Chapter 2 - Section 9

3D Vision and Applications

Dr. Li Hongyang

Friday, April 22, 2022

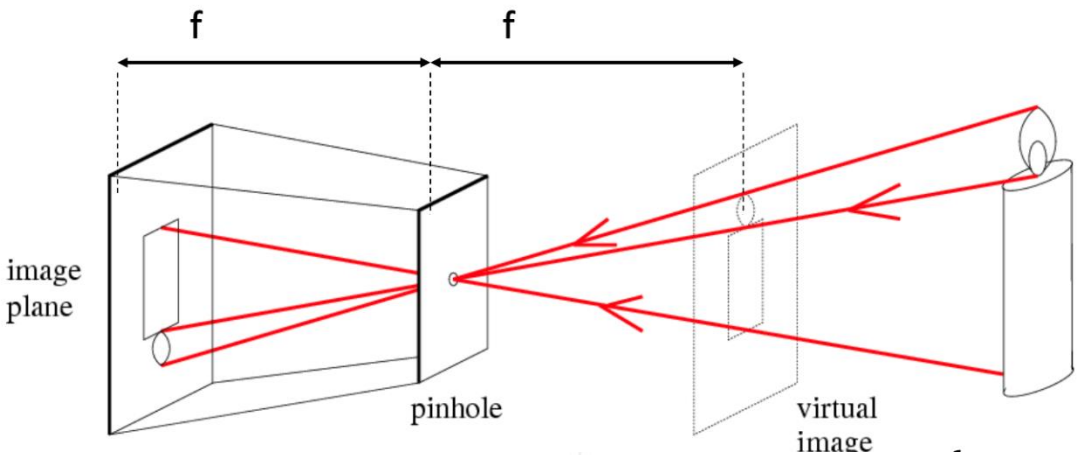
商汤 sensetime

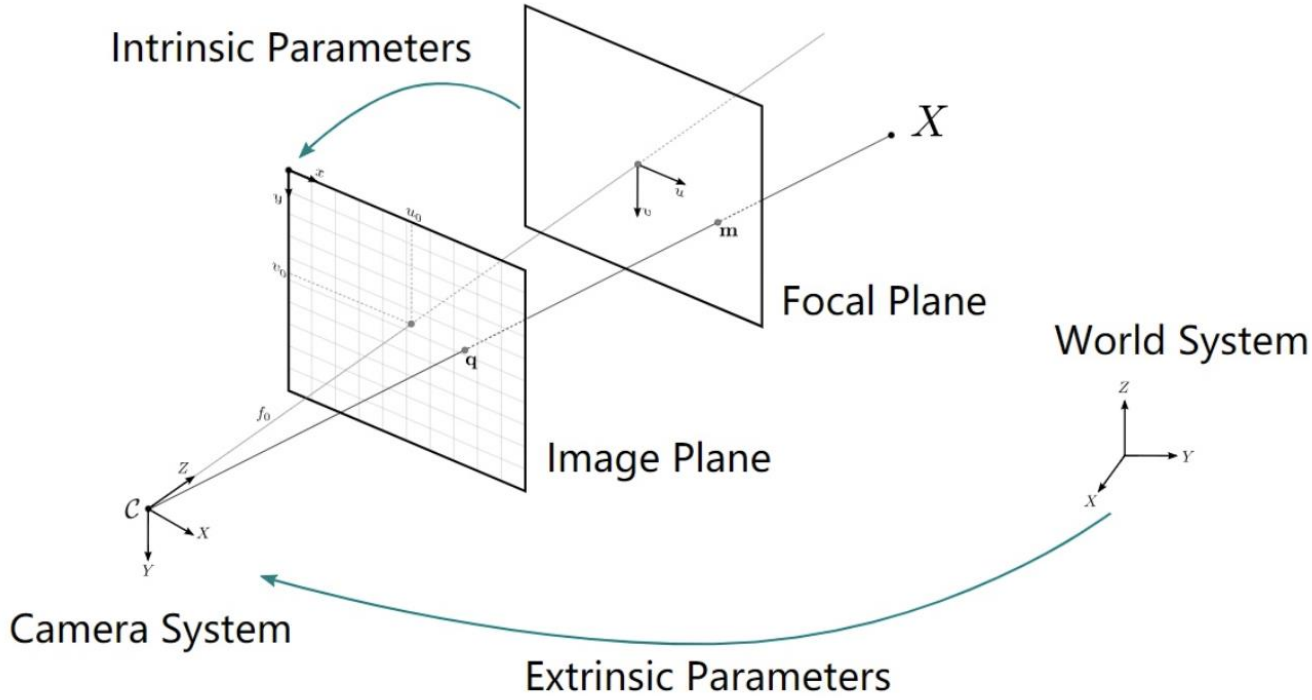
清華大學 Tsinghua University

Outline

Prerequisite: camera model/intrinsics/extrinsics

- The pinhole model





https://zhuanlan.zhihu.com/p/54139614

Prerequisite: camera model/intrinsics/extrinsics

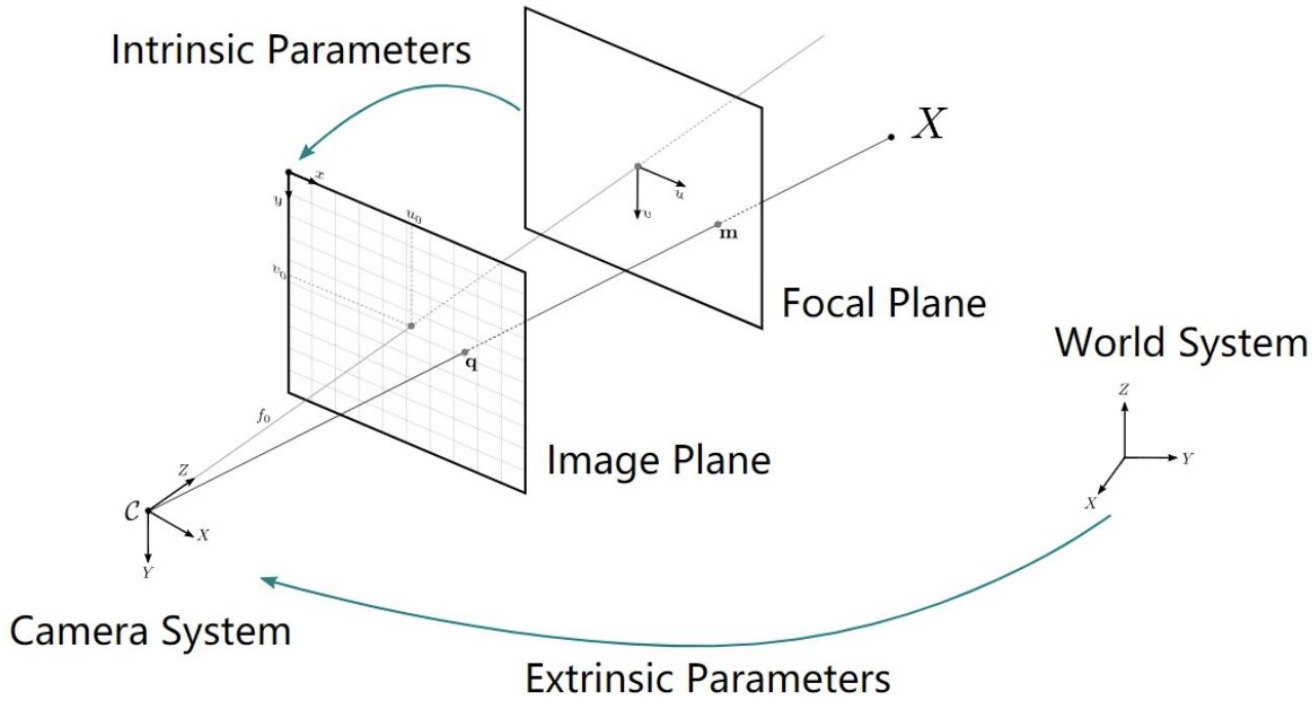
- The pinhole model

世界坐标系(world coordinate system)：用户定义的三维世界的坐标系，为了描述目标物在真实世界里的位置而被引入。单位为m。

相机坐标系(camera coordinate system)：在相机上建立的坐标系，为了从相机的角度描述物体位置而定义，作为沟通世界坐标系和图像/像素坐标系的中间一环。单位为m。

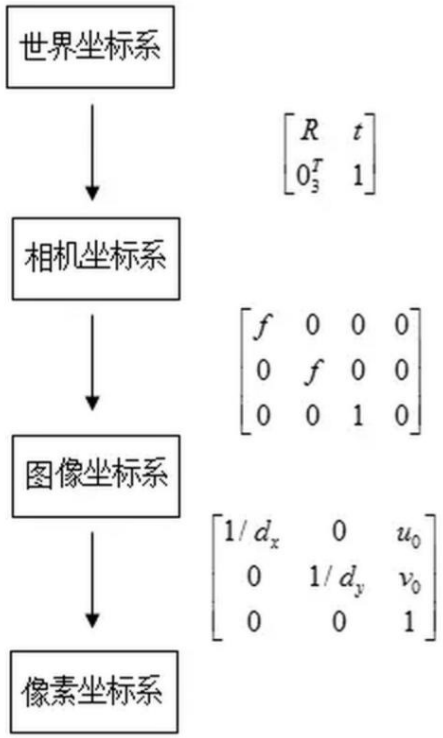
图像坐标系(image coordinate system)：为了描述成像过程中物体从相机坐标系到图像坐标系的投影透射关系而引入，方便进一步得到像素坐标系下的坐标。 单位为m。

像素坐标系(pixel coordinate system)：为了描述物体成像后的像点在数字图像上（相片）的坐标而引入，是我们真正从相机内读取到的信息所在的坐标系。单位为个（像素数目）。



Prerequisite: camera model/intrinsics/extrinsics

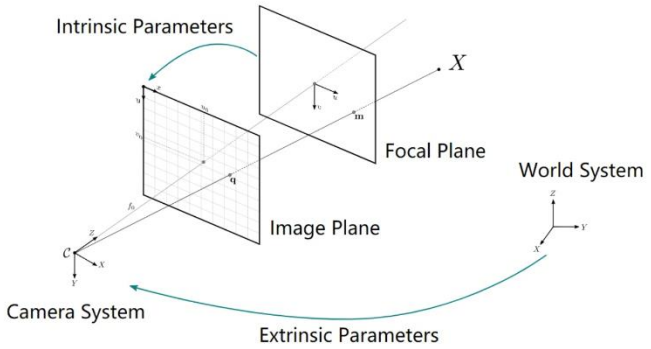
- Camera model



像素坐标和世界坐标的关系：

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = s \begin{bmatrix} f_x & \gamma & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_1 & r_2 & t \end{bmatrix} \begin{bmatrix} x_W \\ y_W \\ 1 \end{bmatrix}$$

u、v表示像素坐标系中的坐标，s表示尺度因子，fx、fy、u0、v0、γ（由于制造误差产生的两个坐标轴偏斜参数，通常很小）表示5个相机内参，R,t表示相机外参，Xw、Yw、Zw（假设标定棋盘位于世界坐标系中Zw=0的平面）表示世界坐标系中的坐标。



Prerequisite: camera model/intrinsics/extrinsics

- Camera model – H matrix **它同时包含了相机内参和外参。**

像素坐标和世界坐标的关系：

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = s \begin{bmatrix} f_x & \gamma & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_1 & r_2 & t \end{bmatrix} \begin{bmatrix} x_W \\ y_W \\ 1 \end{bmatrix}$$

单应性（Homography）变换。可以简单的理解为它用来**描述物体在世界坐标系和像素坐标系之间的位置映射关系。**对应的变换矩阵称为**单应性矩阵。**在上述式子中，单应性矩阵定义为：

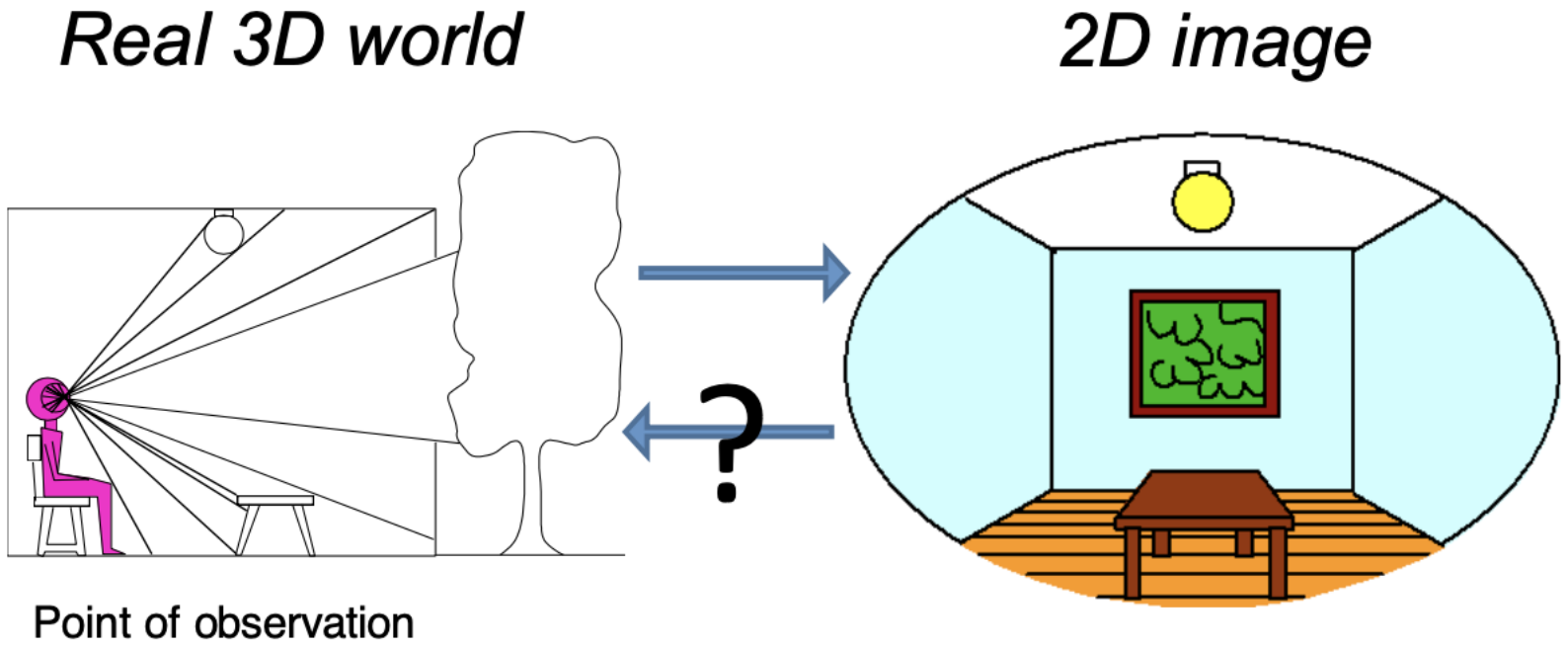
$$H = s \begin{bmatrix} f_x & \gamma & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_1 & r_2 & t \end{bmatrix} = sM \begin{bmatrix} r_1 & r_2 & t \end{bmatrix}$$

其中M是内参矩阵

- H矩阵在计算机视觉中有很广泛的应用
- 如何估算H？相机标定
- 张正友标定法
- **这里从略**

Recover 3D from images

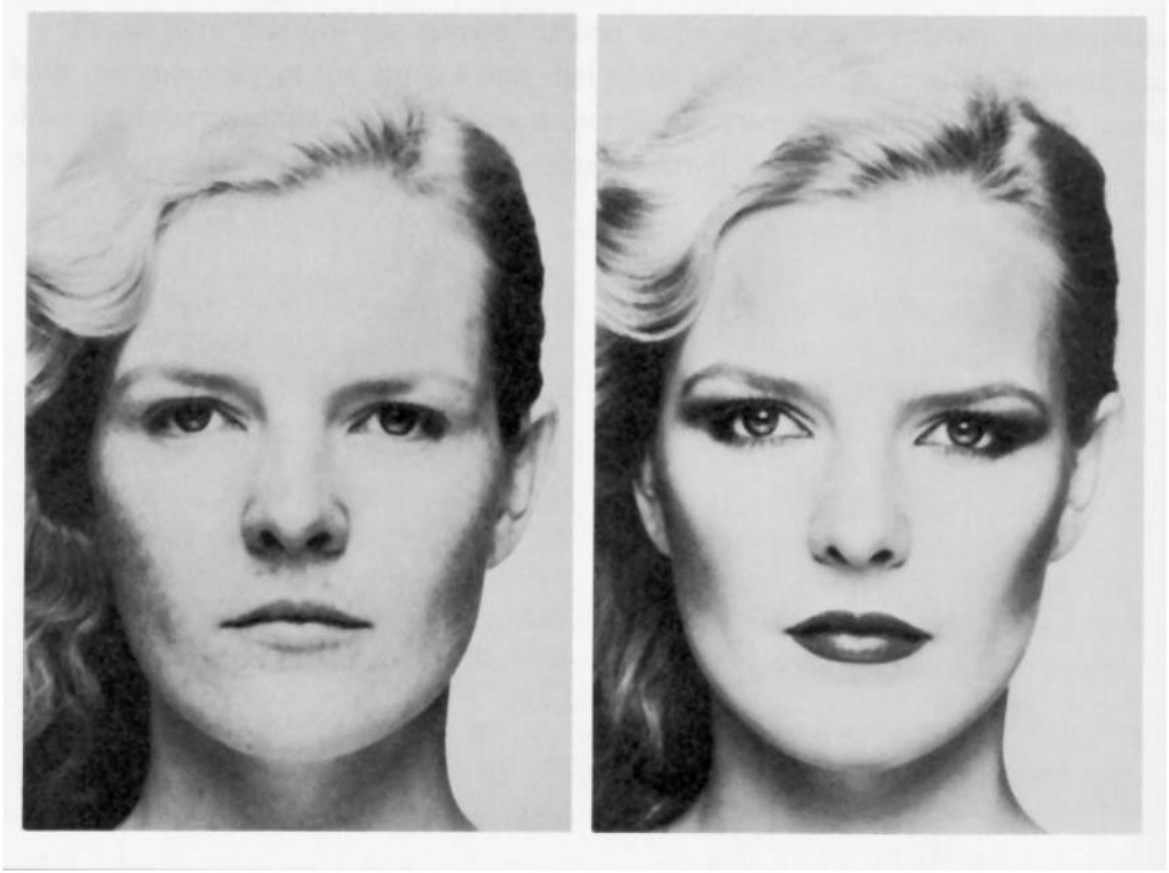
How can we automatically compute 3D geometry from images?
– What cues in the image provide 3D information?



Credit: Fei-fei Li

Recover 3D from images: cues for 3D

Shading

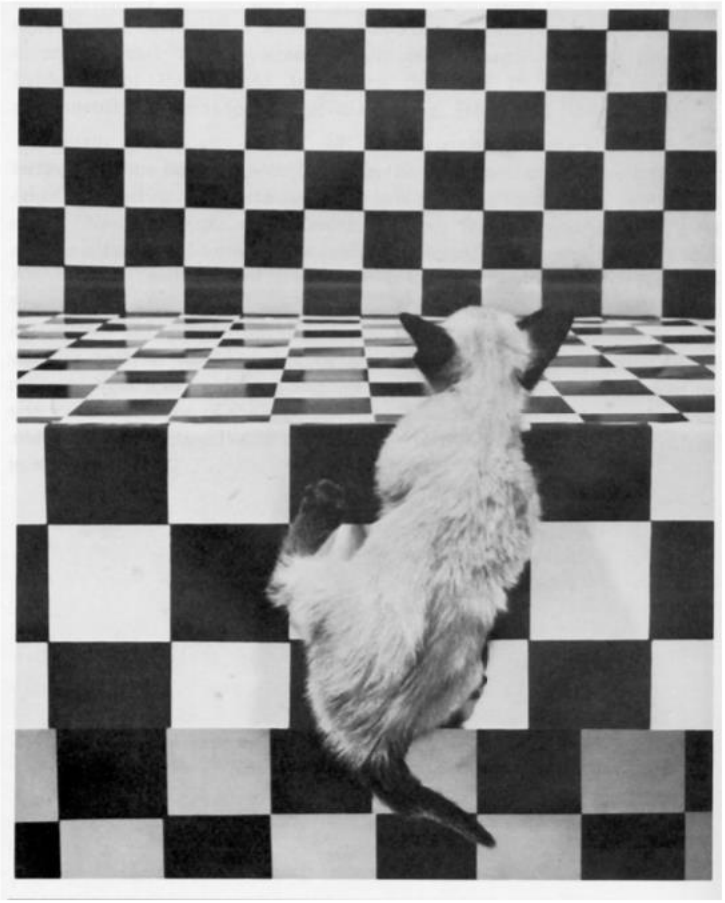


Merle Norman Cosmetics, Los Angeles

Credit: Fei-fei Li

Recover 3D from images: cues for 3D

- Shading
- Texture



The Visual Cliff, by William Vandivert, 1960

Credit: Fei-fei Li

Recover 3D from images: cues for 3D

- Shading
- Texture
- Focus

From *The Art of Photography*, Canon

Credit: Fei-fei Li

Recover 3D from images: cues for 3D

- Shading
- Texture
- Focus
- **Motion**

Credit: Fei-fei Li

Recover 3D from images: cues for 3D

- Shading
- Texture
- Focus
- Motion

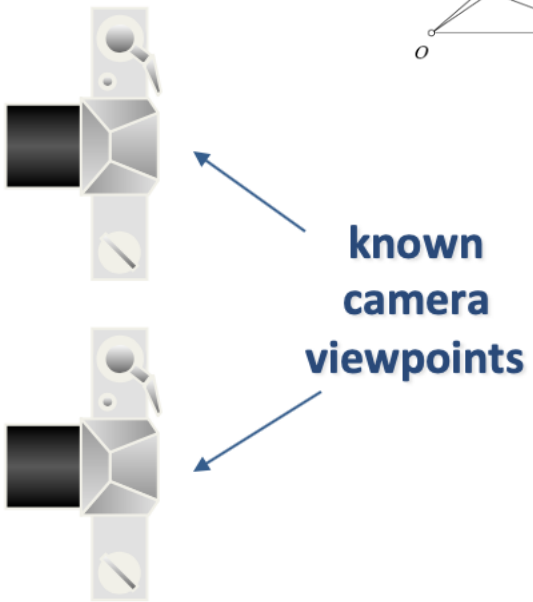
- Others:
  - Highlights
  - Shadows
  - Silhouettes
  - Inter-reflections
  - Symmetry
  - Light Polarization
  - …

Shape From X

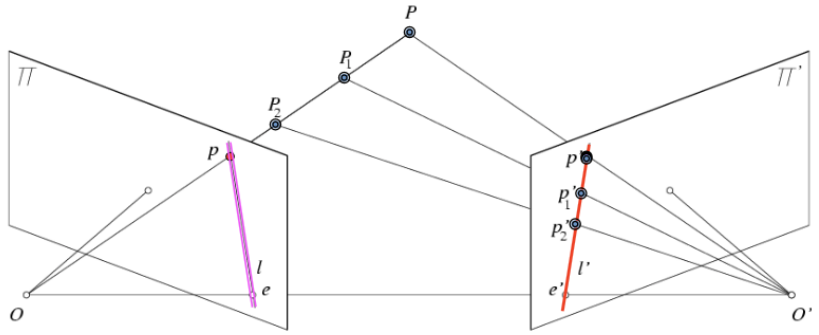
- X = shading, texture, focus, motion, …
- We'll focus on the motion cue

Credit: Fei-fei Li

The Stereo Problem
– Shape from two (or more) images
– Biological motivation



Epipolar Constraint



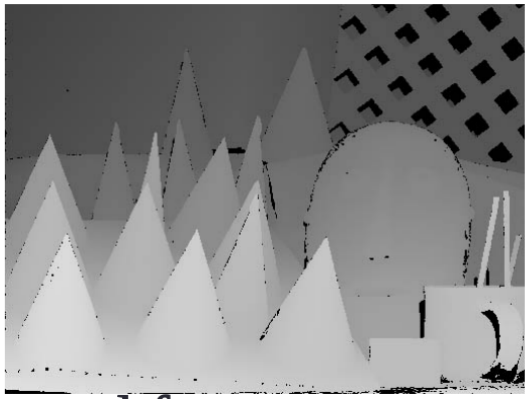
Credit: Fei-fei Li

- Comparison: 2D vs 3D

Analysis Tools	2D	3D
Representation	Image (u,v)	• Depth image (u,v,d) • Point cloud (x,y,z)
1st order differential geometry	Image gradients	Surface normals
2nd order differential geometry	Second moment matrix	Principle curvature
Corner detection	Harris image	Surface variation
Feature extraction	HOG	• Point Feature Histograms • Spin Images
Geometric model fitting	Hough transform	Clustering + RANSAC
Alignment	SSD window filter	Iterative Closest Point (ICP)

Credit: Hawkins

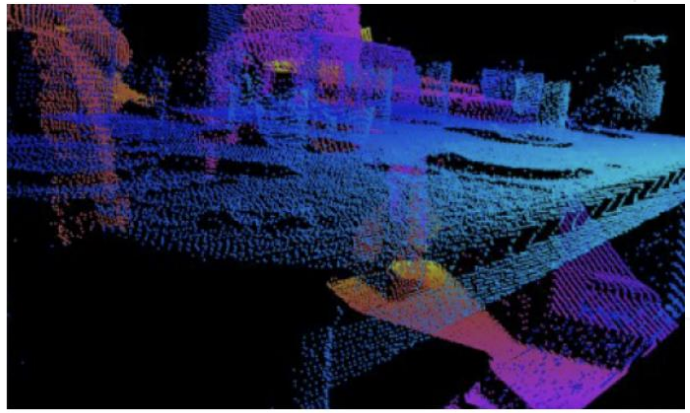
Depth Images



- Advantages
  - Dense representation
  - Gives intuition about occlusion and free space
  - Depth discontinuities are just edges on the image
- Disadvantages
  - Viewpoint dependent, can't merge
  - Doesn't capture physical geometry
  - Need actual 3D locations

Credit: Hawkins

Point Clouds



- Advantages
  - Viewpoint independent
  - Captures surface geometry
  - Points represent physical locations
- Disadvantages
  - Sparse representation
  - Lost information about free space and unknown space
  - Variable density based on distance from sensor

Credit: Hawkins

商汤 sensetime

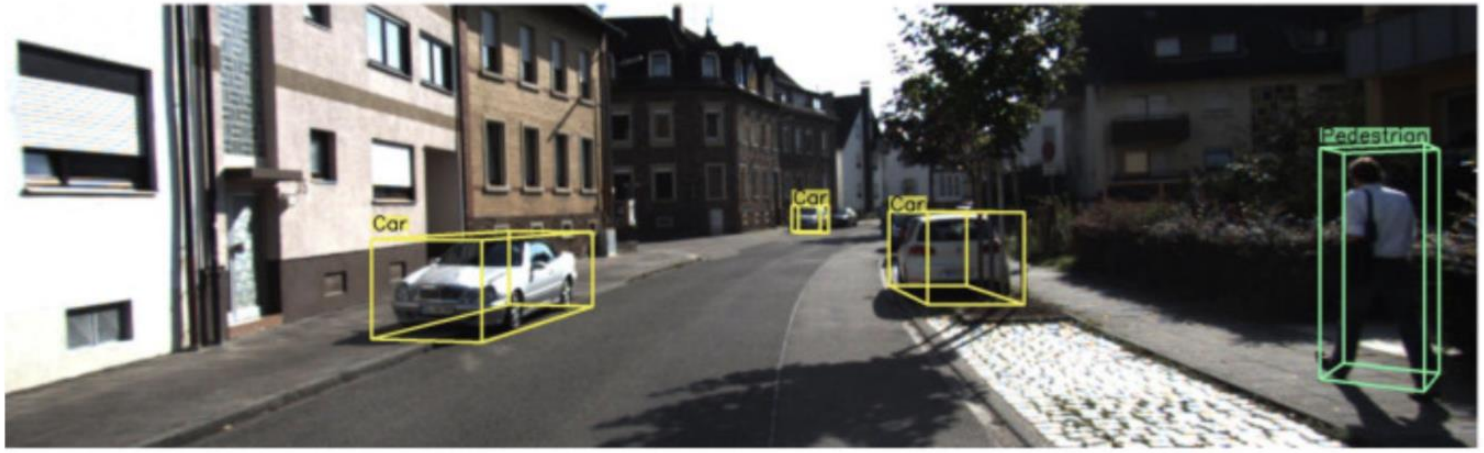
清華大學 Tsinghua University

Outline

Part 1	3D Vision Basics
Part 2	**3D Object Detection**
Part 3	3D Lane Detection

3D Object Detection

- 3D object detection is a crucial task for autonomous driving. Many important fields in autonomous driving such as **prediction, planning, and motion control** generally require a faithful representation of the 3D space *around the ego vehicle.*

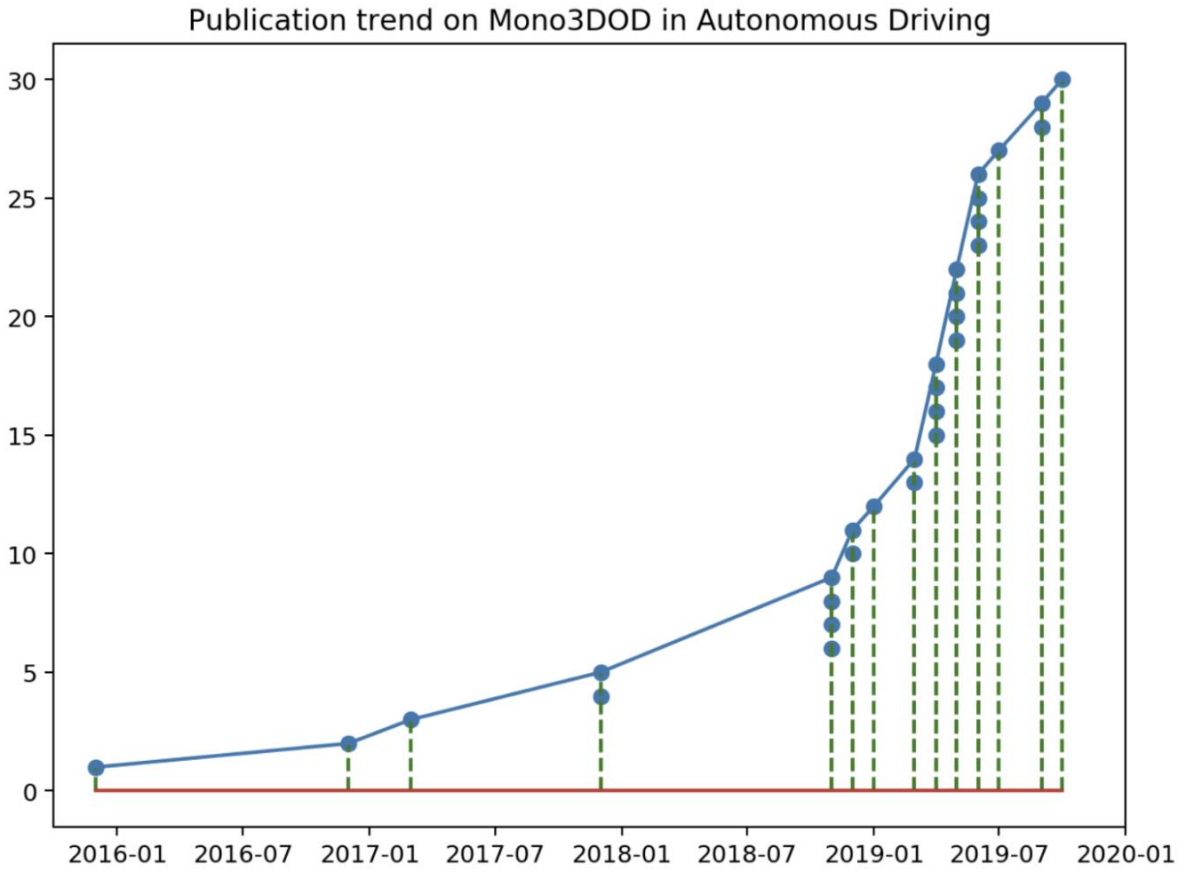


Monocular 3D Object Detection draws 3D bounding boxes on RGB images (source: M3D-RPN)

- **One Solution:**
  - LiDAR point cloud: PointNet
  - High cost
  - Sensitivity to adverse weather conditions

3D Object Detection

- Monocular 3D object detection with RGB image



Increasing amount of efforts in literature on monocular 3D object detection in Autonomous Driving

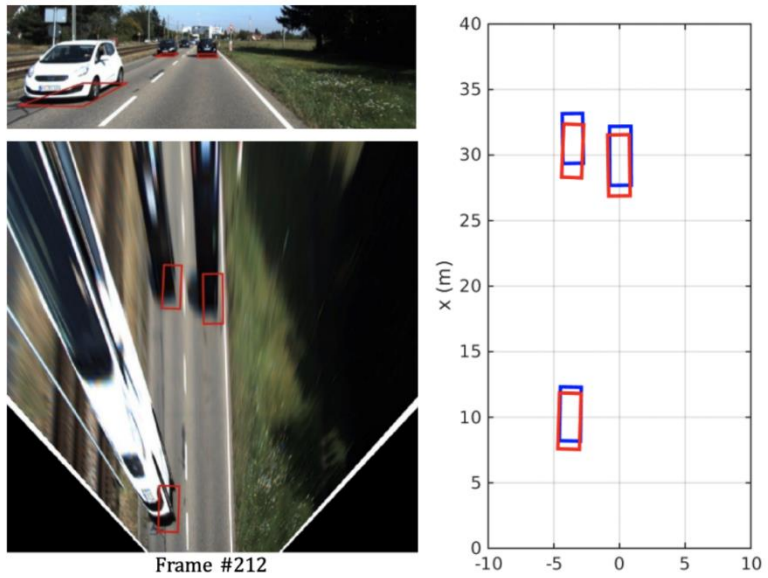
Roughly into 4 classes:

- Representation transformation
- Keypoints and shape
- Geometric reasoning based on 2D/3D constraint
- Direct generation of 3D bbox

Note: one method usually spans multiple categories and thus the grouping criterion is loose.

1. Representation transformation: BEV/pseudo-lidar

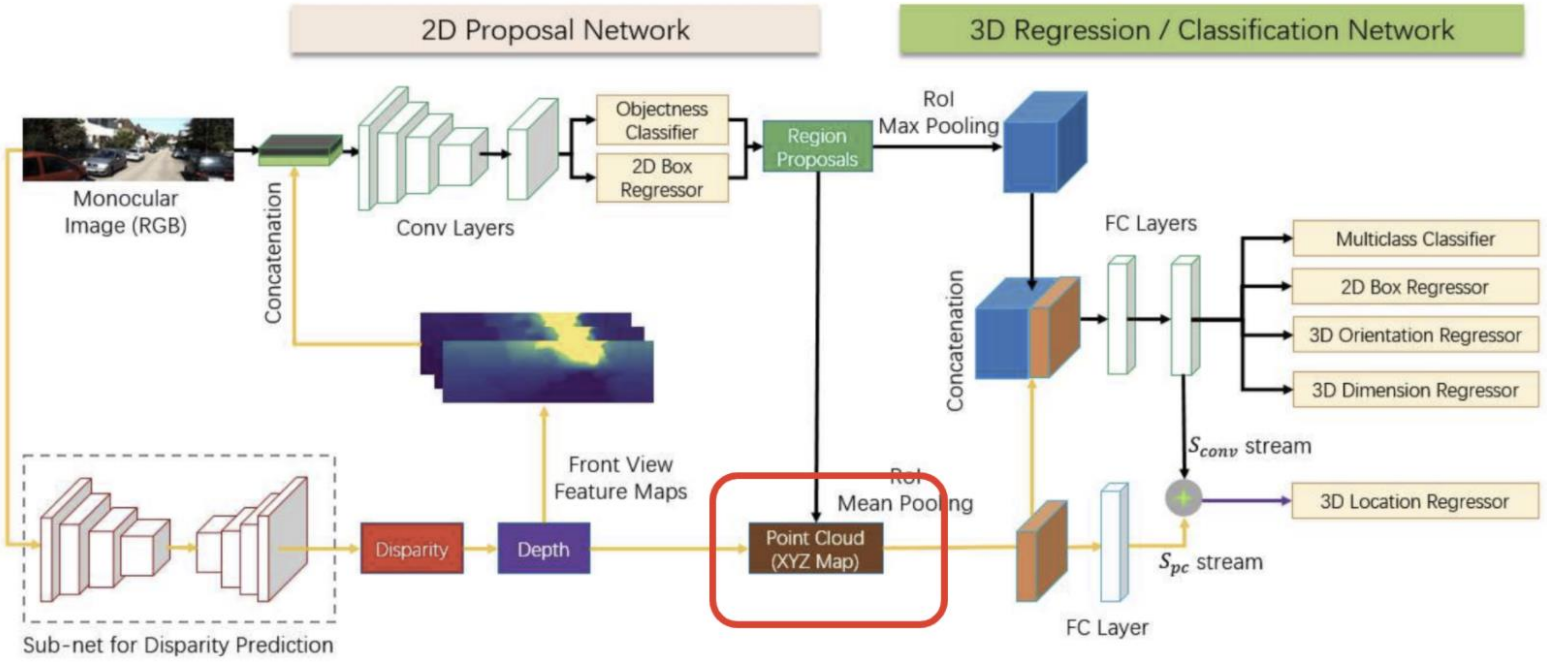
- BEV
  - Why BEV? Scale or occlusion
  - BEV method: IPM, or inverse perspective mapping
- Pseudo-lidar
  - Idea: to generate point cloud based on the estimated depth from the image
  - Details: using pseudo depth **as the fourth channel** and apply the normal lidar detection networks on this input, with minimal changes to the first layer.



Convert perspective image to BEV (from BEV-IPM)

1. Representation transformation: BEV/pseudo-lidar

- Pseudo-lidar
  - Idea: to generate point cloud based on the estimated depth from the image
  - Details: using pseudo depth **as the fourth channel** and apply the normal lidar detection networks on this input, with minimal changes to the first layer.



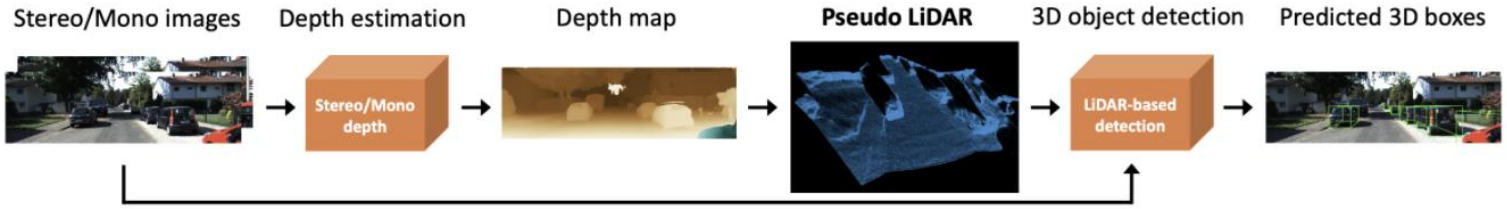
Architecture of Multi-level fusion (MLF), with pseudo-lidar generation circled in red (source)

Does it have a good motivation to convolve on the depth maps?

As neighboring pixels on depth images may be physically far away in 3D space.

1. Representation transformation: BEV/pseudo-lidar

- Pseudo-lidar



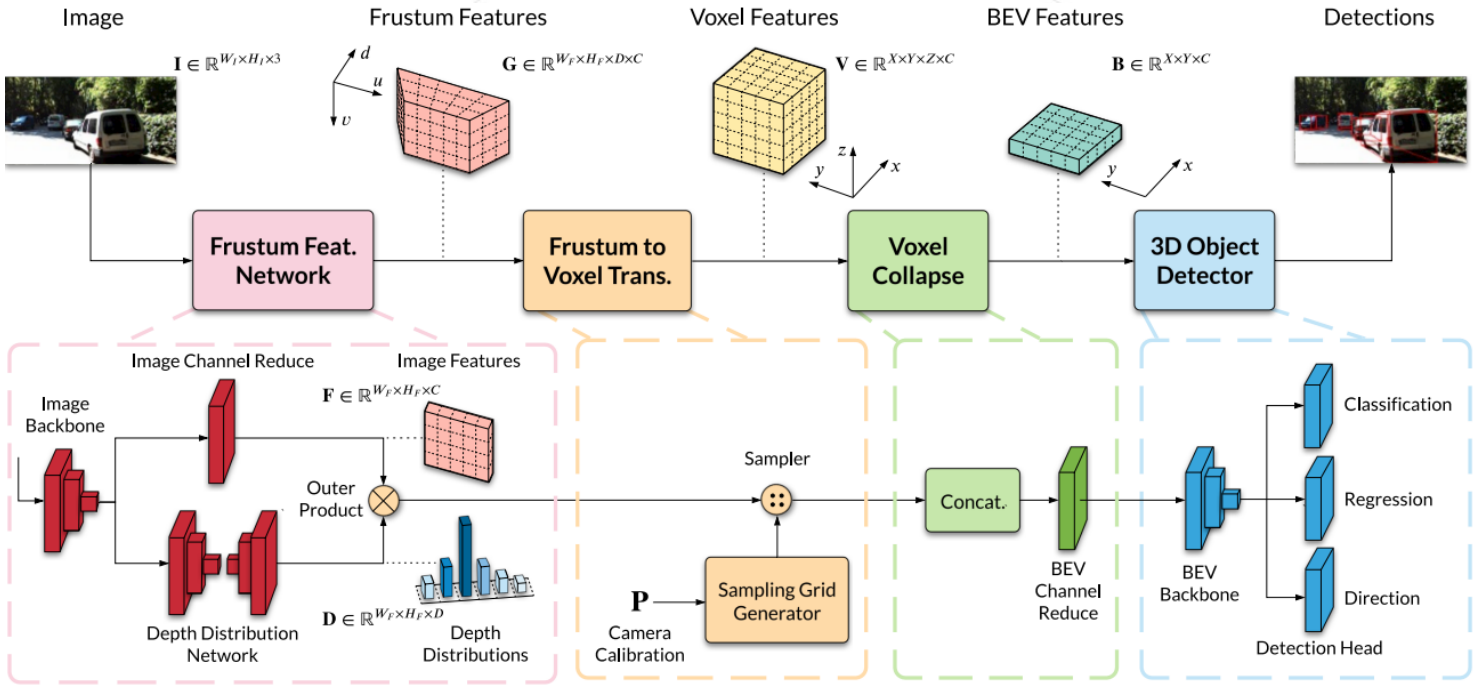
The general pipeline of the pseudo-lidar approach (source)

Caveat note:
there is some **overlap** between the training data of DORN, the off-the-shelf depth estimator, and the validation data of pseudo-lidar 3DOD.

1. Representation transformation: BEV/pseudo-lidar

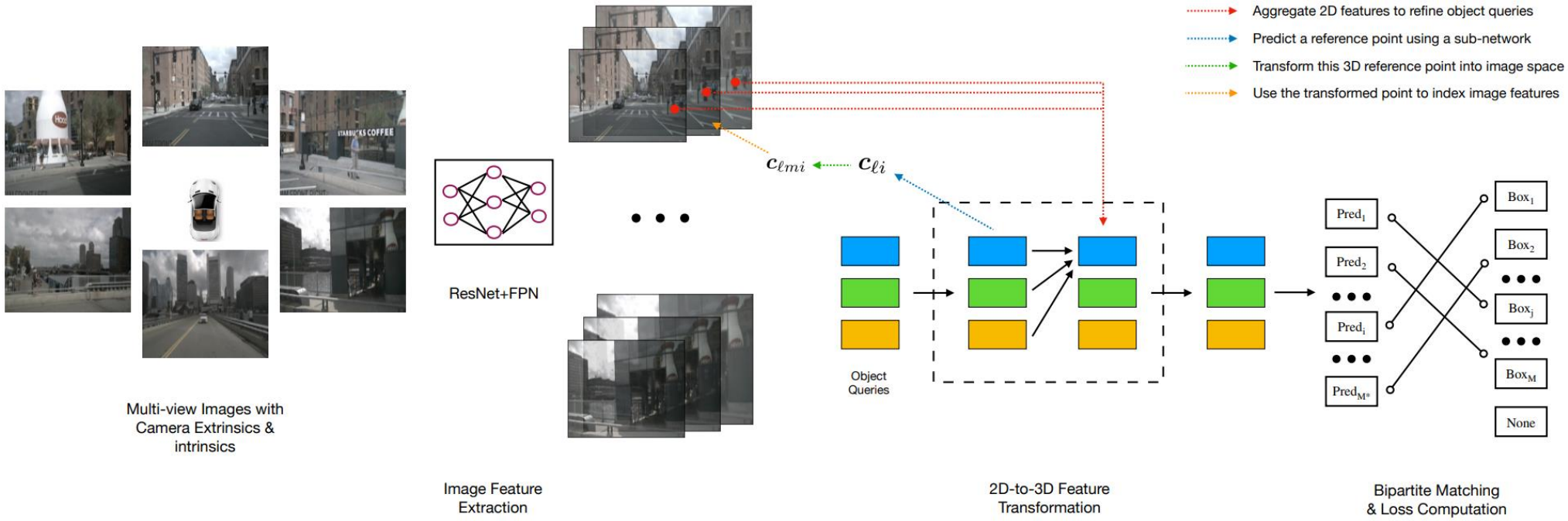
- BEV
  - Performance of monocular methods on the same 3D object detection benchmarks lags significantly relative to LiDAR and stereo ones, due to the loss of depth information when scene information is projected onto the image plane.

- Using categorical distributions allows our feature encoding to capture the inherent depth estimation uncertainty to reduce the impact of erroneous depth estimates



1. Representation transformation: BEV/pseudo-lidar

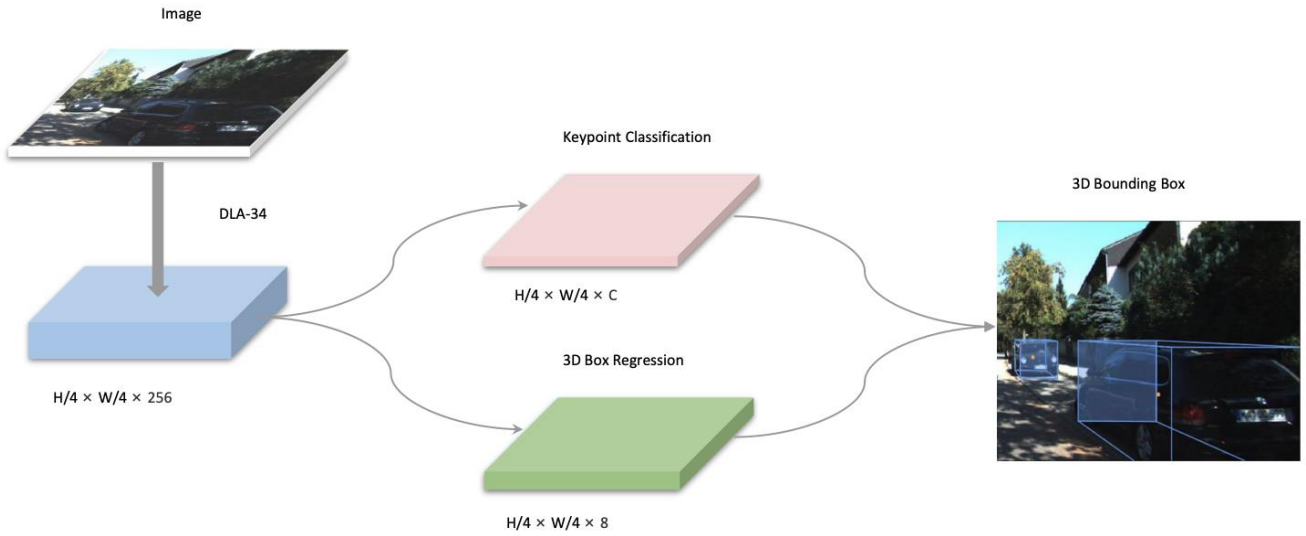
- Multi-camera BEV
  - One advantage of multi-camera BEV perception over Mono3D is in the camera overlap regions where objects are more likely to be cropped by camera field-of-view.
  - DETR3D specifically evaluated on such cropped objects at image boundaries (about 9% of the entire dataset) and found significant improvement of DETR3D over mono3D methods.



2. Keypoints and shapes

Motivation:

Vehicles are **rigid bodies** with distinctive common parts that can be used as landmarks/keypoints for detection, classification and re-identification. In addition, the dimension of the objects of interest (vehicle, pedestrians, etc) are objects **with largely known sizes**, including overall sizes and inter-keypoint sizes. The size information can be effectively leveraged to **estimate the distance** to ego-vehicle.



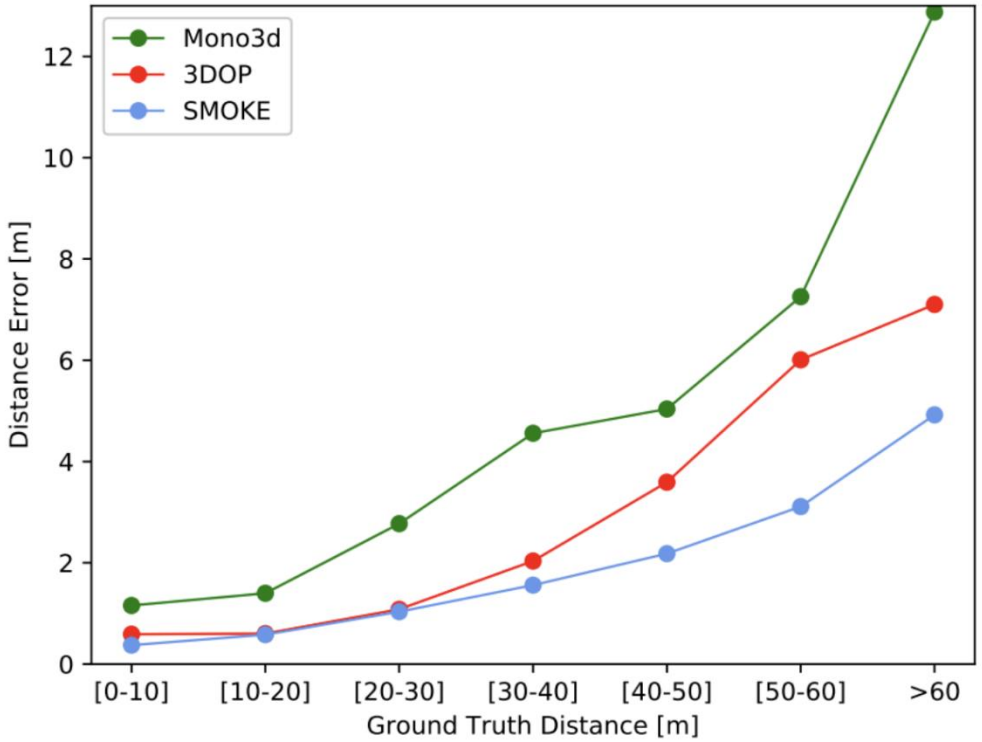
2. Keypoints and shapes

Popular methods:
SMOKE, CVPRW 2020

- Inspired by CenterNet.
- eliminates the regression of 2D bbox altogether
- directly predicts the 3D bbox.

It encodes a 3D bounding box as a point at the projection of the 3D cuboid center, with other parameters (size, distance, yaw) as its additional property.

Loss: 3D corner loss optimized using the disentangled L1 loss, inspired by **MonoDIS**.

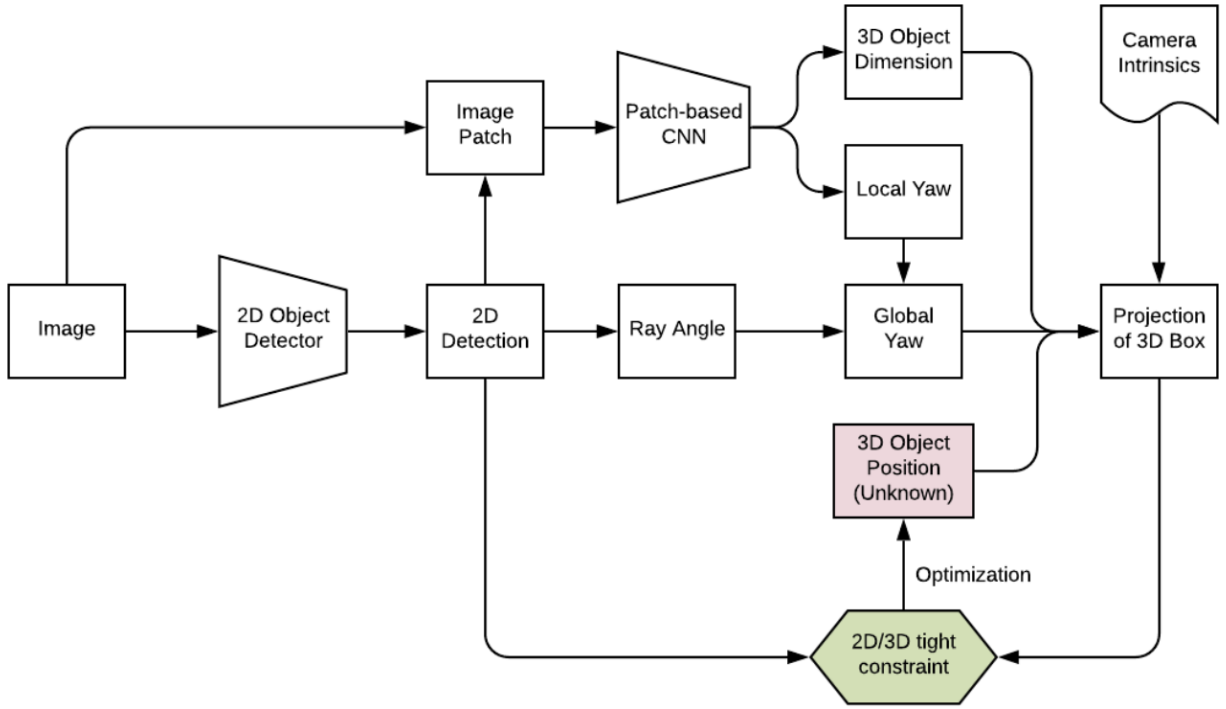


Average depth estimation error by SMOKE

3. Distance estimation through 2D/3D constraints

Popular methods:
Deep3DBox, CVPR 2016

- Extends 2D object detection framework by adding a branch regressing the local yaw (or observation angle) and the dimension offset from the subtype average.



The architecture of deep3DBox, representative of many other similar works (source)

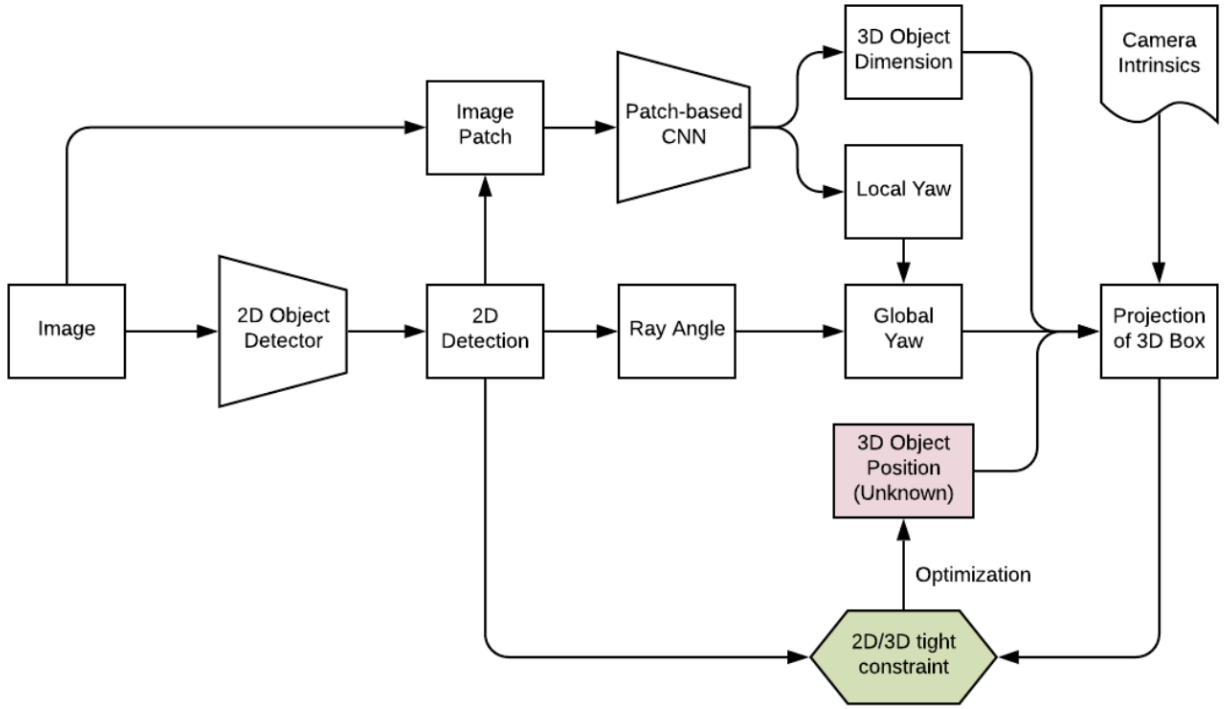
3. Distance estimation through 2D/3D constraints

Popular methods:
Deep3DBox, CVPR 2016

Two drawbacks

- ***relies on accurate detection of 2D bbox*** — *if there are moderate errors in the 2D bbox detection, there could be large errors in the estimated 3D bounding box*

- ***The optimization is purely based on the size and position of bounding boxes, and image appearance cue is not used***. *Thus it cannot benefit from a large number of labeled data in the training set.*

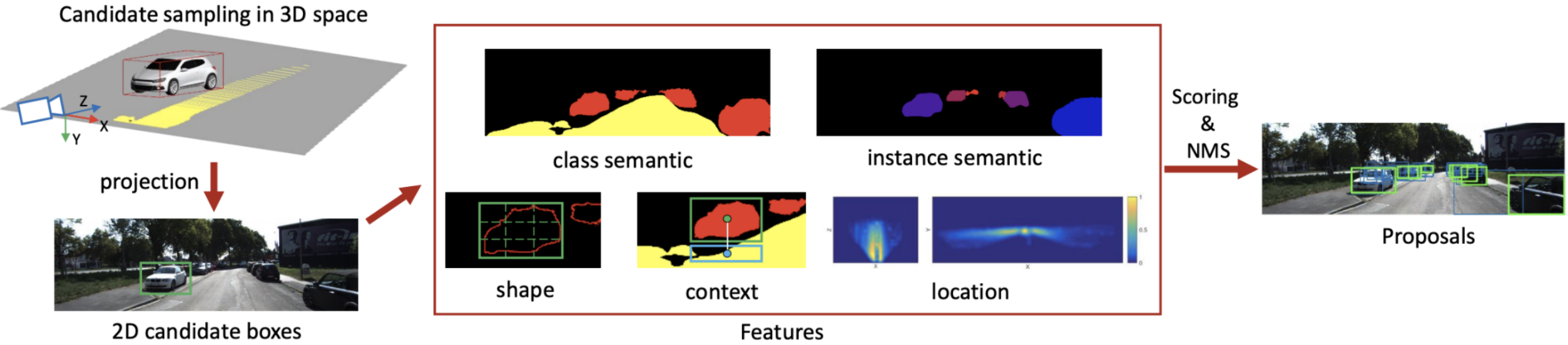


The architecture of deep3DBox, representative of many other similar works (source)

4. Direct Generation of 3D Proposal

Popular methods:
Mono3D, CVPR 2016 / CenterNet

CenterNet *first regresses a heat map indicating the confidence of the object center location and regresses other object properties. It is straightforward to extend CenterNet to include 2D and 3D object detection as the attribute to center points.*



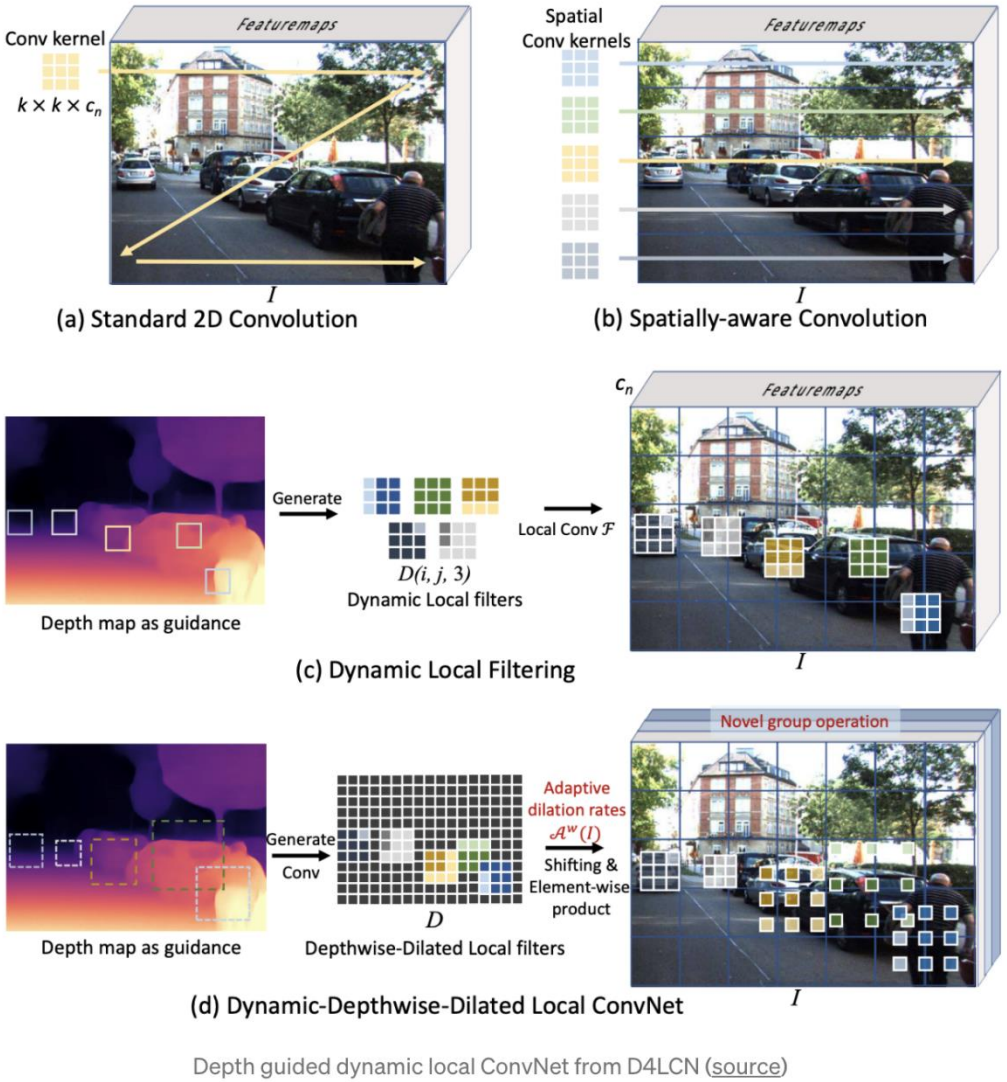
Mono3D places dense 3D proposals on the ground and scores them by manually crafted features (source)

4. Direct Generation of 3D Proposal

Popular methods:
D4LCN, CVPR 2020

Idea: depth-aware convolution from M3D-RPN even further by introducing a **dynamic filter** prediction branch.

This additional branch which takes in the depth prediction as input and generates a filter feature volume, which generates different filters for each specific location in terms of both weights and dilation rates.



Depth guided dynamic local ConvNet from D4LCN (source)

3D Object Detection: takeaways

2D and 3D consistency can help regularize joint 2D and 3D training and can help 3D reasoning as a postprocessing step after the prediction of 2D bounding box and geometric hints.

Monocular depth estimation had significant progress in the past few years. Dense depth estimation lends itself to transform RGB image to pseudo-lidar point cloud, ready to be consumed by state-of-the-art 3D object detection algorithms.

Perspective representation is hard to directly perform 3D detection with. Lifting to Bird's-eye View (BEV) space makes the detection of vehicles a much simpler task scale invariance at different distances.

All the above method assumes **known camera intrinsics.** If the camera intrinsics are unknown, many of the algorithms will still work but only up to a scale factor.

3D Perception Core: Depth Estimation

Input → Output

- **monocular** video frames
- stereo image pairs
- stereo video frames

- **depth**
- relative pose

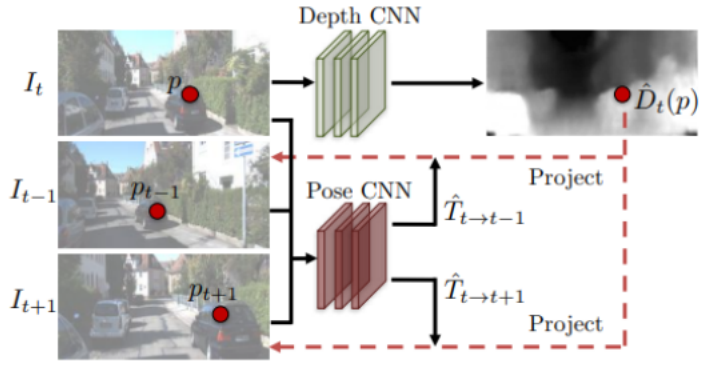
Assumption → Drawbacks

- Static world
- Moving camera
- Rigid body
- Lambert Surface

- **Dynamic objects**
- Scale ambiguity
- occulision
- Low texture regions

3D Perception Core: Depth Estimation

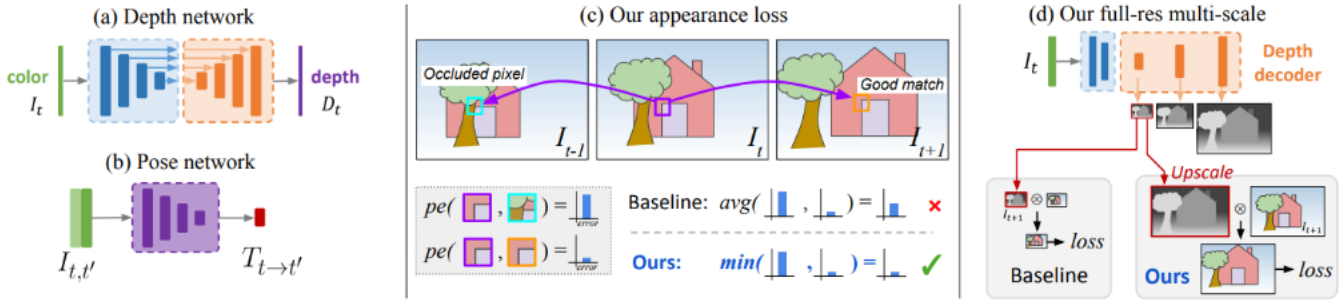
Self Supervised Monocular Depth Estimation



self supervised depth diagram

SFM-Learner[1]

- Unlabeled video input



monodepth2 diagram & it's loss improvment

Monodepth2[2]

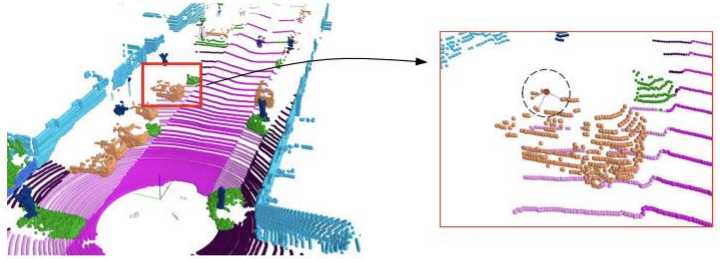
- Minimum reprojection loss
- Multi-scale training
- Auto-mask

[1] Zhou T, Brown M A, Snavely N, al. Unsupervised Learning of Depth and Ego-Motion from Video[J/OL]. 2017 IEEE CVPR, 2017: 6612-6619.

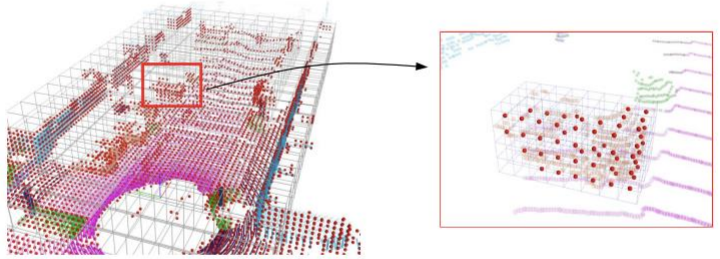
[2] Godard C, Mac Aodha O, Firman M, al. Digging Into Self-Supervised Monocular Depth Estimation[J/OL]. arXiv:1806.01260 [cs, stat], 2019[2021-12-06].

3D Perception: LiDAR Semantic Segmentation

http://www.semantic-kitti.org/tasks.html#semseg



(a) Point-based: disordered



(b) Voxel-based: sparse, quantization loss

(c) Range-based: physical dimensions distorted

Leaderboard. Following leaderboard contains only published approaches, where we at least can provide an arXiv link. (Last updated: August 24, 2021.)

To avoid confusion between the numbers reported in the paper and post-publication results, we report here the numbers from the paper. Please contact us if we missed an updated version with different numbers.

Single Scan | Multiple Scans

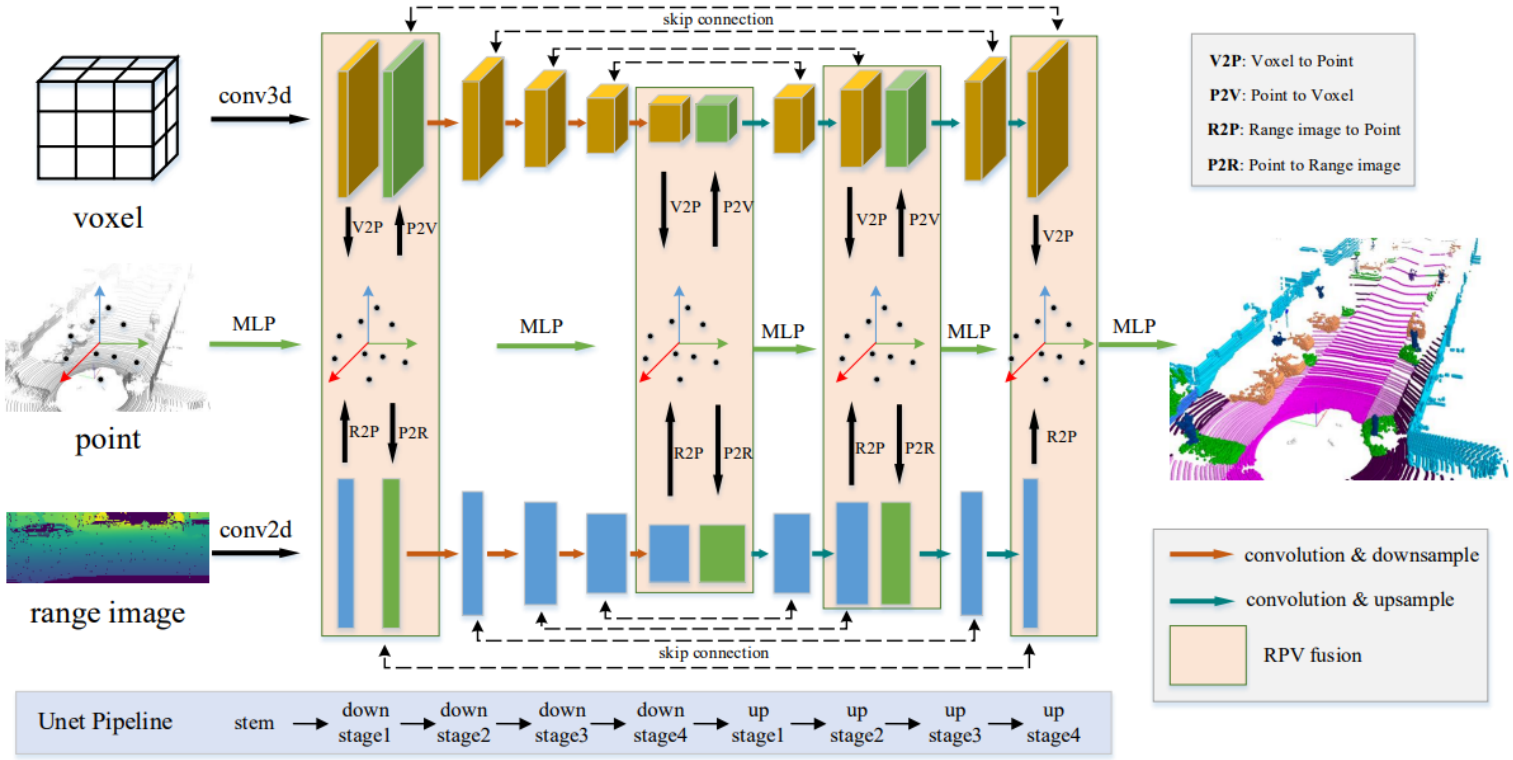
Approach	Paper	Code	mIoU	Classes (IoU)	Details
RPVNet			70.3		
AF2S3Net			69.7		
Cylinder3D			67.8		
SPVNAS			66.4		
JS3C-Net			66.0		
AMVNet			65.3		
Lite-HDSeg			63.8		
TORNADONet			63.1		
KPRNet			63.1		

RPVNet

- Multi-modal representation of LiDAR data
  - Point clouds can be represented in many forms (views), typically, point-based sets, voxel-based cells or range-based images(i.e., panoramic view).
  - Utilize different views advantages and alleviate their own shortcomings in fine-grained segmentation task

- The point-based view is geometrically accurate, while disordered.
- The voxel-based view is regular, while sparse
- The range-based view is regular and generally dense, while distorted

- Summary
  - **R**ange-image, **P**oint, **V**oxel fusion
  - mutual information interactions



Cylinder3D

- Cylindrical Partition
  - varying-density, imbalanced distribution
  - cylinder coordinate
- Asymmetrical 3D Convolution Network
  - specific object shape distribution (cubic objects)
  - asymmetrical residual block (match ~)

- Summary
  - outdoor LiDAR point cloud
  - distribution: point cloud & specific object
  - cylinder coordinate

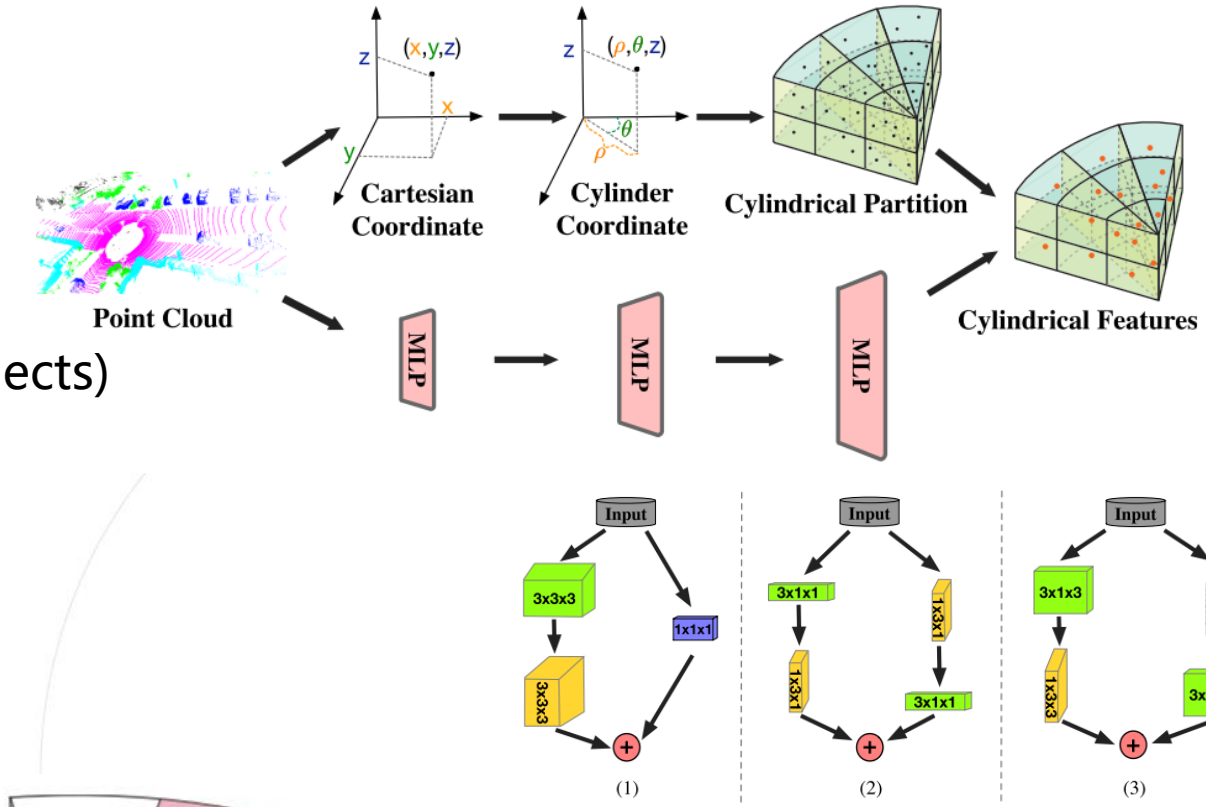
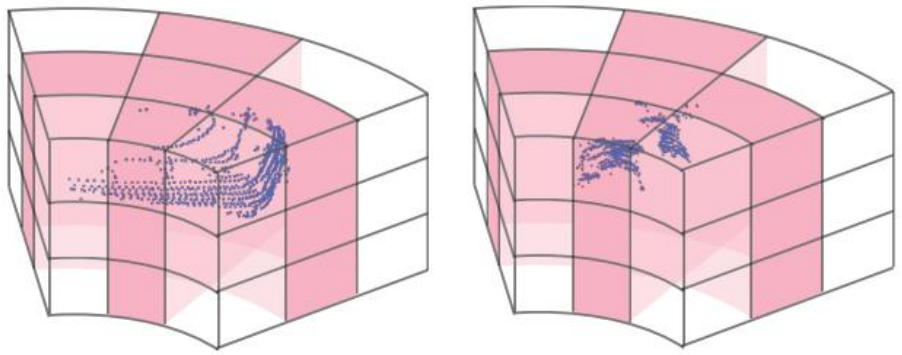


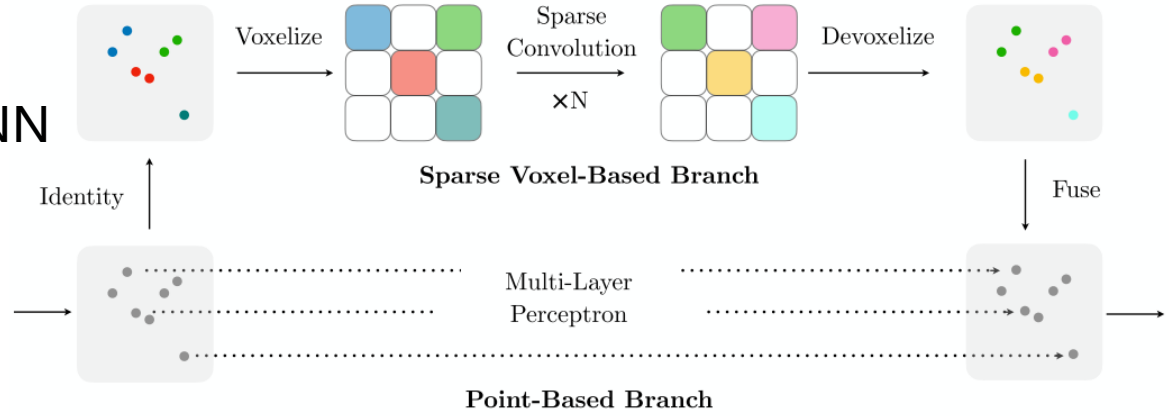
Figure 7: We design three blocks for ablation studies of asymmetrical residual block, including (1) regular residual block , (2) 1D-asymmetrical residual block without height and (3) the proposed asymmetrical residual block.

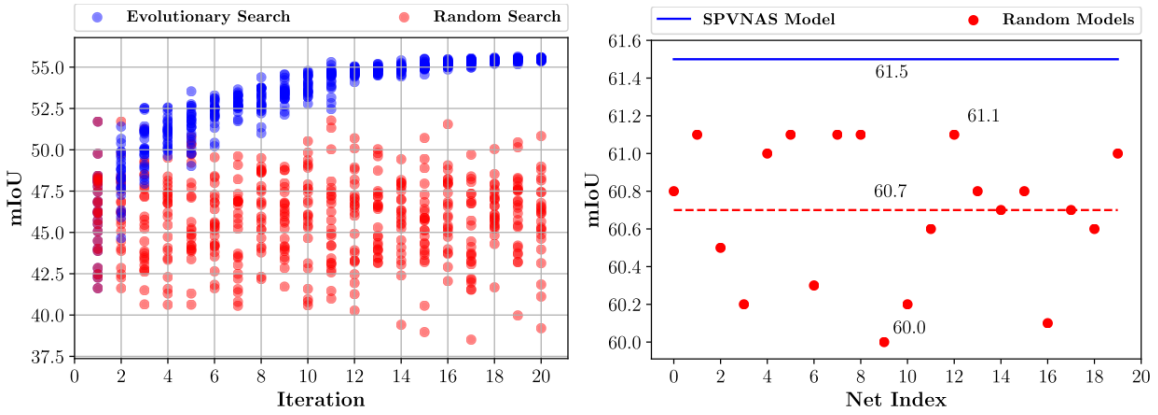


(a) Car (b) Motorcycle

SPVNAS

- Sparse Point-Voxel Convolution
  - Sparse Convolution cannot always keep high-resolution
  - Point-Voxel Convolution does not scale up to large 3D scenes
- 3D Neural Architecture Search
  - architecture search framework for 3D scene
  - improves the efficiency and performance of SPVCNN
- Summary
  - SPVNC: large scenes & high-resolution
  - NAS (evolutionary search)
  - lightweight, fast and powerful





(a) Search curves of ES and RS. (b) Comparison with random models.

商汤 sensetime

清華大學 Tsinghua University

Outline

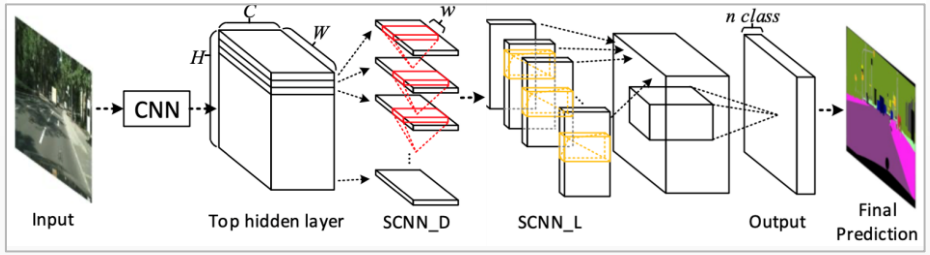
Part 1	3D Vision Basics
Part 2	3D Object Detection
Part 3	**3D Lane Detection**

背景介绍 - 传统车道线检测方法

2D车道线检测算法

分割思路

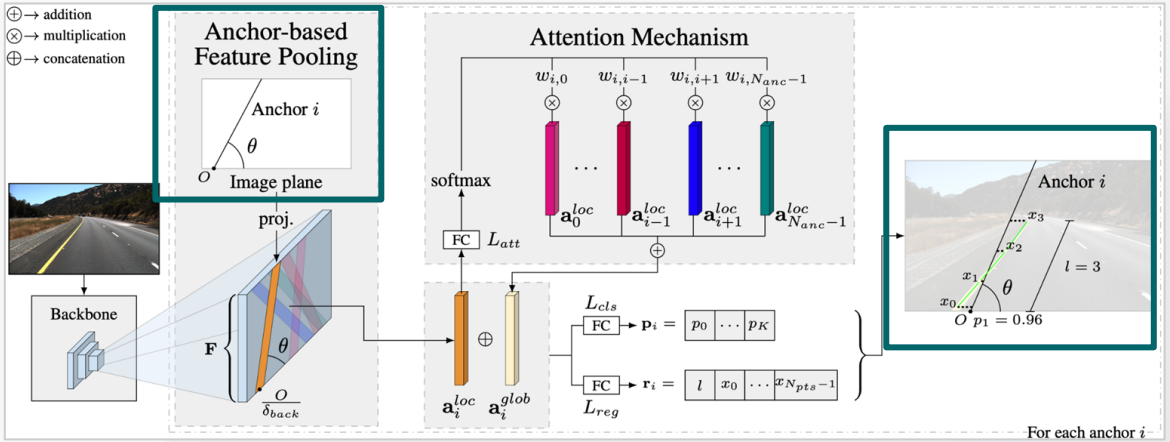
- SCNN [1]
- LaneAF [2]



Anchor-based思路

- LaneATT [3]
- CondLaneNet (Row-wise) [4]

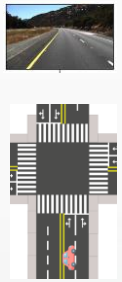
- pro：
  - 利用 lane shape信息
  - instance discrimination



[1] SCNN, AAAI 2018
[2] LaneAF, preprint
[3] LaneATT, CVPR 2021
[4] CondLaneNet, CVPR 2021
[5] BEV IPM OD, IV 2019
[6] Pseudo-LiDAR end2end, ICCV 2019
[7] CaDDN, CVPR 2021
[8] Deep Multi-Sensor LD, IROS 2018

鸟瞰图BEV (Bird's Eye View) 下的车道线检测

一般步骤



2D车道线检测 → 拟合，sample → Inverse Perspective Mapping (IPM) → 拟合，后处理

相机内外参 *Extrinsic/ Intrinsic*

IPM：图像平面到鸟瞰平面的投影

Why BEV - 学术界越来越流行

1. BEV下没有 **遮挡occlusion** 和 **近大远小scale** 的问题 [5-8]
2. 方便后续规划控制模块的开发

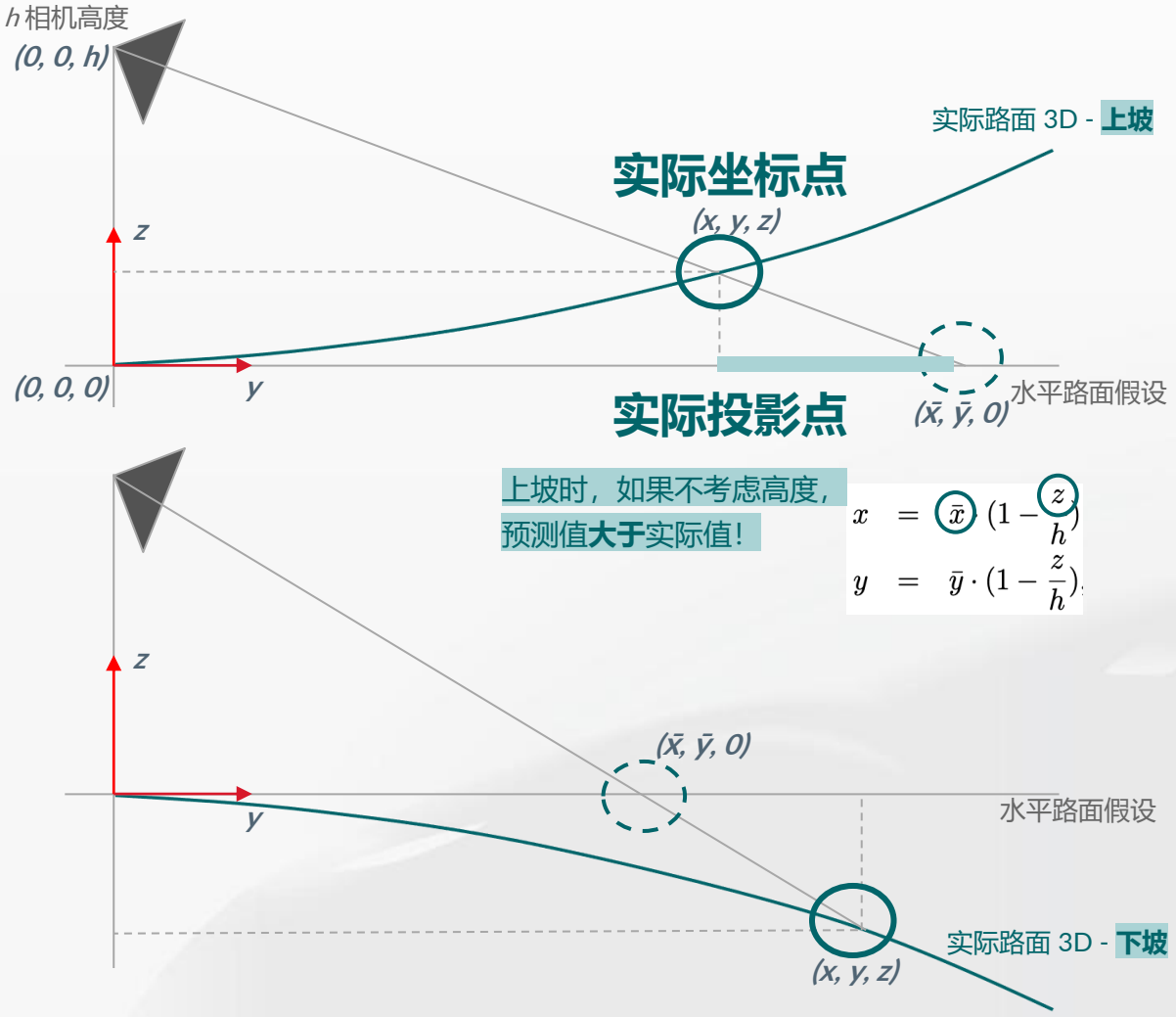
规划控制模块对车道线算法的需求

1. 需要线型(Lane Type)、路沿(Curb)检测功能；路口车道线补全功能
2. 需要满足不同距离下的水平误差

Distance	0m	10m	30m	50m
项目允许误差	+/- 6.5 cm	+/- 6.5 cm	+/- 8.0 cm	+/- 9.0 cm
目前算法 [SenseAuto Mono] 误差	14.6 cm	15.8 cm	20.6 cm	25.34 cm

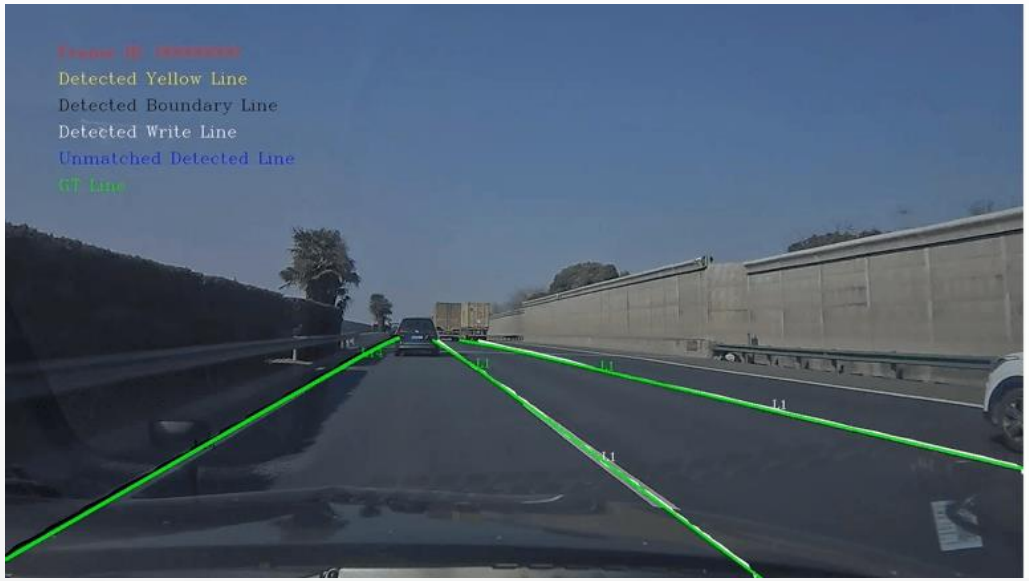
背景介绍 - 3D车道线动机

Why 3D? 在BEV下, 从 (x,y) 到考虑高度 (x,y,z)



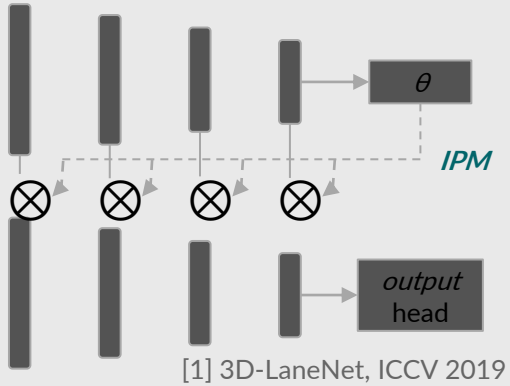
上坡时，如果不考虑高度，预测值**大于**实际值！

$$x = \bar{x} \cdot (1 - \frac{z}{h})$$
$$y = \bar{y} \cdot (1 - \frac{z}{h})$$

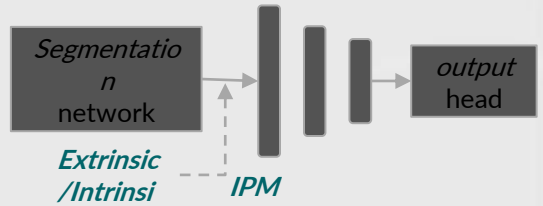


“水平路面假设”在复杂场景 (e.g., 上下坡, 颠簸路面等) 不成立！

该问题在学术界的**已有工作** [1-2]



[1] 3D-LaneNet, ICCV 2019



- **依赖分割结果**
- **基于Synthetic data**

[2] Gen-LaneNet, ECCV 2020

| 关键创新点1：

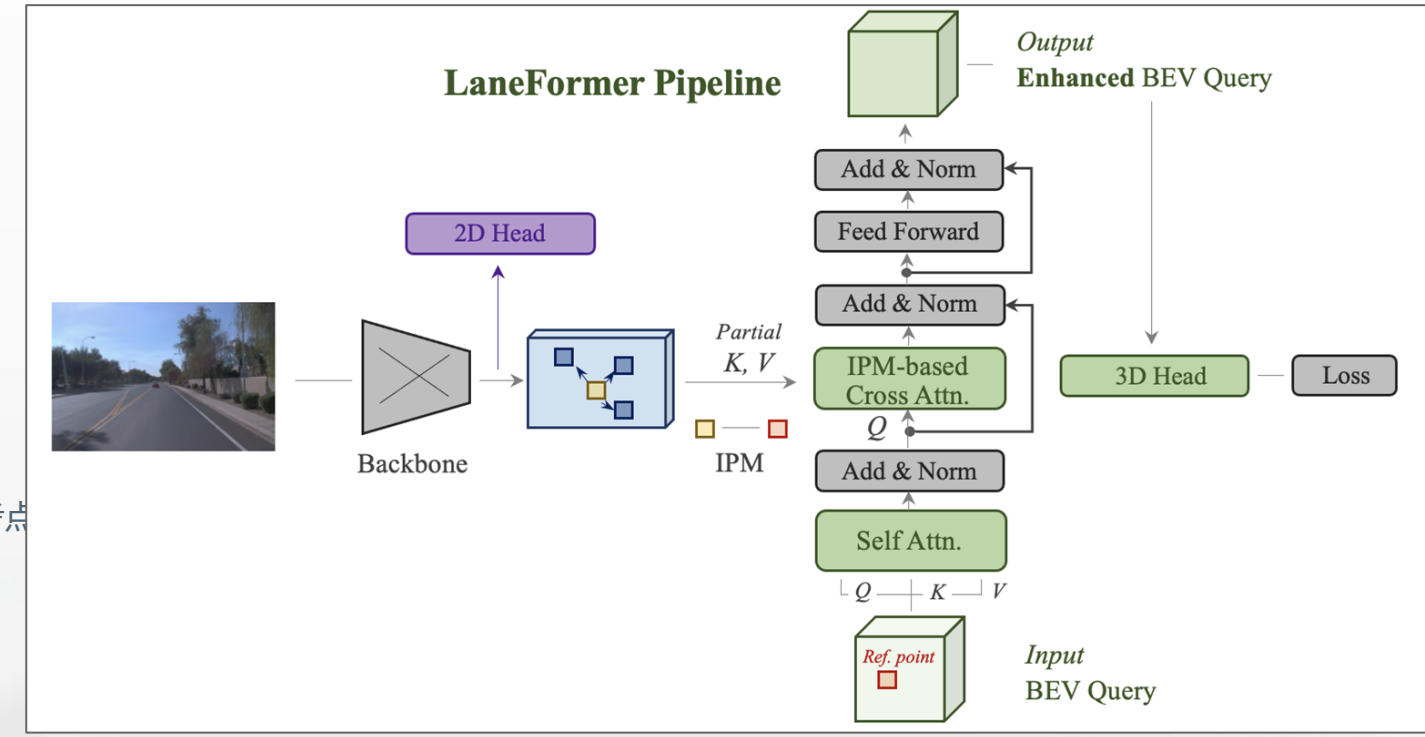
方案技术原创 - Revised Transformer结构提出 – PersFormer[1]

Motivation

1. 传统IPM生成的BEV特征图是**单纯的采样生成**，采样点由相机内外参计算得到，特征来源相对**稀疏**。
2. **注意力机制**可以涵盖一定范围内的局部特征，模型可以自动的去寻找更丰富的特征来源，特征来源相对**稠密**。

Novelty in LaneFormer

1. 使用Deformable Attention做**大尺度**(200*100)的BEV查询
2. 使用相机内外参得到***前视特征图***到***BEV特征图***的对应坐标参考点
3. 加入BEV下车道线分割任务，软监督生成的BEV特征图



Key Results

In terms of F1-Score, on SenseMentor v2.0 benchmark

Methods	Overall	Curve	Intersection	Extreme Weather	Night	Merge & Split	Up & Down
3D-LaneNet [2]	40.2	43.2	29.3	43.0	39.3	36.5	37.7
Gen-LaneNet [3]	29.7	31.1	19.7	26.4	17.5	27.4	24.2
Ours	**47.8**	**52.8**	**37.9**	**48.7**	**46.0**	**44.6**	**42.4**

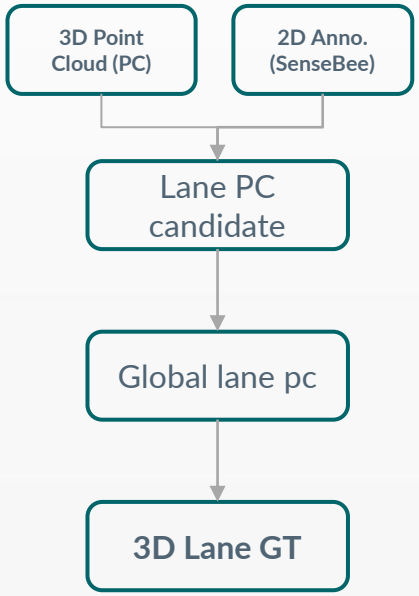
比目前业界最好的方法，性能提升7.6%！

[1] PersFormer, arXiv, https://arxiv.org/abs/2203.11089
[2] 3D-LaneNet, ICCV 2019, https://arxiv.org/abs/1811.10203
[3] Gen-LaneNet, ECCV 2020, https://arxiv.org/abs/2003.10656

| 关键创新点2：

数据壁垒原创 - 高质量3D真值与大规模数据集生成

车道线3D真值生成过程

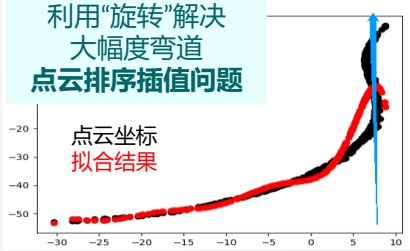


真值生成用到的优化技术

- PVB点云过滤
- 点云插值
- 离群点过滤
- 全局pose拼接
- 有效长度可见性判断
- 不匹配线过滤
- 全局插值
- 滤波平滑

技术难点举例
拟合方法失效

利用“旋转”解决大幅度弯道**点云排序插值问题**



大规模数据集生成

标注环节

1. 离线质检工具开发
2. 标注沟通文档，**提升质检环节效率**
3. 为数据仓库开发铺垫

提升质检操作效率 500%
一步式定位问题图片

分布式真值生成 - 对标亚马逊SageMaker框架

1. 多并发云端数据处理
2. 数据集检验模型/Loader

提升数据生成效率 100倍

3D车道线真值生成过程

| 关键创新点2：

数据壁垒原创 - 高质量3D真值与大规模数据集生成（续）

大规模数据集生成 - SenseMentor 2.0

公开车道线数据集 - 详细对比

Dataset	Size			Diversity			
	#Video	#Frames	Average Length	Instance-level Annotation	Maximum #Lanes	Line-type 线型	Scenario 场景种类
TuSimple 2017	6.4K	6.4K/128K	1s	✓	5	-	Light-traffic, day
CULane 2018	-	133K/133K	-	✓	4	-	Multi-weather, Multi-traffic
ApolloScape 2019	235	115K/115K	16s	✗	-	13	Multi-weather, Multi-traffic
BDD100K 2020	100K	100K/120M	40s	✗	-	11	Multi-weather, Multi-traffic
VIL-100 2021	100	10K/10K	10s	✓	6	10	Multi-weather, Multi-traffic
SenseMentor 2022 Q1	**1.85K**	**240K/240K**	**20s**	✓	**25**	**14**	Multi-weather, Multi-traffic, **Multi-topography, Multi-road structure**

提出意义

- 这是community **首次** 推出真实场景3D车道线标注
- **规模最大**、车道线种类最多、场景最丰富；符合量产实际需求
- 车道线和物体检测在同一个数据集上(Waymo)，不改变用户使用习惯；方便后续**感知任务扩展**

长线价值

21Q3-4	22 Q1	22 1H
SenseMentor v2.0	WIP 多模态、环视数据 数据仓库建设	TODO 数据平台 数据工厂批量生产

2D车道线　　BEV可视化真值

| 关键创新点3：

模块可扩展性 - 多任务、多场景、可落地

同时预测2D/3D车道线

Key Notes

1. 3D（BEV）范围有限（20x100m）但可提供更精确的高度信息，2D蕴含更丰富的场景信息
2. 2D & 3D head，采用一致的anchor设计 (IPM对应关系)
3. 满足不同场景需求

In terms of F1-Score, on SenseMentor v2.0 benchmark

Methods	Overall	Curve	Night	Merge & Split	Up & Down
Ours w/o 2D loss	42.75	47.56	42.26	40.76	37.04
Ours	**46.61**	**50.92**	**55.71**	**43.31**	**41.44**

可落地 - SenseAuto自采数据验证

未训练直接inference结果

服务 HZ-EP40 (左)、GAC-NDA (右)等量产项目

多场景 - 解决自动驾驶Cornercase/长尾分布

2D检测结果 *BEV检测结果* *2D检测结果* *BEV检测结果*

清華大學 Tsinghua University

商汤 sensetime

One last drop(s) for me

Reach me at lihongyang@senseauto.com