**Time: from 13/03/2019 to 20/03/2019**

**Paper reading：**

CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction

Author：Keisuke Tatento

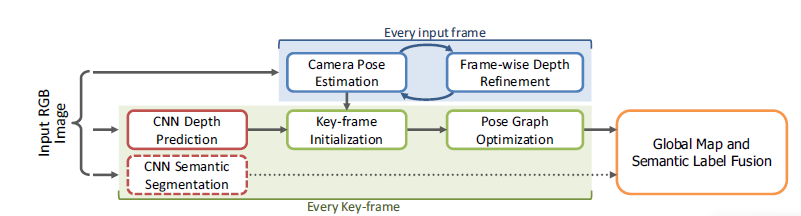
Publication: CVPR2017

**Main idea:**

This paper investigates how predicted depth maps from a deep neural network can be deployed for accurate and dense monocular reconstruction.

**Framework**

The flow diagram in Fig. 1 sketches the pipeline of the framework. Authors employ a key-frame based SLAM paradigm. Within such approach, a subset of visually distinct frames is collected as key-frames, whose pose is subject to global refinement based on pose graph optimization. At the same time, camera pose estimation is carried out at each input frame, by estimating the transformation between the frame and its nearest key-frame. To maintain a high frame-rate, they propose to predict a depth map via CNN only on key-frames. In particular, if the currently estimated pose is far from that of existing keyframes, a new key-frame is created out of the current frame and its depth estimated via CNN.



**CNN model for SLAM**

The depth prediction architecture is based on ResNet50 and initialized with pre-trained weights on ImageNet. Pooling and FC are replaced by a sequence of residual up-sampling blocks composed of a combination of unpooling and convolutional layers. After up-pooling, drop-out is applied. The loss function is based on the **reverse Huber function.**

They also retrained this network for predicting pixel-wise semantic labels for RGB images. In this way, they modified the network so that it has as many output channels as the number of categories and employed a soft-max layer and a cross-entropy loss function to be minimized via back-propagation and SGD.

**Key-frame Creation & Pose Graph Optimization**

There is a problem that sensors for SLAM have different intrinsic parameters from those used to capture the training set, the results will be inaccurate. In this way, they propose to adjust the depth regressed via CNN with the ratio between the focal length of current camera, and that of the sensor used for training, as

Where is the depth map directly regressed by CNN

This transformation is estimated by minimizing the photometric residual between the intensity image of the current frame and the intensity image of the nearest key-frame via weighted Gauss-Newton optimization based on the objective function

Where is Huber norm and is a function measuring the residual uncertainty. And r is the photometric residual defined as

while represents a 3D element of the vertex map computed from the key-frame’s depth map

Once is obtained, the current camera pose in the world coordinate system is computed as

(文章之后介绍了提出的3D地图重建问题中的单目深度估计优化，之后补充)

**Code reading:**

Github.com/gaoxiang12/slambook

**Next week work plan**

Continue to read Github.com/gaoxiang12/slambook

Read paper: Fully Convolutional Networks for Semantic Segmentation