FACULTY OF MATHEMATICS,
PHYSICS AND INFORMATICS
Comenius University
Bratislava

3D Vision

Lecture 9: Monocular Depth Estimation and 3D Object Detection

Ing. Viktor Kocur, PhD.

25.4.2023

Contents

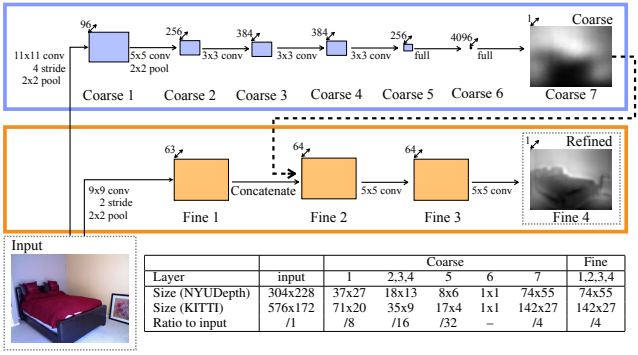
Monocular Depth Estimation

The goal of monocular depth estimation is to take a single RGB image and output a depth image for the given scene. This task lends itself very well to deep learning based approaches as the geometry can be inferred from learned structures. We will briefly cover some interesting approaches to this problem.

The most significant way in which we can divide the methods is based on whether they need depth annotations or not. Based on this we can divide them into

- Supervised methods - require pairs of corresponding RGB and depth data
- Semi-supervised methods - require some other information - usually two stereo RGB images of the same scene
- Unsupervised methods - usually utilize sequences of RGB images

Simple Supervised Approach

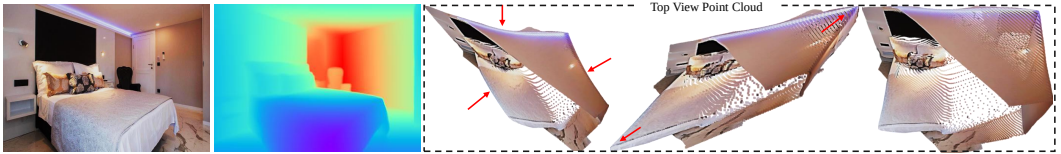


		Coarse					Fine
Layer	input	1	2,3,4	5	6	7	1,2,3,4
Size (NYUDepth)	304x228	37x27	18x13	8x6	1x1	74x55	74x55
Size (KITTI)	576x172	71x20	35x9	17x4	1x1	142x27	142x27
Ratio to input	/1	/8	/16	/32	–	/4	/4

$$L(y, y^*) = \frac{1}{n}\sum_i y_i^2 - \frac{\lambda}{n^2}\left(\sum_i y_i\right)^2 \qquad y_i = \log d_i - \log d_i^* \tag{1}$$

Image adopted from: David Eigen, Christian Puhrsch, and Rob Fergus. “Depth map prediction from a single image using a multi-scale deep network.” In: *Advances in neural information processing systems* 27 (2014)

Recovering Focal Length



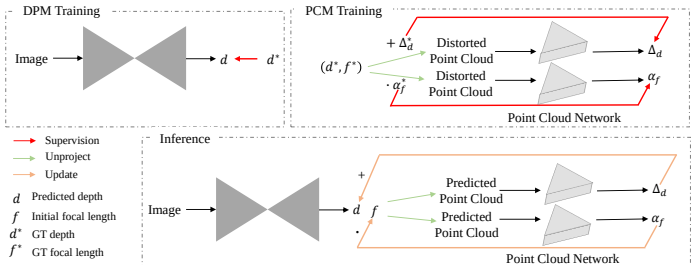


Image adopted from: Wei Yin et al. “Learning to recover 3d scene shape from a single image.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 204–213

Adversarial Training

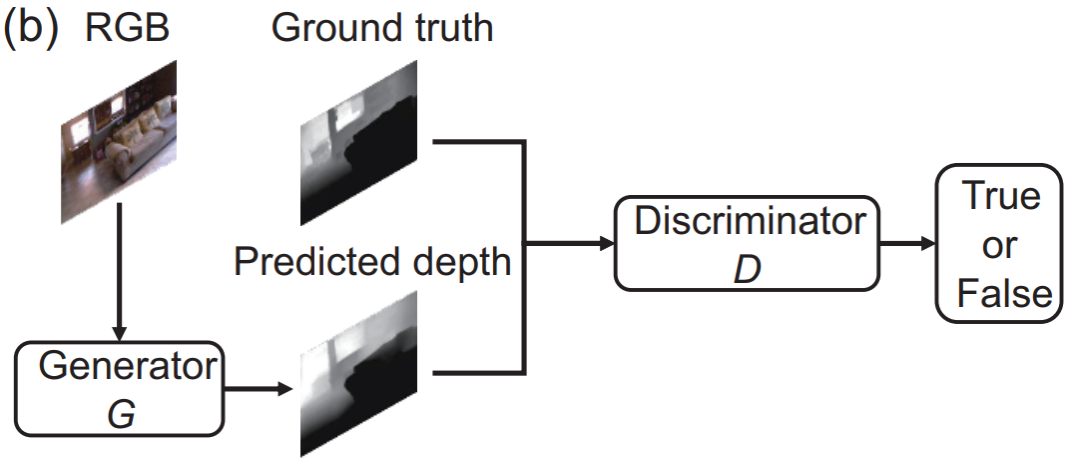


Image adopted from: Chaoqiang Zhao et al. “Monocular depth estimation based on deep learning: An overview.” In: *Science China Technological Sciences* 63.9 (2020), pp. 1612–1627

Semi-supervised Training

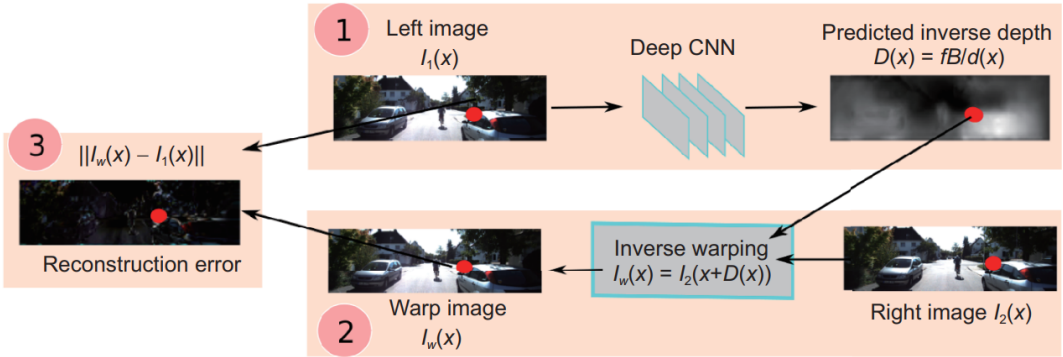


Image adopted from: Chaoqiang Zhao et al. “Monocular depth estimation based on deep learning: An overview.” In: *Science China Technological Sciences* 63.9 (2020), pp. 1612–1627

Sparsely- and Semi-supervised Training

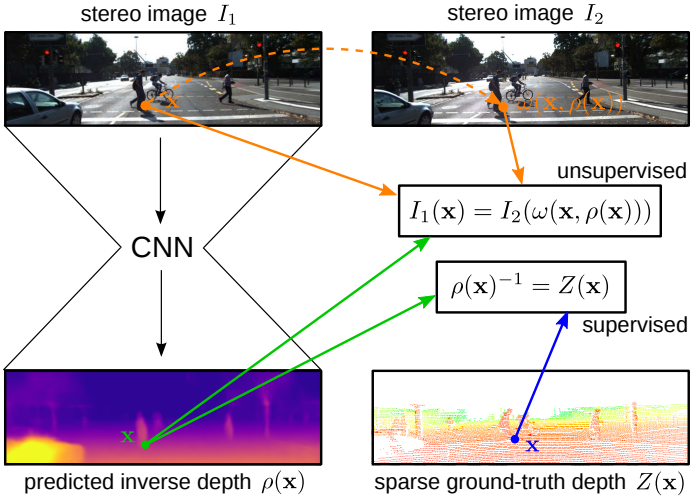


Image adopted from: Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. “Semi-supervised deep learning for monocular depth map prediction.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 6647–6655

Adversarial Semi-supervised Training

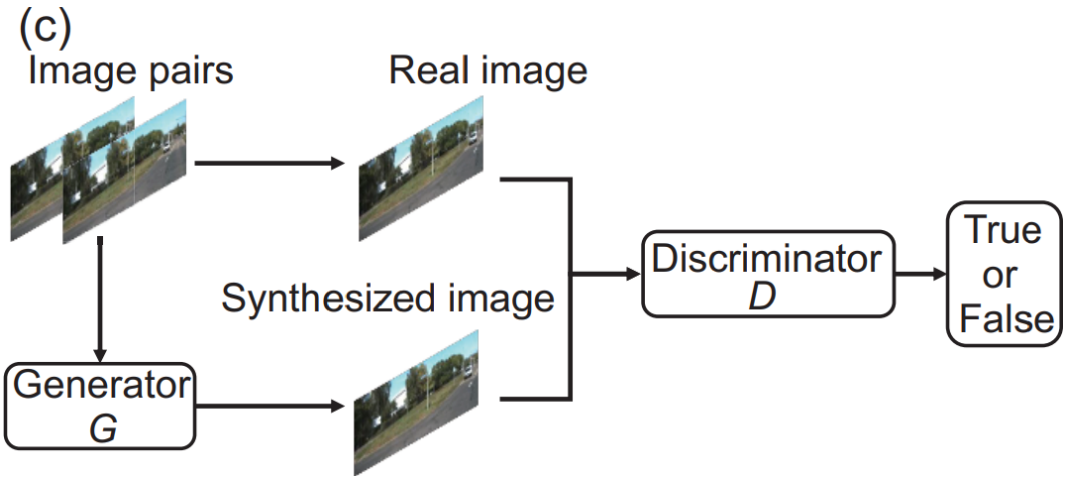


Image adopted from: Chaoqiang Zhao et al. “Monocular depth estimation based on deep learning: An overview.” In: *Science China Technological Sciences* 63.9 (2020), pp. 1612–1627

Unsupervised Training

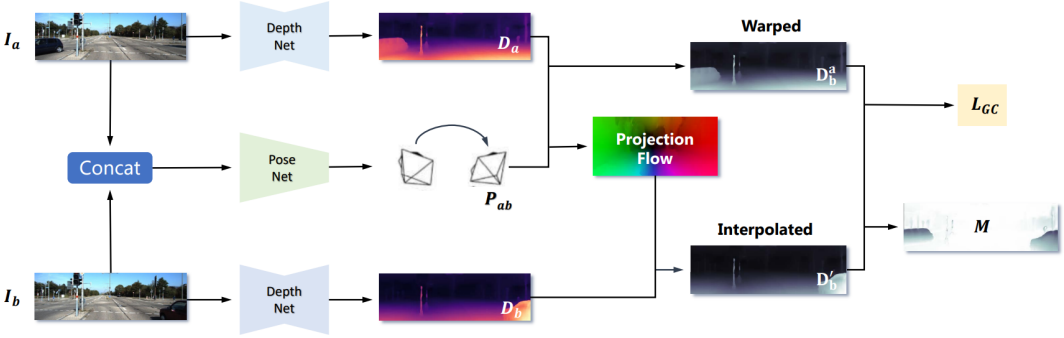


Image adopted from: Jiawang Bian et al. “Unsupervised scale-consistent depth and ego-motion learning from monocular video.” In: *Advances in neural information processing systems* 32 (2019)

Object Detection

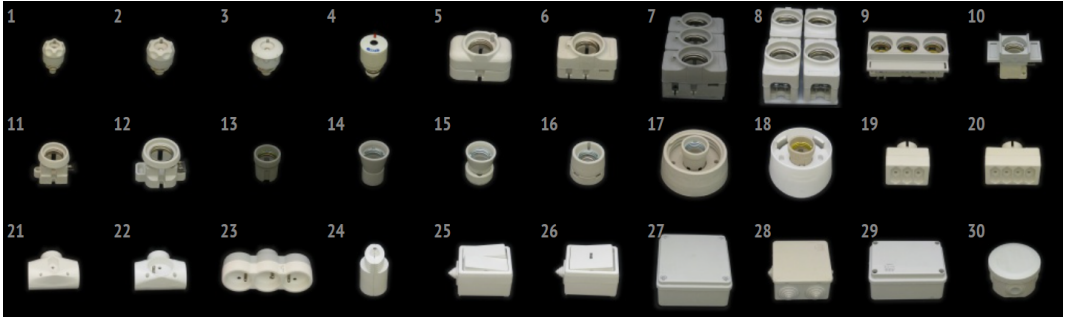
The 3D object detection task has more variants than the standard 2D task on RGB images. In general the goal of the task is to detect 3D positions of objects. Generally there are several different types of this task:

- 6 DoF pose estimation - detect the positions (R and $\boldsymbol{t}$) of a known object
- Categorical object detection - detect a 3D bounding box around an object of a given category
- Bird's eye view detection - detect a 2D bounding boxes of objects in a bird's eye view of the scene

We can also consider multiple different types of input:

- RGB images - one (monocular) or more
- RGB-D or depth maps
- Pointclouds

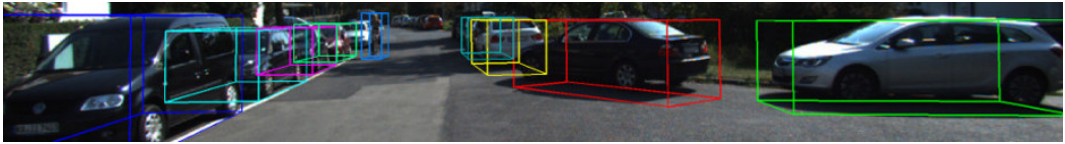
6 DoF Pose Estimation



For this task we may have a CAD model available. In some cases the method assumes that the object will be recorded from multiple views prior to detection.

Image adopted from: Tomáš Hodan et al. “T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects.” In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2017, pp. 880–888

Categorical 3D Object Detection



Each bounding box is defined by its width, height and depth, translation and rotation. In some cases the rotation may be limited fewer than 3 DoF (as in the dataset shown above). There can be multiple categories. There is some ambiguity w.r.t. the correct rotation of some objects.

Image adopted from: Andreas Geiger, Philip Lenz, and Raquel Urtasun. “Are we ready for autonomous driving? the kitti vision benchmark suite.” In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 3354–3361

Annotating 3D Objects

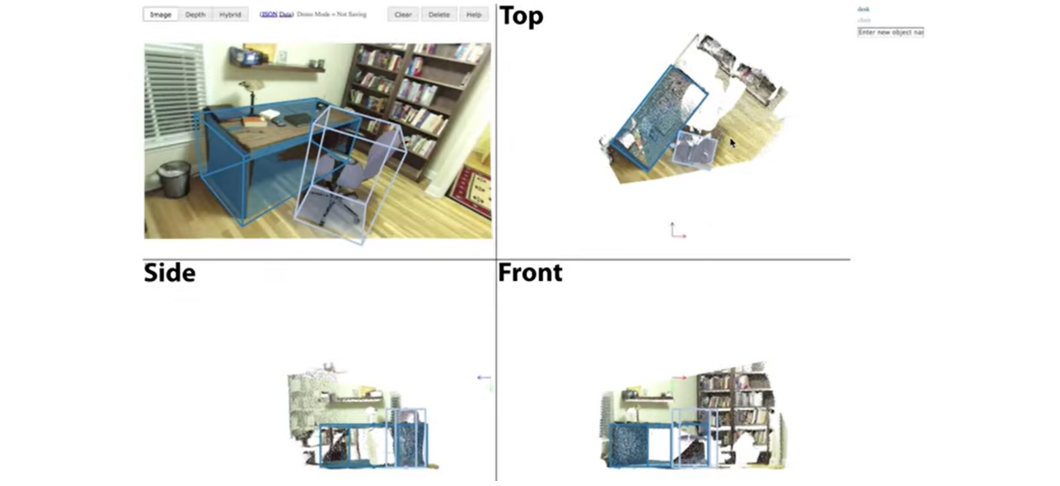


Image adopted from: Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. “Sun rgb-d: A rgb-d scene understanding benchmark suite.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 567–576

Standard SfM vs Deep Learning

If we have a sequence of images with a given object we can obtain its sparse reconstruction. We can then use this to detect the object. We simply do this by finding the R and $\boldsymbol{t}$ of the object instead of the whole scene using a standard PnP algorithm.

Note that there may still be an issue with the R and $\boldsymbol{t}$ based on the sparse model, which may be aligned in an arbitrary manner (e.g. coordinates of the first view used for reconstruction)

Regressing Parameters

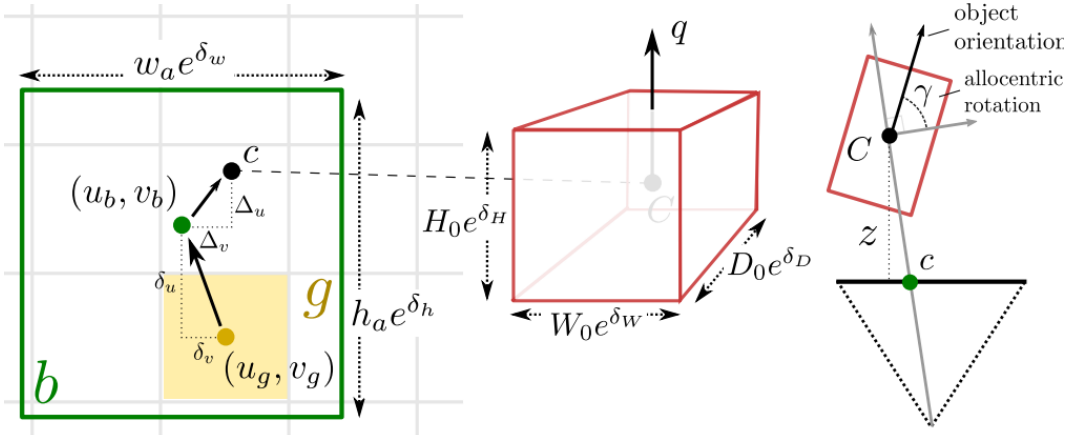
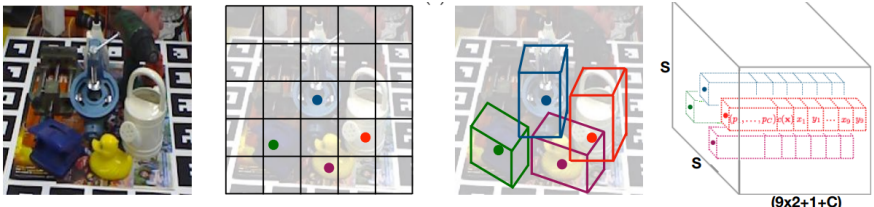


Image adopted from: Andrea Simonelli et al. “Disentangling monocular 3d object detection.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 1991–1999

Corners + PnP



(b) (c) (d) (e)

Figure 1. Overview: (a) The proposed CNN architecture. (b) An example input image with four objects. (c) The $S \times S$ grid showing cells responsible for detecting the four objects. (d) Each cell predicts 2D locations of the corners of the projected 3D bounding boxes in the image. (e) The 3D output tensor from our network, which represents for each cell a vector consisting of the 2D corner locations, the class probabilities and a confidence value associated with the prediction.

Image adopted from: Bugra Tekin, Sudipta N Sinha, and Pascal Fua. “Real-time seamless single shot 6d object pose prediction.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 292–301

3D Location of Points + PnP

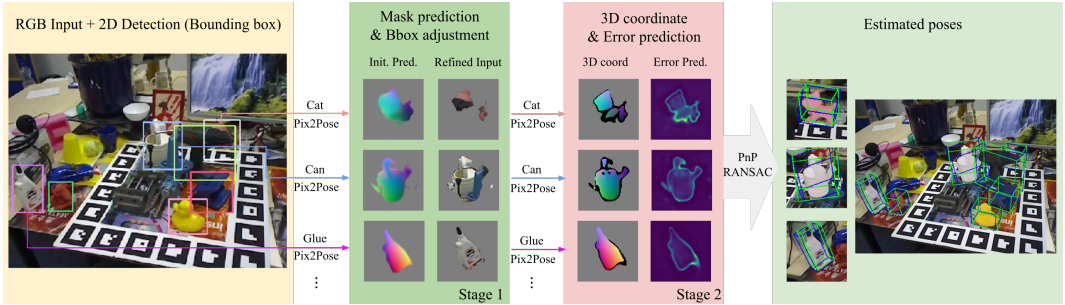


Image adopted from: Kiru Park, Timothy Patten, and Markus Vincze. “Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 7668–7677

Dealing with Symmetries

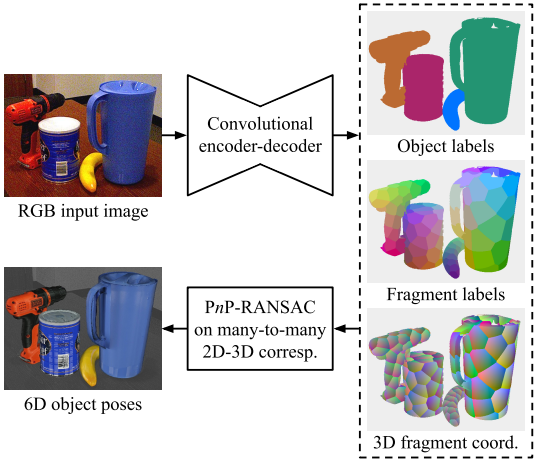


Image adopted from: Tomas Hodan, Daniel Barath, and Jiri Matas. “Epos: Estimating 6d pose of objects with symmetries.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 11703–11712

Faster 2D-3D Correspondences

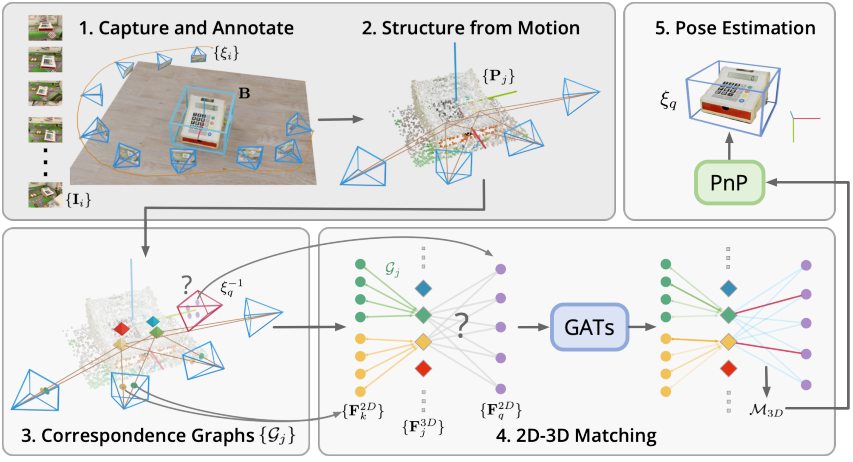


Image adopted from: Jiaming Sun et al. “Onepose: One-shot object pose estimation without cad models.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 6825–6834

Regressing R and $\boldsymbol{t}$

We may also be interested in regressing R and $\boldsymbol{t}$ for the objects. Regressing $\boldsymbol{t}$ is usually straightforward, but working with R has several problems.

The fundamental issue with regressing R from a neural network is that it has 3 DoF, but there is not *continuous* 3-dimensional representation for $\mathrm{SO}(3)$.

Representing R

If we output simply R out of the network we would have problem in that R won't be in $\mathrm{SO}(3)$. We could project a matrix $\hat{R}$ from a network onto $\mathrm{SO}(3)$ (Procrustes problem), but that is not very common.

Some other approaches include:

- Three (Euler) angles - problems with periodicity, can be solved by binning - regression turns into classification
- Quaternions - both q and $-q$ represent the same rotation, discontinuity near small rotations
- Rodrigues' vector - problem with periodicity, $-\boldsymbol{v}$ vs. $\boldsymbol{v}$, discontinuity near small rotations

Two Vector Approach

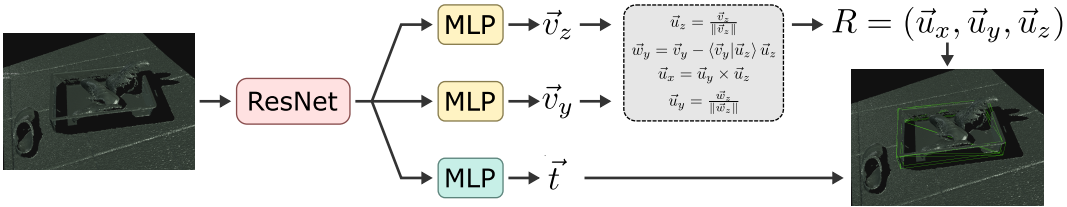


Image adopted from: Lukáš Gajdošech et al. “Towards Deep Learning-based 6D Bin Pose Estimation in 3D Scans.” In: *arXiv preprint arXiv:2112.09598* (2021)

Symmetries and R

Symmetric objects pose a challenge. Let us consider rotations $R_s \in S$ such that rotating the object by R_s keeps the object pose equivalent. In other words $R \sim R_s R$. Note that $I \in S$. Such symmetries can be dealt with by using a loss of the form:

$$\hat{L} = \min_{R_s \in S} L(R_s R, R_{gt}). \tag{2}$$

This loss may have multiple minima or pleateau areas. This can be problematic in terms of training dynamics. It is possible to deal with this directly via paramterization of R or splitting the output to multiple regressors and a classifier.

Frustum PointNet

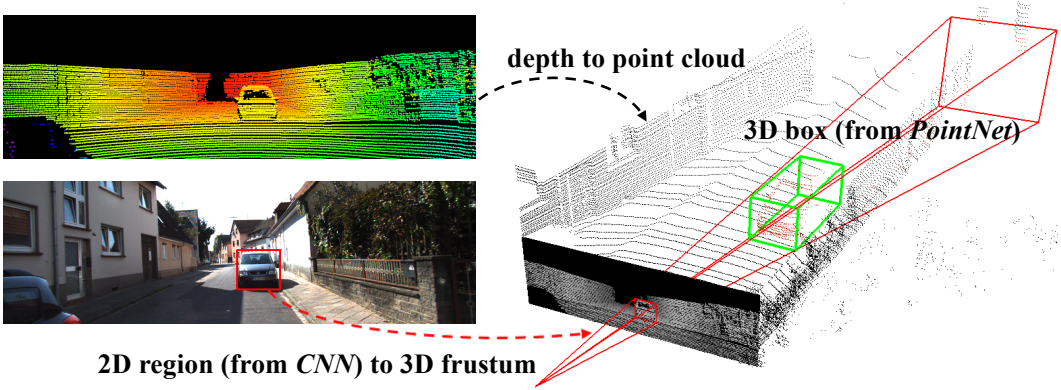


Image adopted from: Charles R Qi et al. “Frustum pointnets for 3d object detection from rgb-d data.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 918–927

Frustum PointNet

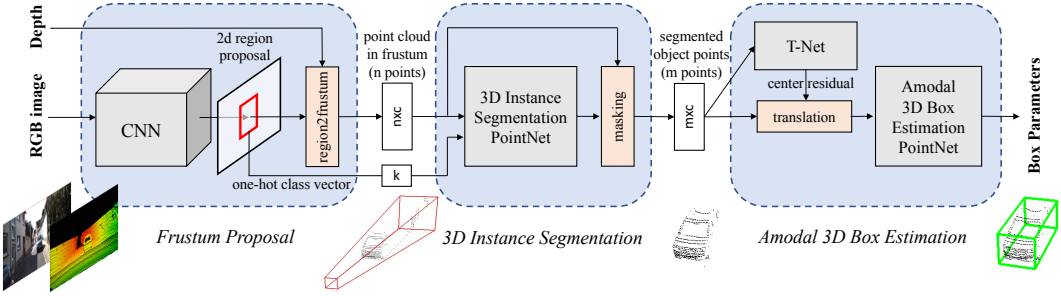


Image adopted from: Charles R Qi et al. “Frustum pointnets for 3d object detection from rgb-d data.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 918–927

Point Pillars

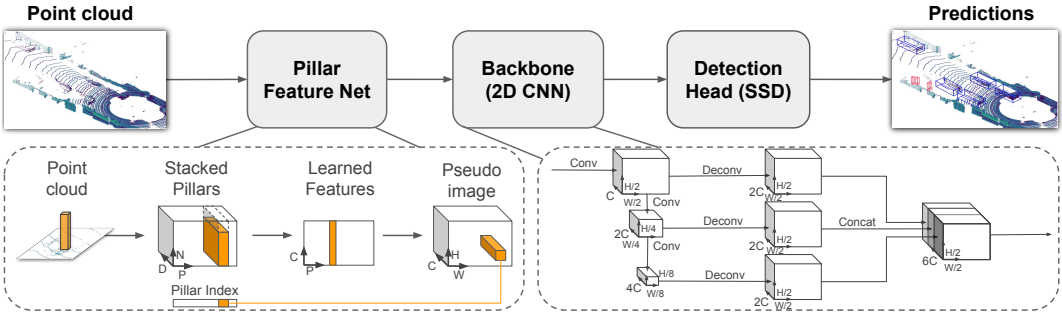


Image adopted from: Alex H Lang et al. “Pointpillars: Fast encoders for object detection from point clouds.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 12697–12705

Pseudolidar

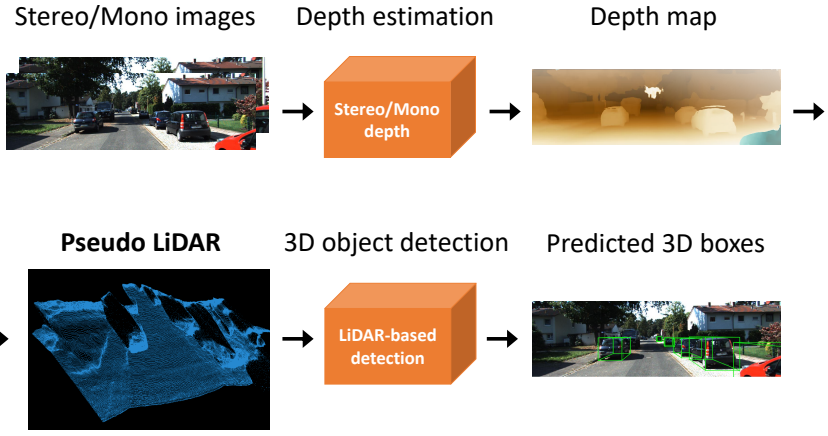


Image adopted from: Yan Wang et al. “Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 8445–8453

Pseudolidar - differentiable

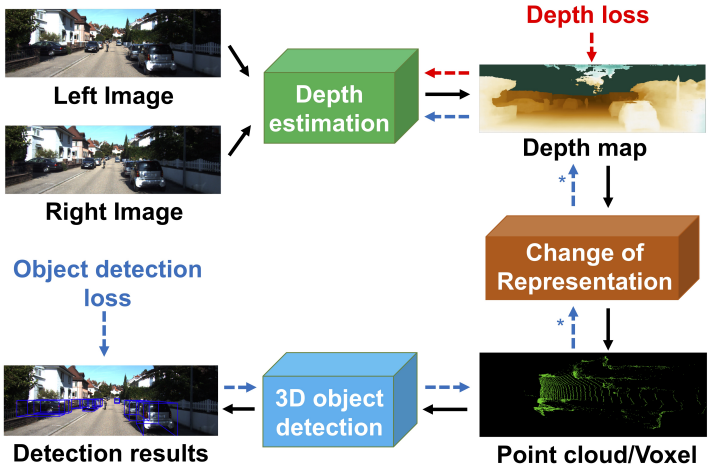


Image adopted from: Rui Qian et al. “End-to-end pseudo-lidar for image-based 3d object detection.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 5881–5890

Depth-aware Transformers

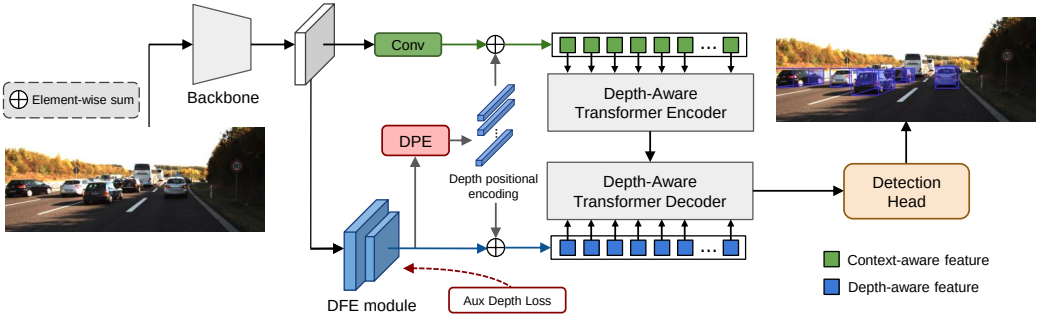


Image adopted from: Kuan-Chih Huang et al. “Monodtr: Monocular 3d object detection with depth-aware transformer.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 4012–4021

Sensor Fusion

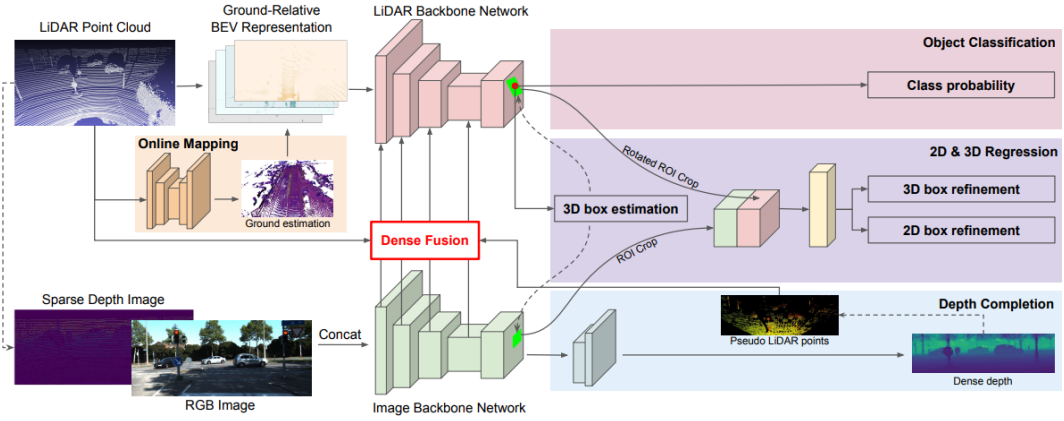


Image adopted from: Ming Liang et al. “Multi-task multi-sensor fusion for 3d object detection.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 7345–7353