# MonoEdge: Monocular 3D Object Detection Using Local Perspectives (Appendix)

Minghan Zhu[1*], Lingting Ge[2], Panqu Wang[2], Huei Peng[1]
[1]University of Michigan
[2]TuSimple Inc
minghanz@umich.edu, lingting.ge@tusimple.ai, panqu.wang@tusimple.ai, hpeng@umich.edu

## 1. Depth and yaw from keyedge-ratios

In the paper, we derived the depth $d_{obj}$ and yaw angle $\theta$ from a pair of keyedge-ratios $(r_{ba}, r_{bc})$. The calculation for other keyedge-ratios are similar. Given all four keyedge-ratios $r_{ab}, r_{bc}, r_{cd}, r_{da}$, we first reorganize them into four tuples $(r_{ad}, r_{ab}), (r_{ba}, r_{bc}), (r_{cb}, r_{cd}), (r_{dc}, r_{da})$. Each tuple has a reference keyedge which occurs in both keyedge-ratios (e.g., $a$ in $(r_{ad}, r_{ab})$). We use a common notation $(r_1, r_2)$ to represent the first and second elements of each tuple. Then for each tuple, the solution of depth and yaw has the same form:

$$\theta = \text{arctan2}\ \left(wR_w^\theta, lR_l^\theta\right) \tag{1}$$

$$d_i = \frac{1}{\sqrt{\frac{{R_w^d}^2}{w^2} + \frac{{R_l^d}^2}{l^2}}} \tag{2}$$

where $\theta$ is the yaw angle and $d_i$ is the depth of the reference keyedge. $R_w^\theta$, $R_l^\theta$, $R_w^d$, and $R_l^d$ are placeholders of which the value follows Tab. 1.

Then the depth of the object center is:

$$d_{obj} = d_i + \frac{1}{2}\Delta_d \tag{3}$$

where the value of $\Delta_d$ is listed in Tab. 2.

As discussed in the paper, the depth and yaw derived from keyedge-ratios only depends on the pair of the keyedge-ratios and the physical length $l$ and width $w$ of the object.

## 2. Camera-centric indexing

With camera-centric indexing, index 1 is always assigned to the keyedge with shortest distance to the camera center. Notice that the distance is not the depth (which is the projected distance in the front direction). When the keyedge with the shortest distance also has the smallest depth, all four keyedge-ratios $[r_{21}, r_{41}, r_{32}, r_{34}]$ are equal or smaller than 1. Otherwise, the keyedge-ratios can go slightly larger than 1. In practice, the keyedge with shortest depth has shortest distance for most objects in our tested datasets. We use the distance for camera-centric indexing because it changes accordingly with the allocentric angle and visible faces of an object, which largely affects the appearance of the object in an image.

| reference keyedge | $R_w^\theta$ | $R_l^\theta$ | $R_w^d$ | $R_l^d$ |
|---|---|---|---|---|
| $a$ | $r_1 - 1$ | $-(r_2 - 1)$ | $r_2 - 1$ | $r_1 - 1$ |
| $b$ | $r_2 - 1$ | $r_1 - 1$ | $r_1 - 1$ | $r_2 - 1$ |
| $c$ | $-(r_1 - 1)$ | $r_2 - 1$ | $r_2 - 1$ | $r_1 - 1$ |
| $d$ | $-(r_2 - 1)$ | $-(r_1 - 1)$ | $r_1 - 1$ | $r_2 - 1$ |

Table 1. Value of the placeholders in Eq. (1) and Eq. (2) for each tuple of keyedge-ratios.

| reference keyedge | $\Delta_d$ |
|---|---|
| $a$ | $l \sin\theta - w \cos\theta$ |
| $b$ | $l \sin\theta + w \cos\theta$ |
| $c$ | $-l \sin\theta + w \cos\theta$ |
| $d$ | $-l \sin\theta - w \cos\theta$ |

Table 2. Value of $\Delta_d$ in Eq. (3) for each tuple of keyedge-ratios.

## 3. More details on the experimented networks

Given that our proposed local perspective module can be plugged-in to various network structures, here we give more details on the regressed variables and corresponding loss functions on the experimented networks outside of the local perspective module, so that readers have a better idea of the overall architecture.

### 3.1. MonoFlex [1]

The regressed variables are:

- Classification score (focal loss [2])

[*] Work done during internship at TuSimple Inc.

- 2D bounding box (GIoU loss [3])
- 2D projected center (L1 loss)
- Projected keypoints (L1 loss)
- Physical size (L1 loss)
- Depth and its uncertainty (Uncertainty-aware loss, same form as Eq. (9))

The size of the regression head of the local perspective module follows the design of other regression heads of this network, i.e., 1 FC layer with 256 dimensional features.

### 3.2. MonoRCNN [4]

The regressed variables are:

- Classification score (cross entropy loss)
- 2D bounding box (L1 loss)
- 2D projected center (L1 loss)
- Projected keypoints (L1 loss)
- Physical size (L1 loss)
- Physical height and inverse visual height and their uncertainty for depth estimation (Uncertainty-aware loss, same form as Eq. (9))

The size of the regression head of the local perspective module follows the design of other regression heads of this network, i.e., 2 FC layer with 1024 dimensional features.

### 3.3. PGD [5]

The regressed variables are:

- Classification score (cross entropy loss)
- Centerness (BCE loss)
- 2D projected center (smooth L1 loss)
- 2D bounding box (smooth L1 loss)
- Physical size (smooth L1 loss)
- Velocity (smooth L1 loss)
- Depth and its weight (smooth L1 loss)

The size of the regression head of the local perspective module follows the design of other regression heads of this network, i.e., 1 FC layer with 256 dimensional features.

## References

[1] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3289–3298, 2021.

[2] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[3] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[4] Xuepeng Shi, Qi Ye, Xiaozhi Chen, Chuangrong Chen, Zhixiang Chen, and Tae-Kyun Kim. Geometry-based distance decomposition for monocular 3d object detection. *arXiv preprint arXiv:2104.03775*, 2021.

[5] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. *arXiv preprint arXiv:2107.14160*, 2021.