Supplementary Material: Stereo R-CNN based 3D Object Detection for Autonomous Driving

Peiliang Li[1], Xiaozhi Chen[2], and Shaojie Shen[1]
[1]The Hong Kong University of Science and Technology, [2]DJI
pliap@connect.ust.hk, cxz.thu@gmail.com, eeshaojie@ust.hk

1. Network Architecture Details

The network structure is illustrated in Fig. 2, where we describe the spatial size of each data block in detail. The batch-size is ignored for simplifying the representation. We use weight-share ResNet-101 [2] and Feature Pyramid Network (FPN) [4] for left-right feature extraction. For stereo RPN, we use five scales of feature maps $\{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64}\}$, which are corresponding to five scales of anchors $\{32, 64, 128, 256, 512\}$. The output channel of RPN classification is six corresponding to three anchor ratios $\{0.5, 1, 2\}$ with "object or background" categories. Similarly, we have 18 output channels in stereo regression for three anchor ratios with six box offsets $[\Delta u, \Delta w, \Delta u', \Delta w', \Delta v, \Delta h]$. For stereo R-CNN, we apply 7×7 and 14×14 RoI Align [1] on four scales of feature maps $\{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$ for regression and keypoint prediction respectively. The R-CNN branch produces object classes, stereo 2D bounding boxes, dimensions and viewpoints. We also describe the output size for each term in Fig. 2. For keypoint prediction, the final output size is 28×6, where the 28×4 part is aggregated for predicting the type and location of the perspective keypoint, and the 28×2 part is for predicting the location of two boundary keypoint.

2. Sparse Constraints for 3D Box Estimation

We illustrate four types of perspective keypoint in Fig. 1, where we define the camera frame and object frame, and each sub-image represents a typical case of one perspective keypoint type. For different types of the perspective keypoint, 2D box edges correspond to different 3D box corners. Therefore, constraining equations need to changed appropriately according to relations of 3D box corners and 2D edges. We formulate seven equations of 3D box estimation in detail for each type of the perspective keypoint:

$$
\begin{aligned}
v_t &= (y - \tfrac{h}{2})/(z + \tfrac{w}{2}sin\theta + \tfrac{l}{2}cos\theta), \\
v_b &= (y + \tfrac{h}{2})/(z + \tfrac{w}{2}sin\theta + \tfrac{l}{2}cos\theta), \\
u_l &= (x + \tfrac{w}{2}cos\theta + \tfrac{l}{2}sin\theta)/(z - \tfrac{w}{2}sin\theta + \tfrac{l}{2}cos\theta),
\end{aligned}
$$

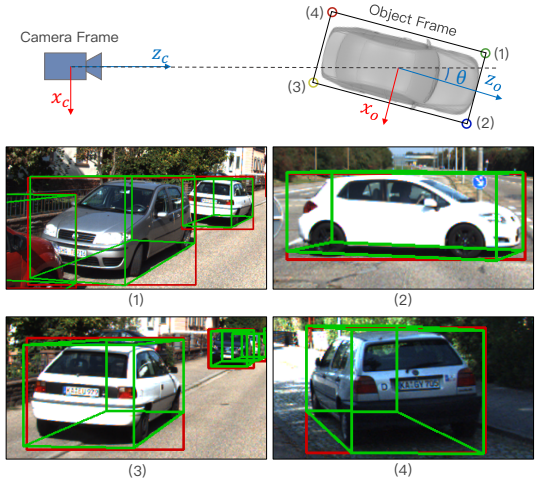

Figure 1: Four types of perspective keypoints illustration and corresponding projections.

$$
\begin{aligned}
u_p &= (x - \tfrac{w}{2}cos\theta + \tfrac{l}{2}sin\theta)/(z + \tfrac{w}{2}sin\theta + \tfrac{l}{2}cos\theta), \\
u_r &= (x - \tfrac{w}{2}cos\theta - \tfrac{l}{2}sin\theta)/(z + \tfrac{w}{2}sin\theta - \tfrac{l}{2}cos\theta), \\
u_l &= (x - b + \tfrac{w}{2}cos\theta + \tfrac{l}{2}sin\theta)/(z - \tfrac{w}{2}sin\theta + \tfrac{l}{2}cos\theta), \\
u'_r &= (x - b - \tfrac{w}{2}cos\theta - \tfrac{l}{2}sin\theta)/(z + \tfrac{w}{2}sin\theta - \tfrac{l}{2}cos\theta).
\end{aligned} \tag{1}
$$

$$
\begin{aligned}
v_t &= (y - \tfrac{h}{2})/(z - \tfrac{w}{2}sin\theta + \tfrac{l}{2}cos\theta), \\
v_b &= (y + \tfrac{h}{2})/(z - \tfrac{w}{2}sin\theta + \tfrac{l}{2}cos\theta), \\
u_l &= (x + \tfrac{w}{2}cos\theta - \tfrac{l}{2}sin\theta)/(z - \tfrac{w}{2}sin\theta - \tfrac{l}{2}cos\theta), \\
u_p &= (x + \tfrac{w}{2}cos\theta + \tfrac{l}{2}sin\theta)/(z - \tfrac{w}{2}sin\theta + \tfrac{l}{2}cos\theta), \\
u_r &= (x - \tfrac{w}{2}cos\theta + \tfrac{l}{2}sin\theta)/(z + \tfrac{w}{2}sin\theta + \tfrac{l}{2}cos\theta), \\
u_l &= (x - b + \tfrac{w}{2}cos\theta - \tfrac{l}{2}sin\theta)/(z - \tfrac{w}{2}sin\theta - \tfrac{l}{2}cos\theta), \\
u'_r &= (x - b - \tfrac{w}{2}cos\theta + \tfrac{l}{2}sin\theta)/(z + \tfrac{w}{2}sin\theta + \tfrac{l}{2}cos\theta).
\end{aligned} \tag{2}
$$

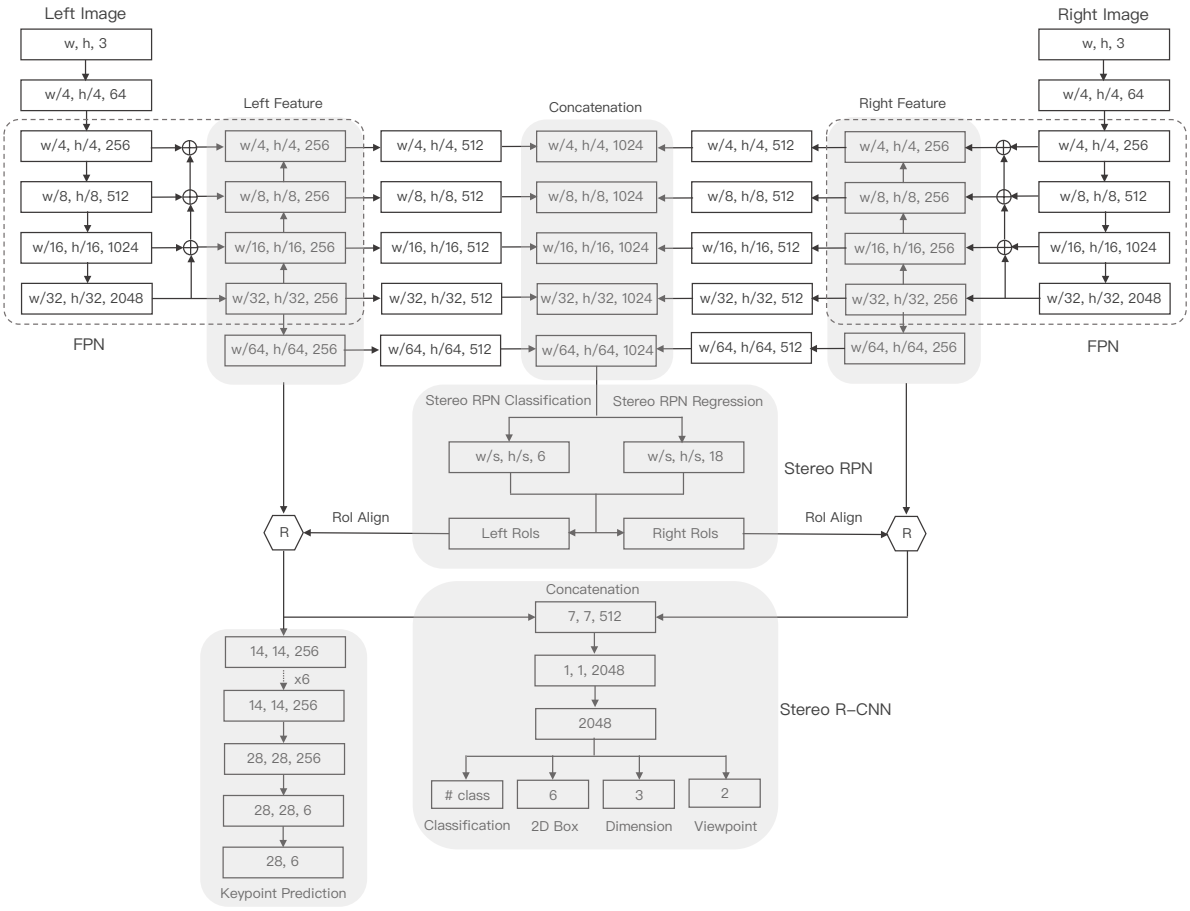

Figure 2: Stereo R-CNN architecture, where we describe the spatial size of each data block. We use five scales s in RPN.

$$
\begin{aligned}
v_t &= (y - \tfrac{h}{2})/(z - \tfrac{w}{2}sin\theta - \tfrac{l}{2}cos\theta), \\
v_b &= (y + \tfrac{h}{2})/(z - \tfrac{w}{2}sin\theta - \tfrac{l}{2}cos\theta), \\
u_l &= (x - \tfrac{w}{2}cos\theta - \tfrac{l}{2}sin\theta)/(z + \tfrac{w}{2}sin\theta - \tfrac{l}{2}cos\theta), \\
u_p &= (x + \tfrac{w}{2}cos\theta - \tfrac{l}{2}sin\theta)/(z - \tfrac{w}{2}sin\theta - \tfrac{l}{2}cos\theta), \\
u_r &= (x + \tfrac{w}{2}cos\theta + \tfrac{l}{2}sin\theta)/(z - \tfrac{w}{2}sin\theta + \tfrac{l}{2}cos\theta), \\
u_l &= (x - b - \tfrac{w}{2}cos\theta - \tfrac{l}{2}sin\theta)/(z + \tfrac{w}{2}sin\theta - \tfrac{l}{2}cos\theta), \\
u'_r &= (x - b + \tfrac{w}{2}cos\theta + \tfrac{l}{2}sin\theta)/(z - \tfrac{w}{2}sin\theta + \tfrac{l}{2}cos\theta).
\end{aligned} \tag{3}
$$

$$
\begin{aligned}
v_t &= (y - \tfrac{h}{2})/(z + \tfrac{w}{2}sin\theta - \tfrac{l}{2}cos\theta), \\
v_b &= (y + \tfrac{h}{2})/(z + \tfrac{w}{2}sin\theta - \tfrac{l}{2}cos\theta), \\
u_l &= (x - \tfrac{w}{2}cos\theta + \tfrac{l}{2}sin\theta)/(z + \tfrac{w}{2}sin\theta + \tfrac{l}{2}cos\theta), \\
u_p &= (x - \tfrac{w}{2}cos\theta - \tfrac{l}{2}sin\theta)/(z + \tfrac{w}{2}sin\theta - \tfrac{l}{2}cos\theta), \\
u_r &= (x + \tfrac{w}{2}cos\theta - \tfrac{l}{2}sin\theta)/(z - \tfrac{w}{2}sin\theta - \tfrac{l}{2}cos\theta), \\
u_l &= (x - b - \tfrac{w}{2}cos\theta + \tfrac{l}{2}sin\theta)/(z + \tfrac{w}{2}sin\theta + \tfrac{l}{2}cos\theta), \\
u'_r &= (x - b + \tfrac{w}{2}cos\theta - \tfrac{l}{2}sin\theta)/(z - \tfrac{w}{2}sin\theta - \tfrac{l}{2}cos\theta).
\end{aligned} \tag{4}
$$

Where the Eq. 1, Eq. 2, Eq. 3, Eq. 4 describe seven constraints for four types of perspective keypoint in Fig. 1 respectively.

3. Additional Experimental Results

In this section, we report additional experimental result for more detailed evaluation. Specifically, We draw the Recall vs IoU overlap and Precision vs Recall (PR) curves corresponding to the Average Recall (AR) and Average Precision (AP) evaluation in our paper. Curves for the 2D detection and association are shown in Fig. 3, where we visualize Recall vs IoU curves for stereo RPN and PR curves for stereo R-CNN in left and right sub-figure respectively. As we can see in Fig. 3, the stereo recall is slightly lower than the recall for the single image. However, the PR curves for left, right, and stereo are almost aligned after R-CNN, which again evidences our consistent detection performance on the left and right image and nearly all true positive detections in the left image have corresponding true-positive right detections.

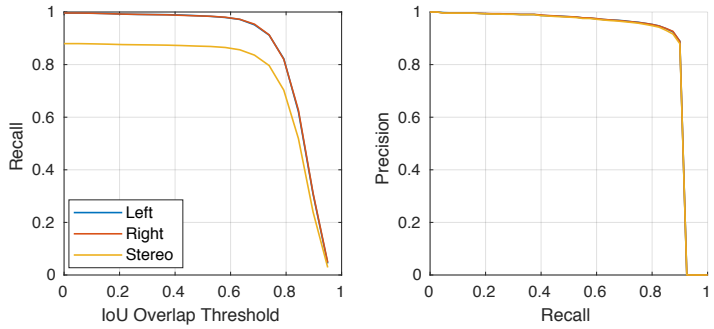

Figure 3: From left to right: Recall vs IoU overlap threshold, 2D Precision vs Recall at IoU threshold of 0.7, evaluated on the *moderate* regimes of the KITTI *validation set*.

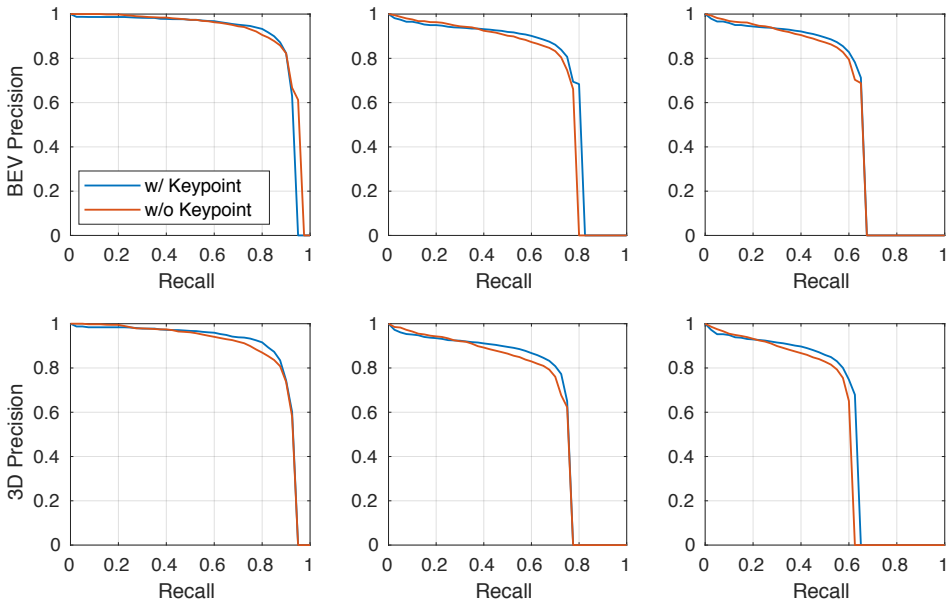

Figure 4: Comparing the effects of using keypoint. From left to right: Precision vs Recall at IoU threshold of 0.5 for the *easy*, *moderate*, and *hard* regimes, evaluated on the the KITTI *validation set*. The top and bottom represents the bird's eye view precision and the 3D box precision respectively.

We also provide PR curves for comparisons of w/ or w/o keypoint, w/ or w/o the 3D alignment in the Fig. 4 and Fig. 5 respectively. Ablation evaluation results for effects of the stereo-flip augmentation and the uncertainty weight [3] are shown in Fig. 6.

We visualize more qualitative examples in Fig. 7.

References

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[3] A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *CVPR*, volume 1, page 4, 2017.

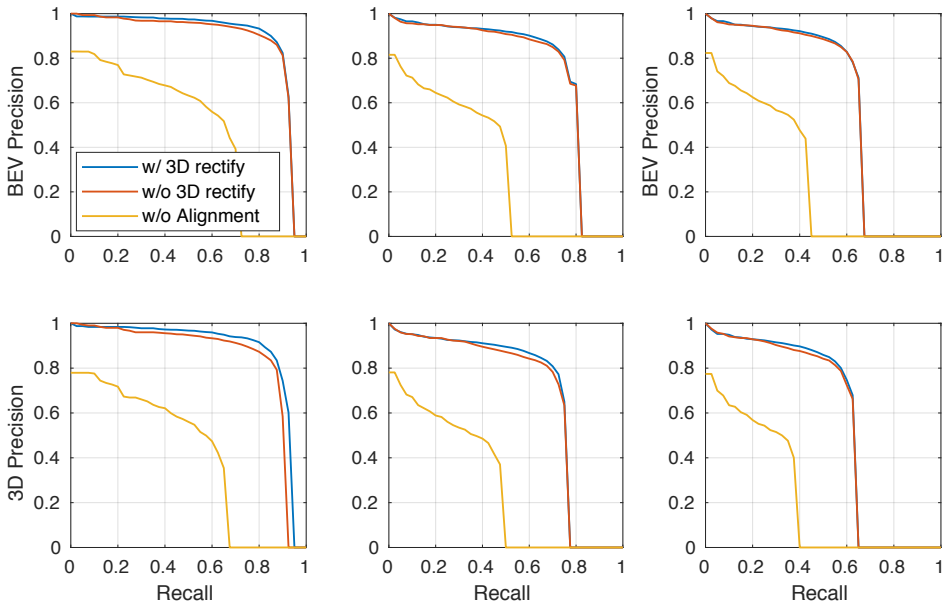

Figure 5: Comparing the effects of using the 3D alignment and the 3D rectify. From left to right: Precision vs Recall at IoU threshold of 0.5 for the *easy*, *moderate*, and *hard* regimes, evaluated on the the KITTI *validation set*. The top and bottom represent the bird's eye view precision and the 3D box precision respectively.

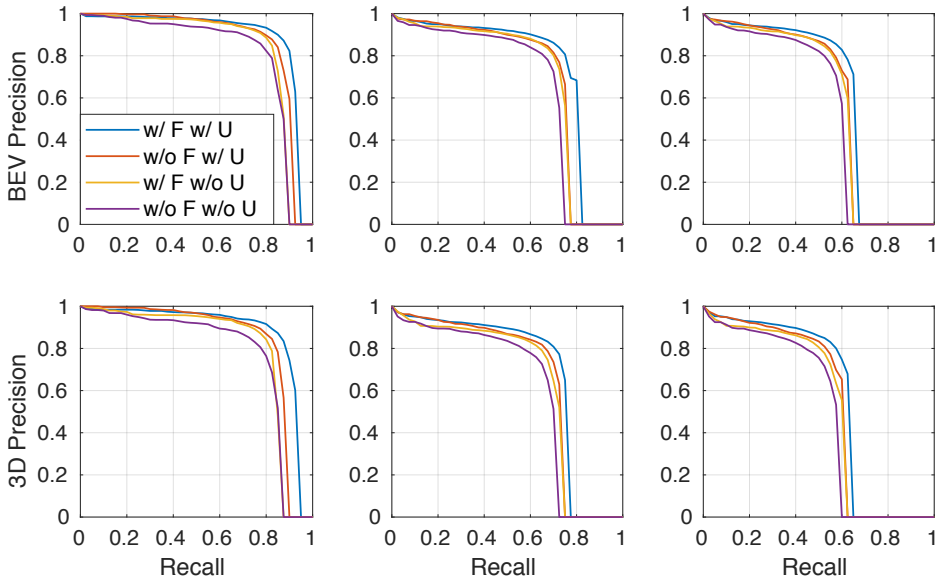

Figure 6: Ablation comparing. We use F and U to denote the stereo-flip augmentation and the uncertainty weight respectively. From left to right: Precision vs Recall at IoU threshold of 0.5 for the *easy*, *moderate*, and *hard* regimes, evaluated on the the KITTI *validation set*. The top and bottom represent the bird's eye view precision and the 3D box precision respectively.

Figure 7: More qualitative results. From top to bottom for every three rows: detections on the left image, right image, and bird's eye view image. Our method provides accurate 3D detection and localization performance even for high occluded and faraway objects.