The overall trend is similar, this highlights that the extracted features are enough for the task independently of the backbone used.

TABLE IV
MAPNET BACKEND COMPARISON

Model	Position error	Rotation error	Total parameters	Trainable parameters
GoogLeNet	0.225	0.0876	-	-
ResNet-18	0.202	0.0658	14,853,703	3,677,191
ResNet-34	0.187	0.0757	24,961,863	3,677,191
ResNet-50	0.220	0.0969	30,330,951	6,822,919
ResNet-152	0.233	0.0869	64,966,727	3,677,191
EfficientNet-B7	0.210	0.0848	71,658,455	7,871,495

Another point that emerges is the importance of the final encoder, it works similarly of the *bag of words* used by structure from motion (link al paper). Extracted features are mapped in a space that is used later as a comparison tool for new images for which the pose is asked. For this reason the final encoder was modified from the original one (link al paper) in order to increase the latent space in which data can be stored.

C. Comparison

D. Dashboard

A dashboard was developed with the aim to easily allow users to interact with model inference through a web-server. In Figure 11 is presented the *UI* where red zones are not walkable areas.

VI. MATERIALS

Every material used in the project have been uploaded respectively:

- the datasets have been uploaded on the Google Drive folder;
- the code is available in the GitHub repository.

The project has been developed in Python 3, using common data science libraries, such as numpy, pandas, PyTorch, matplotlib, scipy, and many others.

A. Repository organization

The repository follows the structure:

- `camera-pose-estimation/`
 - `model/` contains everything related to the deep learning part of the project. It also includes the code used for implementing the web server under `webserver.py` and `static/`.
 - `tools/` contains scripts used for the dataset generation pipeline.
- `config_parser/`: Python package written by us that allows to create configuration files, with the idea of improving reproducibility in our experiments. Each configuration file can be subdivided in sections: for each section you can define variables with the syntax `label=value`,

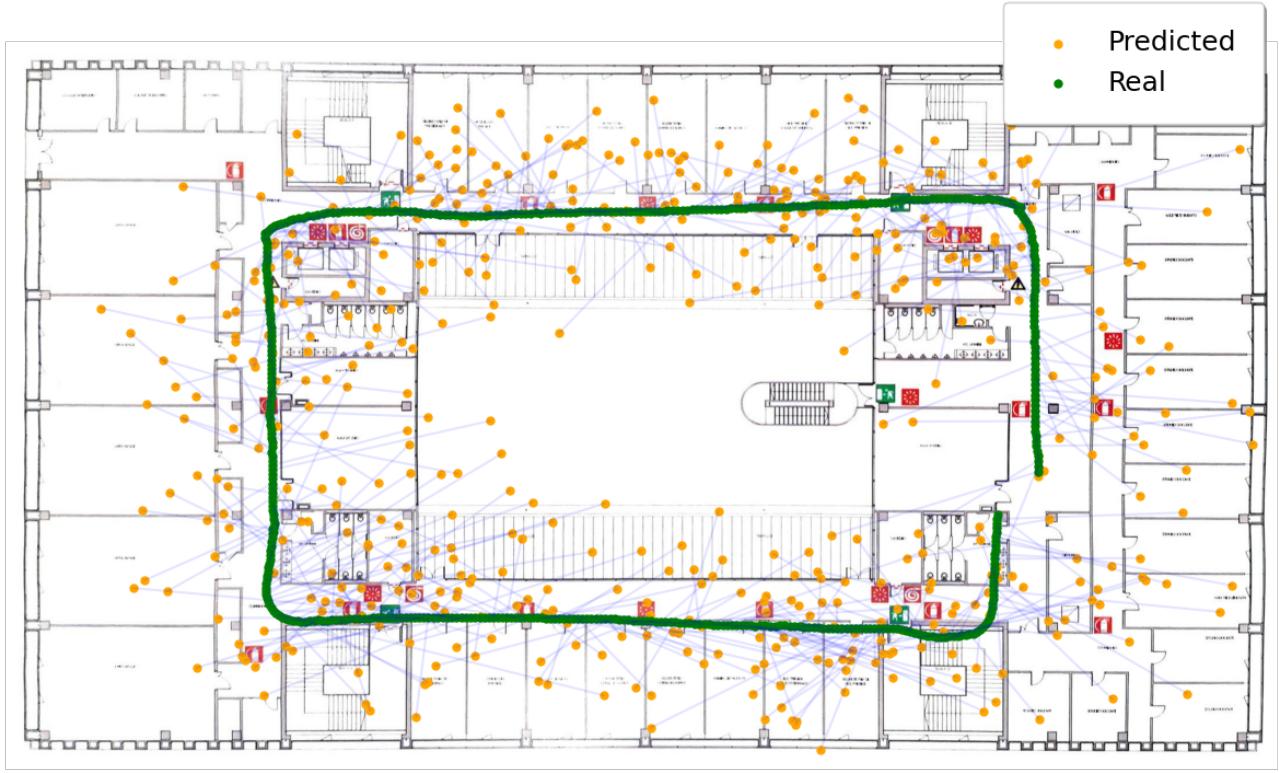


Fig. 7. Predicted trajectory PoseNet

where `value` is a parsable JSON object (boolean, int, float, list, object).

- `notebooks/` contains some Python Jupyter Notebooks that have been used for data exploration, validation, and post-processing of the model predictions.

B. Data organization

For each footage, a folder has been created:

- `imgs/` contains the video frames exported with `ffmpeg`;
- `processed_dataset/` contains the train, validation, and test datasets that can be reused during different trainings: this helps speeding up the loading procedure from ...minutes to ...seconds;
- `workspace/` contains the models generated by COLMAP;
- each of `train.csv`, `validation.csv`, and `test.csv` contains a table for specifying the pose for each image frame. This are the files generated with the `video_to_dataset.sh` script.

VII. CONCLUSION

REFERENCES

- [1] A. Elmoogy, X. Dong, T. Lu, R. Westendorp, and K. Reddy, “Linear-posenet: A real-time camera pose estimation system using linear regression and principal component analysis,” in *2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*, 2020, pp. 1–6.

- [2] S. Brahmbhatt, J. Gu, K. Kim, J. Hays, and J. Kautz, “Mapnet: Geometry-aware learning of maps for camera localization,” *CoRR*, vol. abs/1712.03342, 2017. [Online]. Available: <http://arxiv.org/abs/1712.03342>
- [3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” *CoRR*, vol. abs/1409.4842, 2014. [Online]. Available: <http://arxiv.org/abs/1409.4842>
- [4] A. Fisher, R. Cannizzaro, M. Cochrane, C. Nagahawatte, and J. L. Palmer, “Colmap: A memory-efficient occupancy grid mapping framework,” *Robotics and Autonomous Systems*, vol. 142, p. 103755, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0921889021000403>
- [5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: <http://arxiv.org/abs/1512.03385>
- [7] M. Tan and Q. V. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” *CoRR*, vol. abs/1905.11946, 2019. [Online]. Available: <http://arxiv.org/abs/1905.11946>

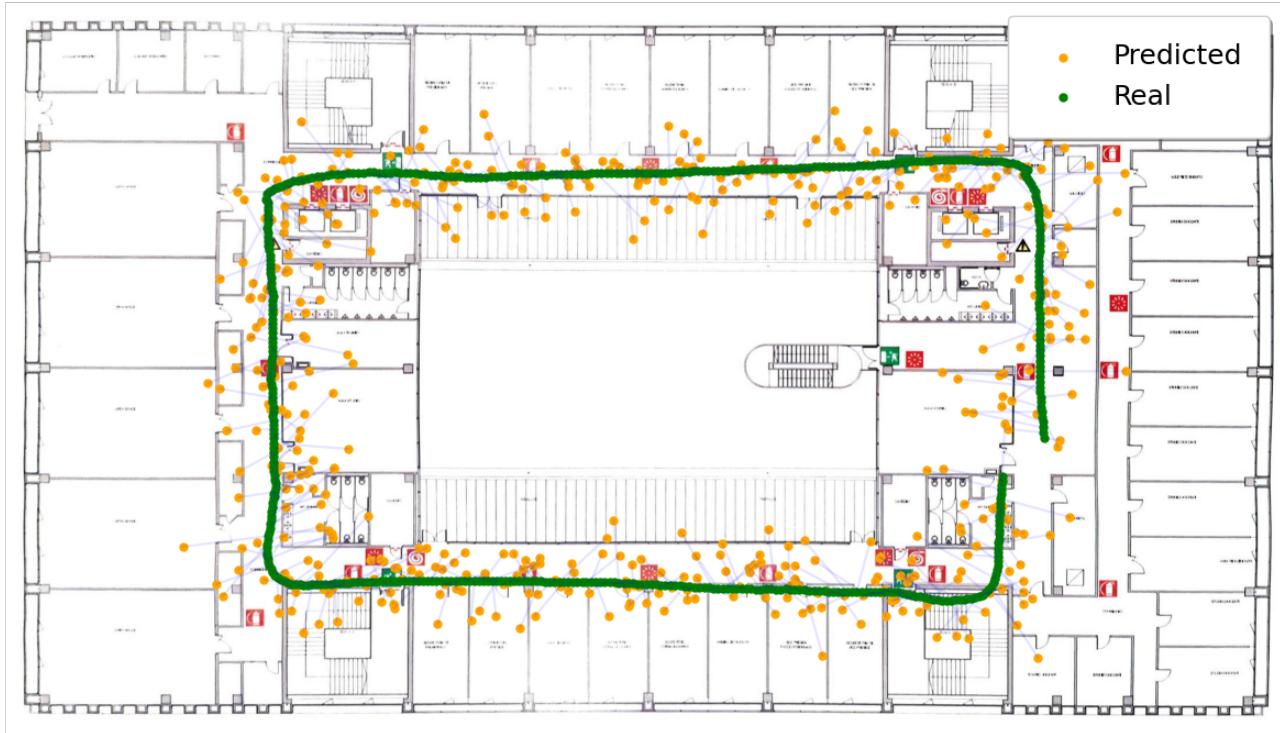


Fig. 8. Predicted trajectory MapNet

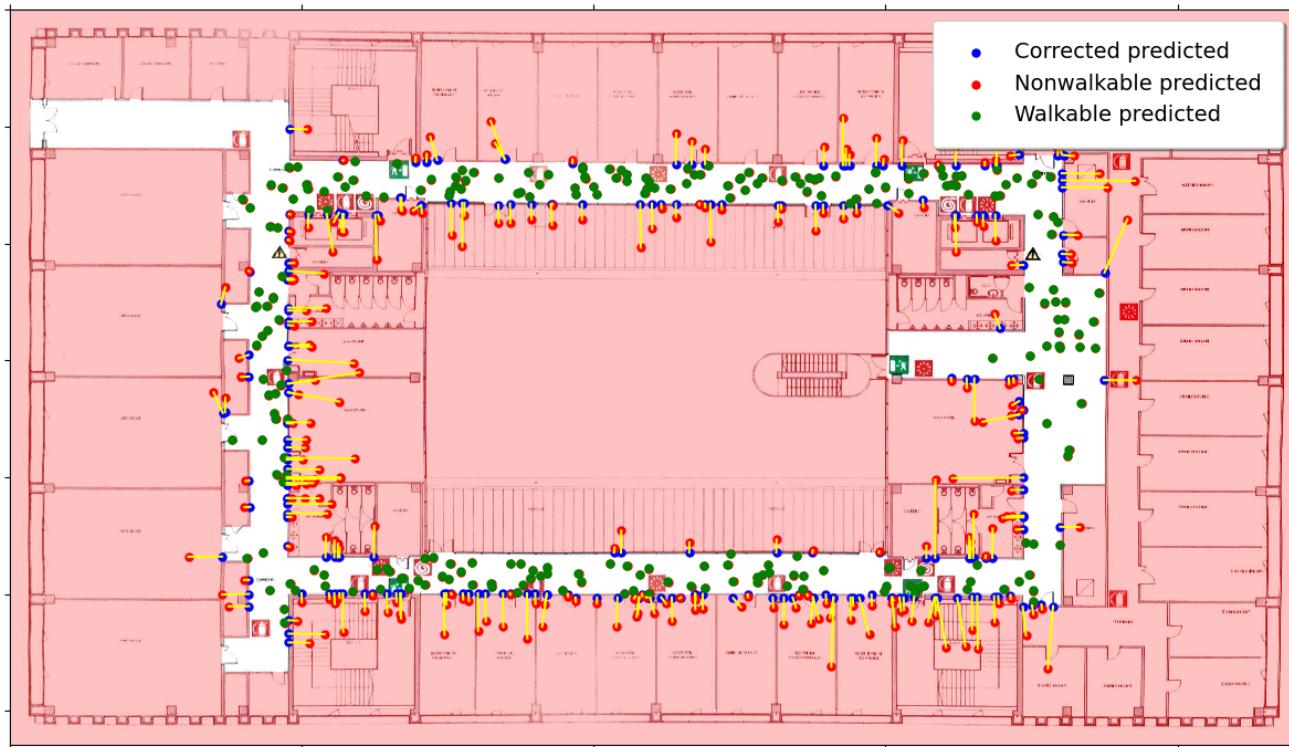


Fig. 9. Predictions post-processed on walkable areas

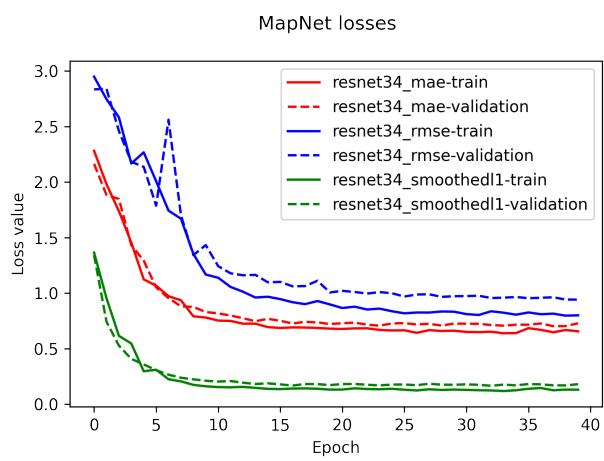


Fig. 10. MapNet (ResNet-34) losses



Submit ▾



Fig. 11. Inference dashboard