G2L-Net: Global to Local Network for Real-time 6D Pose Estimation with Embedding Vector Features

Wei Chen [1,2] Xi Jia[1] Hyung Jin Chang[1] Jinming Duan[1] Ales Leonardis[1]
[1] School of Computer Science, University of Birmingham
[2] School of Computer Science, National University of Defense Technology

A. Overview

This document provides more details of our G2L-Net. Section B describes how we train our G2L-Net. Section C compares another state-of-the-art method [4] with ours. Finally, we attach a demo video, which suggests the effectiveness of the G2L-Net in real scene.

B. Training details

In training stage, the architecture is different from that of inference stage, since we do not deploy block C (in Figure A) in inference stage.

The training stage consists of two steps. In the first step, we fine-tune the 2D detector, *i.e*. YOLO-V3 [3] on the ImageNet [1] to locate the 2D region of interest and access the class probability map.

In the second step, we jointly train the translation and rotation localization networks. The translation localization network performs two related tasks: 3D segmentation and translation residual estimation. For 3D segmentation, we employ cross-entropy loss $\mathcal{L}_{seg}$. For translation residual estimation, the loss function is defined as:

$$\mathcal{L}_{tran} = \frac{\sum_{i=1}^{3}(x_i - \widetilde{x}_i)^2}{3}, \tag{1}$$

where x_i is the coordinate of the ground truth, x_i^p is the estimated translation residual.

The rotation localization network consists of three blocks: A, B, and C (as shown in Figure A). The loss function of block A is defined as the mean square error between the predicted and ground truth directional vectors:

$$\mathcal{L}_A = \min_{\boldsymbol{\theta}} \frac{1}{K\,|\mathcal{P}|} \sum_{k=1}^{K} \sum_{i} \|\widetilde{v}_k(\mathcal{P}_i;\boldsymbol{\theta}) - v_k(\mathcal{P}_i)\|_2^2, \tag{2}$$

where K is the number of 3D keypoints. $\boldsymbol{\theta}$ is the network parameters. $\widetilde{v}_k(\mathcal{P}_i;\boldsymbol{\theta})$ and $v_k(\mathcal{P}_i)$ are the predicted vector and the ground truth vector, respectively. $\mathcal{P} \in \mathbb{R}^{n\times 3}$ denotes the object points in the camera space. $|\mathcal{P}|$ denotes the number of object points.

For block B and C, we use the modified corner loss proposed in [2]:

$$\mathcal{L}_{B,C} = \frac{1}{8}\sum_{k=1}^{8} \left\| P_k - \widetilde{P}_k \right\|^2, \tag{3}$$

where P_k and $\widetilde{P}_k$ are ground truth and prediction of the k_{th} keypoint, respectively.

The ground truth P is generated by the pre-defined 3D bounding box and the ground truth pose in block B. However, the ground truth of block C is the residual between ground truth of block B and the prediction of block B.

We combine all the losses together to simultaneously optimize all networks:

$$\mathcal{L} = \mathcal{L}_{tran} + \lambda_1\mathcal{L}_{seg} + \lambda_2\mathcal{L}_A + \lambda_3\mathcal{L}_B + \lambda_4\mathcal{L}_C, \tag{4}$$

where $\lambda_i = \{0.01, 1, 0.001, 0.001\}$.

C. Comparison with state-of-the-art

We compare our method with another state-of-the-art method, $Densefusion$ [4], which also used point cloud for 6D object pose estimation. $Densefusion$ estimated the 6D object pose from point clouds and RGB information in a global space. However, we estimate the 6D pose from point clouds in a local space. Figure B suggests that our method is more robust to viewpoints changes, and can handle some failure cases of $Densefusion$.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[2] Charles R. Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J. Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[3] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1

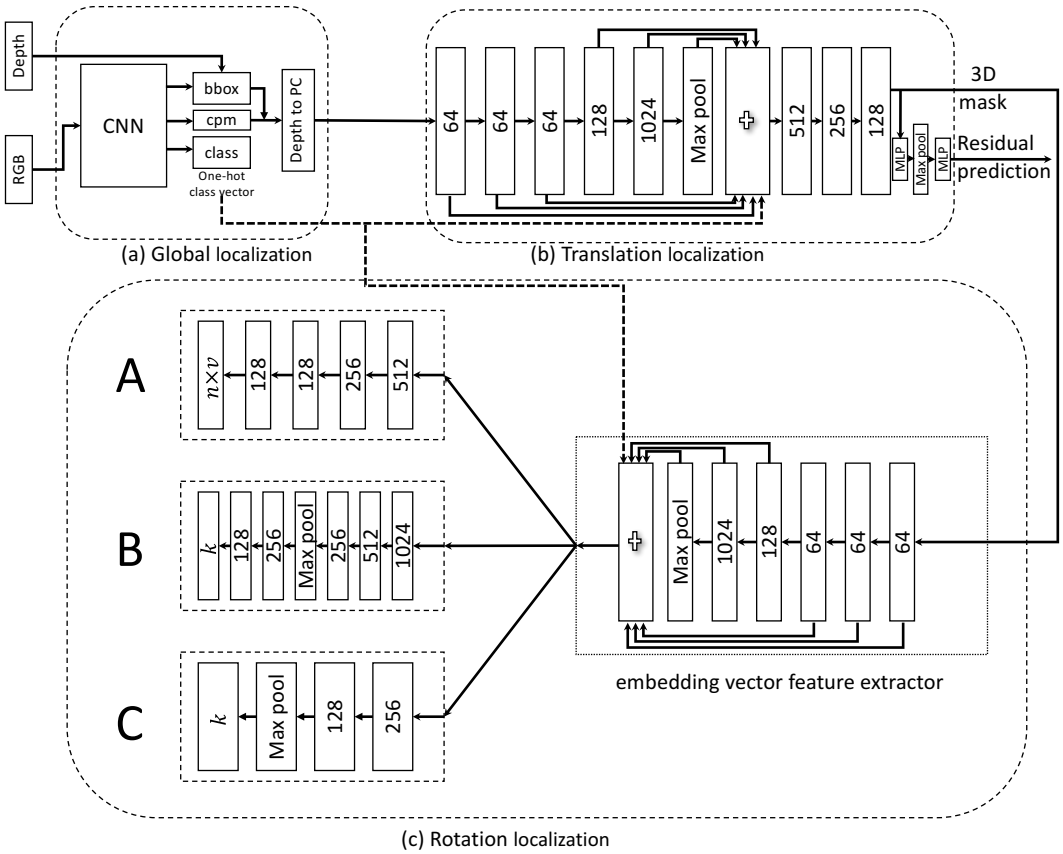

Figure A. **Training architecture of G2L-Net**. In training stage, we have three different blocks of rotation localization network, which is different from the inference architecture.

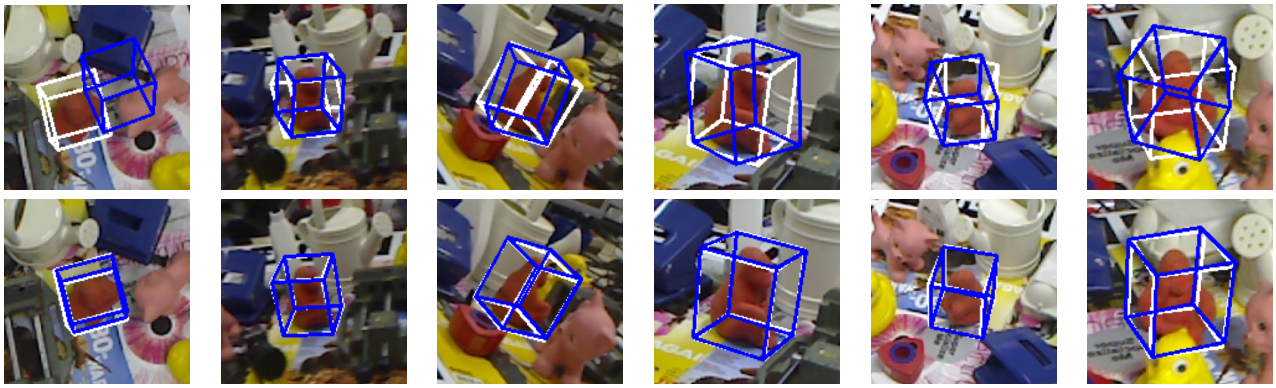

Figure B. **Visualization comparison**. The top row lists some failure cases of the state-of-the-art method $Densefusion$ [4]. The bottom row presents the corresponding detection results of our G2L-Net. For all images, the white bounding boxes are the ground truth, while the blue bounding boxes are prediction results.

[4] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. 2019. 1, 2