Chapter 2 - Section 12

3D Vision and Applications

Dr. Li Hongyang

Friday, May 14, 2021

商汤
sensetime

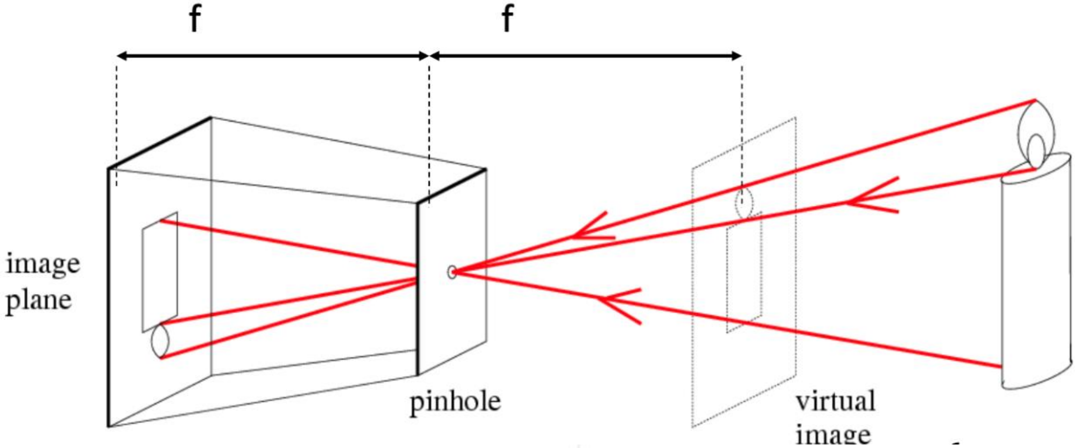
清華大學
Tsinghua University

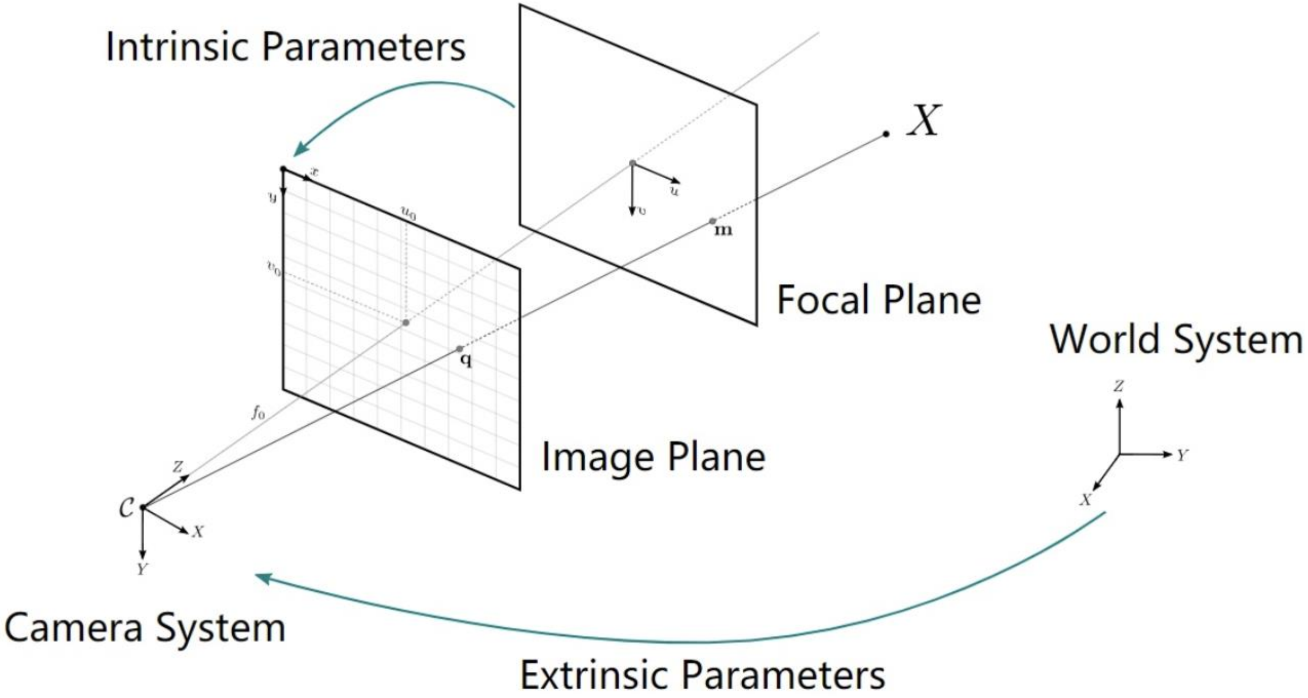
Outline

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

Prerequisite: camera model/intrinsics/extrinsics

- The pinhole model





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

Prerequisite: camera model/intrinsics/extrinsics

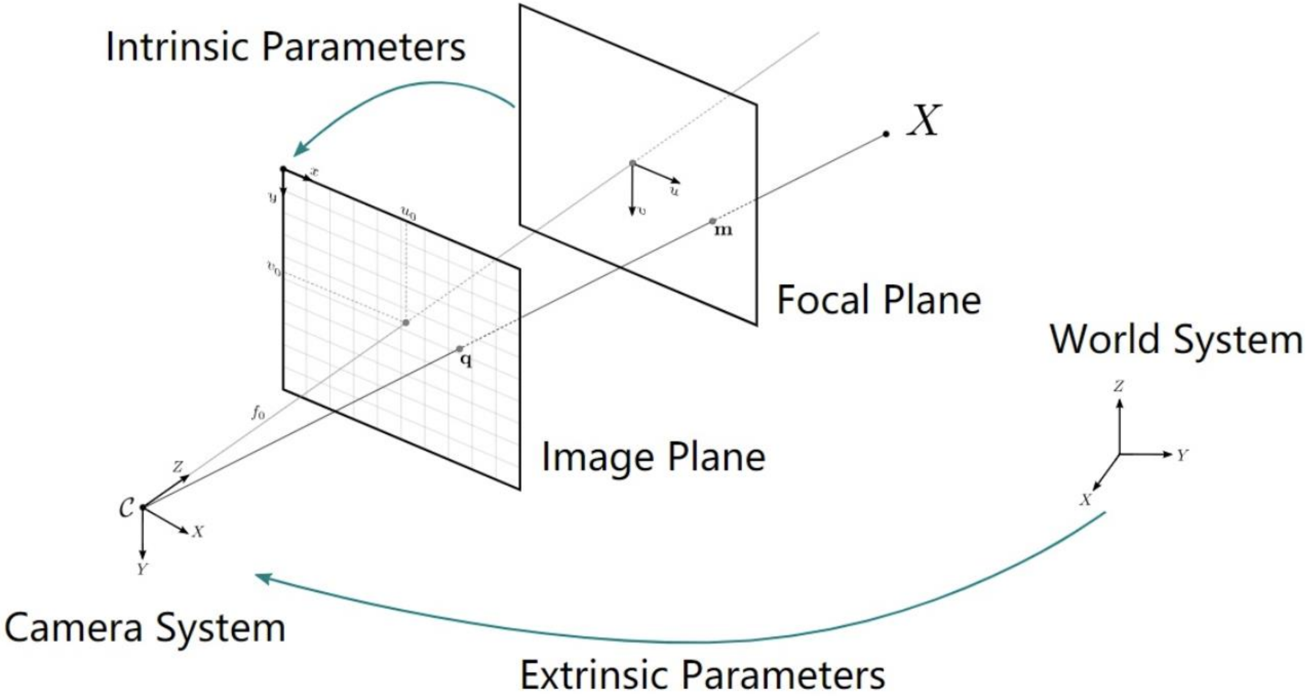
- The pinhole model

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。

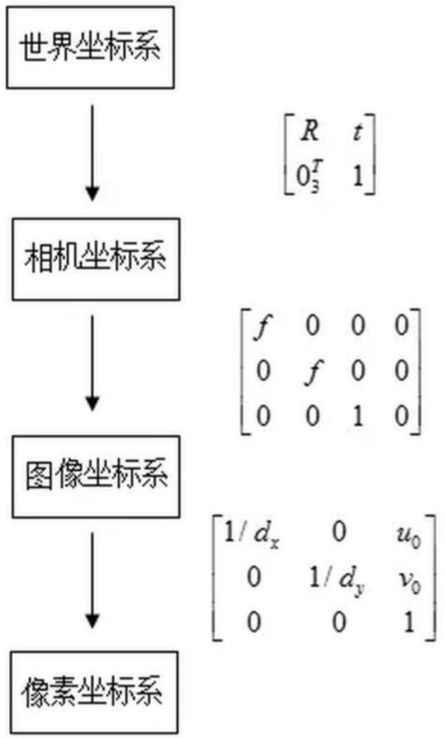
图像坐标系(image coordinate system)：为了描述成像过程中物体从相机坐标系到图像坐标系的投影透射关系而引入，方便进一步得到像素坐标系下的坐标。 单位为m。

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



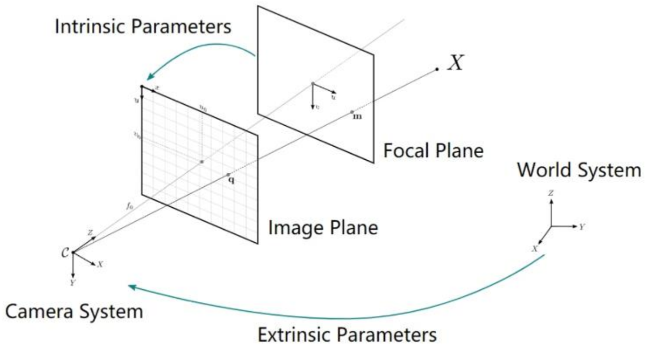
$$\begin{bmatrix} R & t \\ 0_3^T & 1 \end{bmatrix}$$

$$\begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1/d_x & 0 & u_0 \\ 0 & 1/d_y & v_0 \\ 0 & 0 & 1 \end{bmatrix}$$

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：

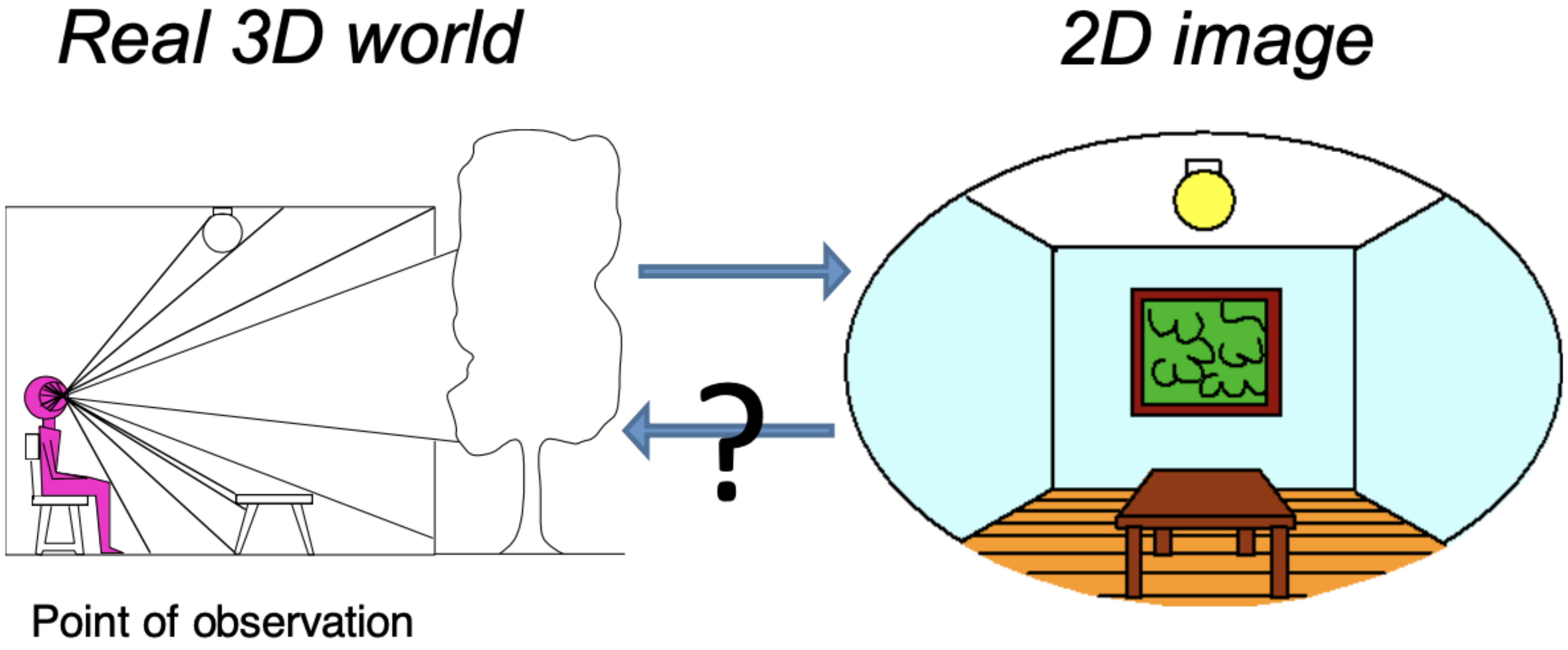
$$H = s \begin{bmatrix} f_x & \gamma & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_1 & r_2 & t \end{bmatrix} = sM \begin{bmatrix} r_1 & r_2 & t \end{bmatrix}$$

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

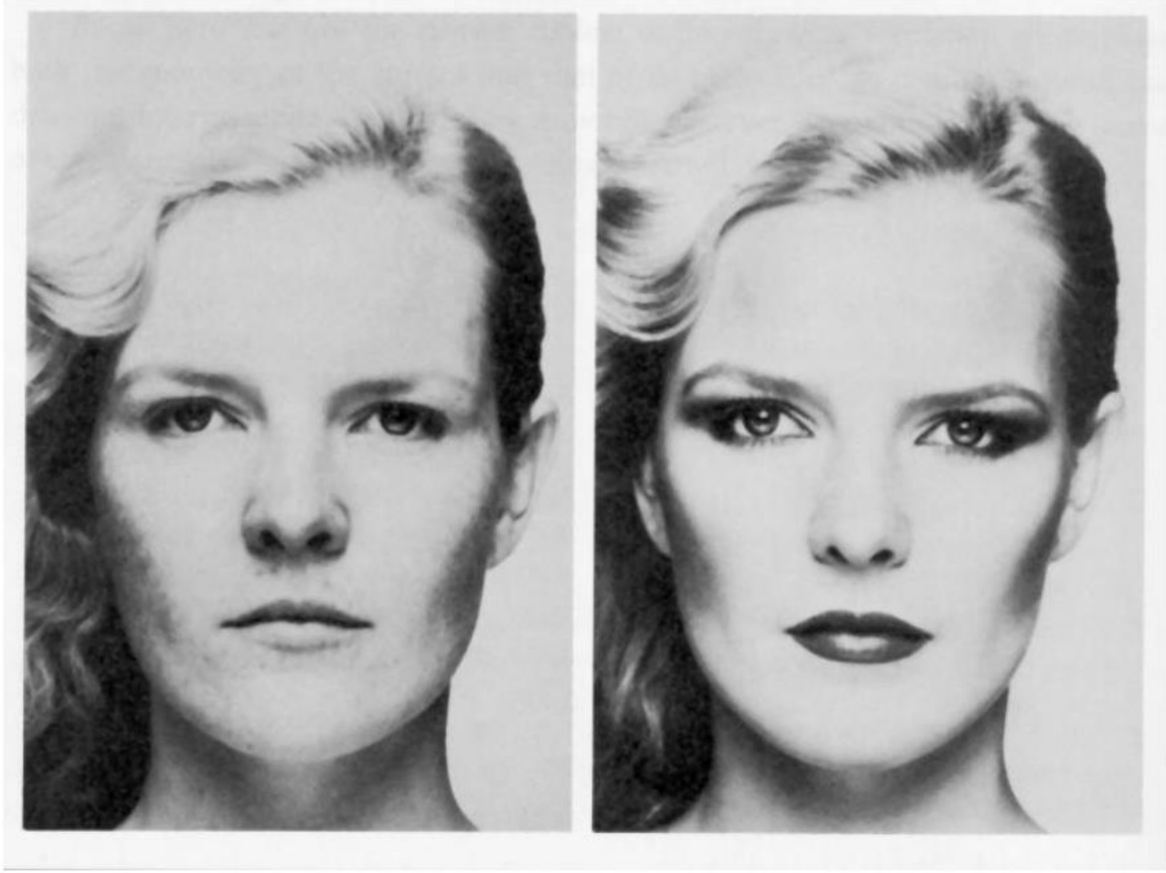
How can we automatically compute 3D geometry from images?
– What cues in the image provide 3D information?



Credit: Fei-fei Li

Recover 3D from images: cues for 3D

Shading

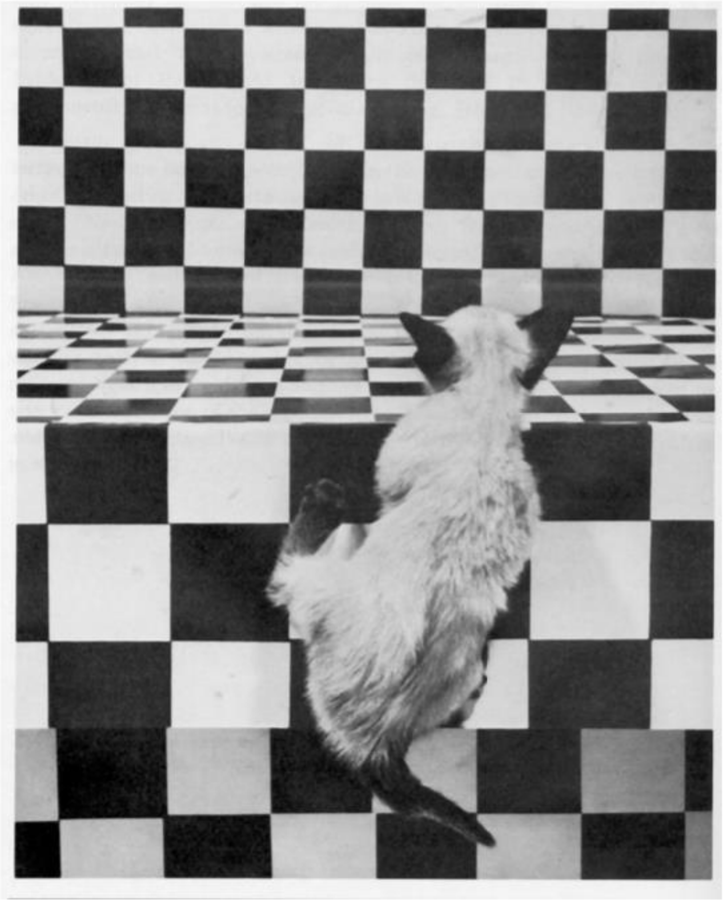


Merle Norman Cosmetics, Los Angeles

Credit: Fei-fei Li

Recover 3D from images: cues for 3D

- Shading
- Texture

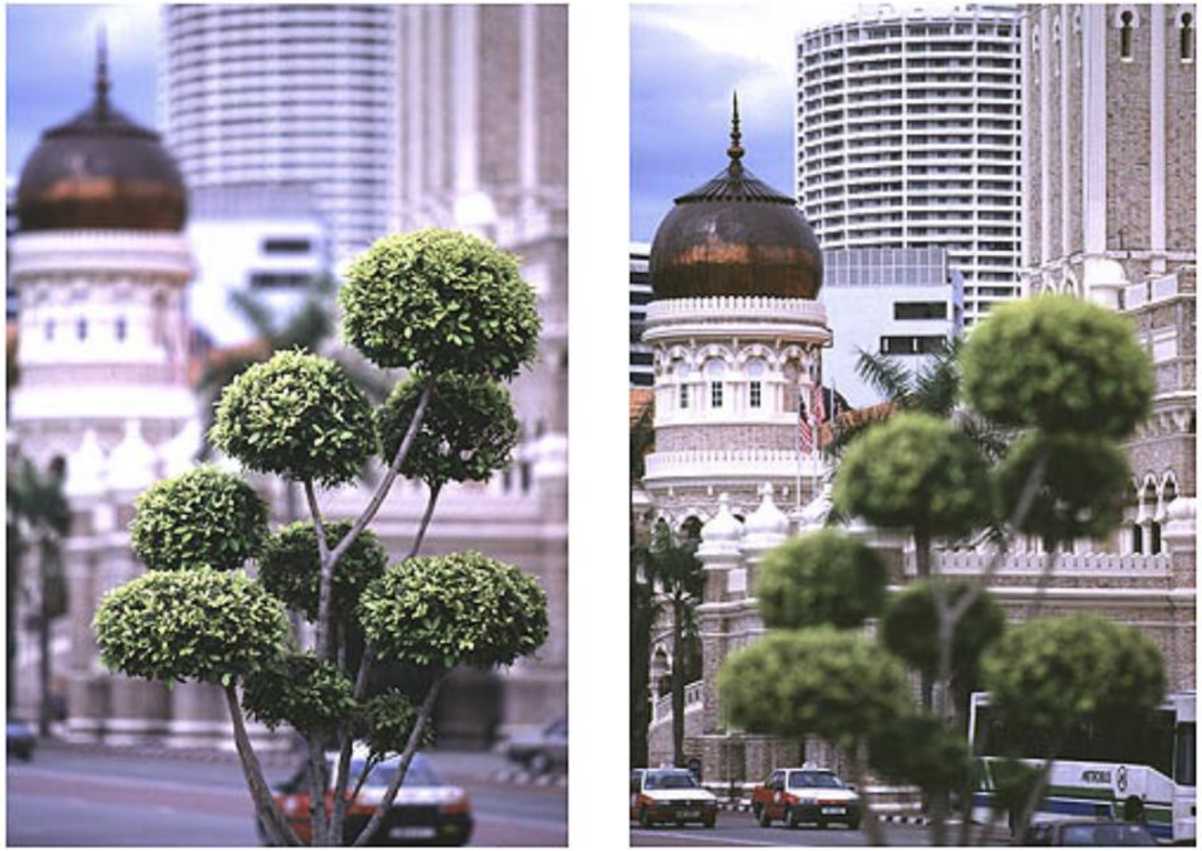


***The Visual Cliff*, by William Vandivert, 1960**

Credit: Fei-fei Li

Recover 3D from images: cues for 3D

- Shading
- Texture
- **Focus**

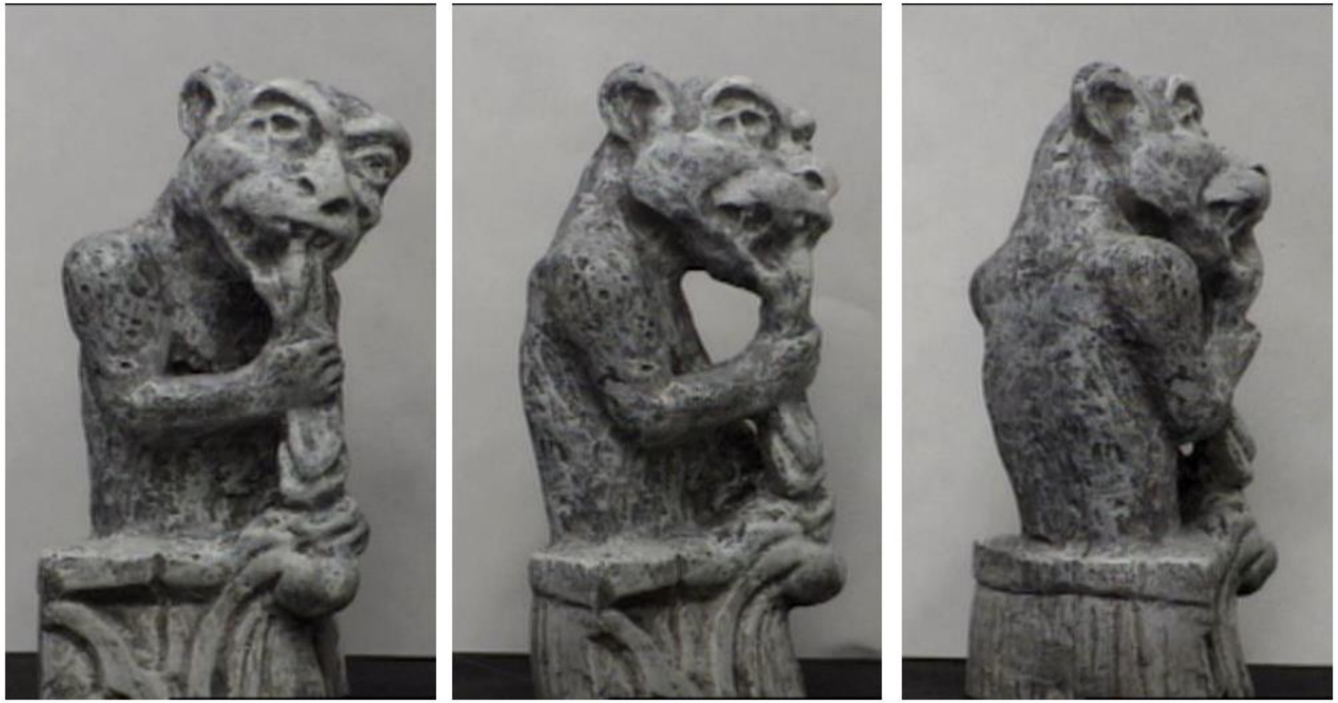


From *The Art of Photography*, Canon

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

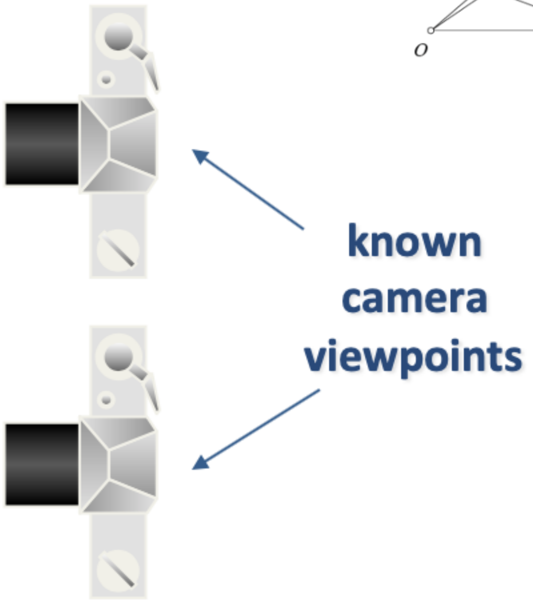
Credit: Fei-fei Li

Stereo Reconstruction

The Stereo Problem

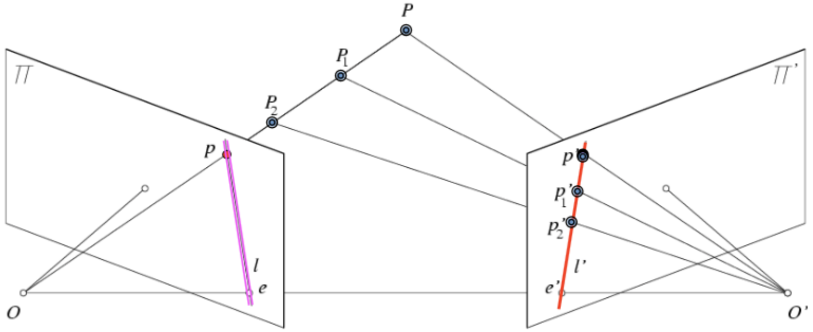
– Shape from two (or more) images

– Biological motivation



known camera viewpoints

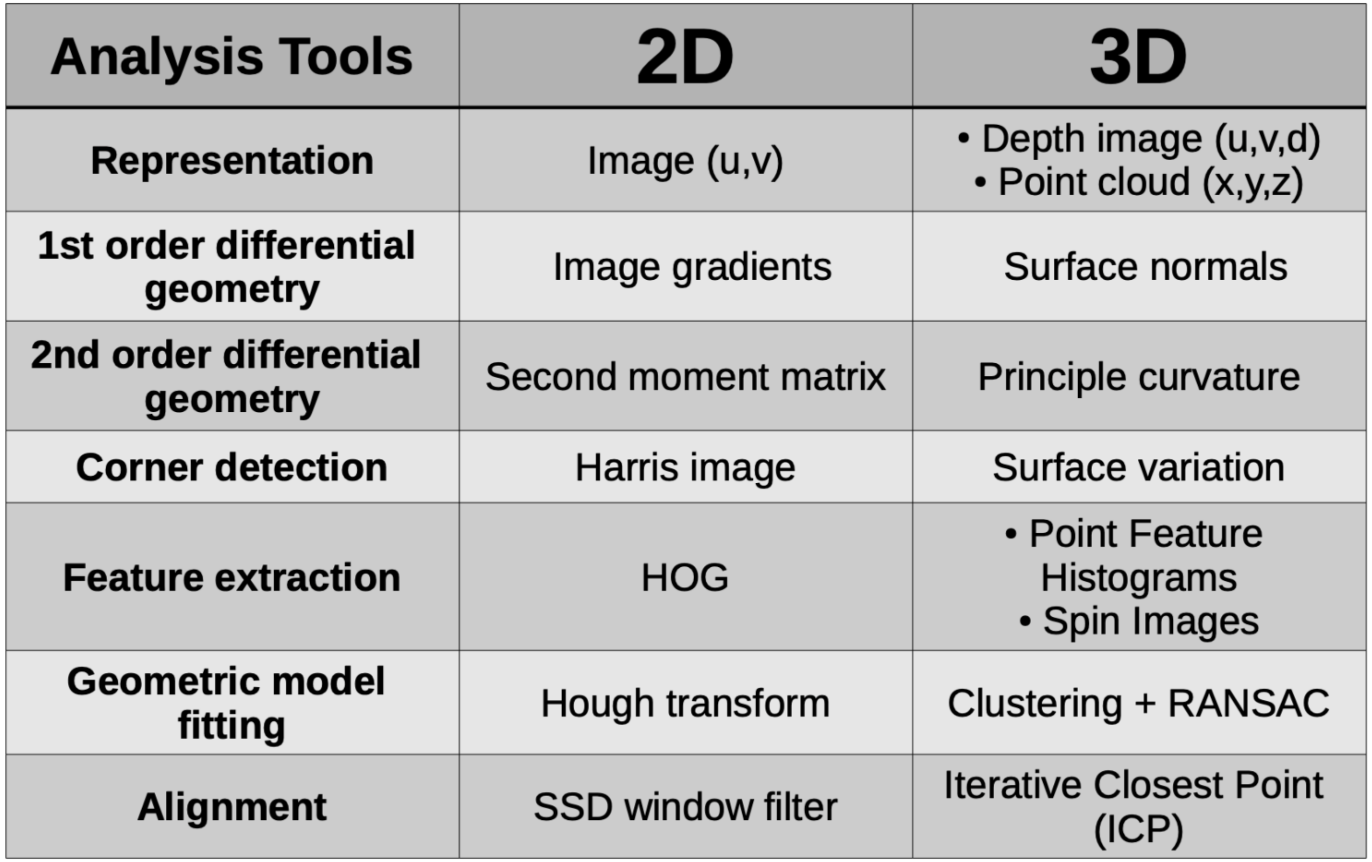
Epipolar Constraint



Credit: Fei-fei Li

Introduction to Mono/Stereo/3D Vision

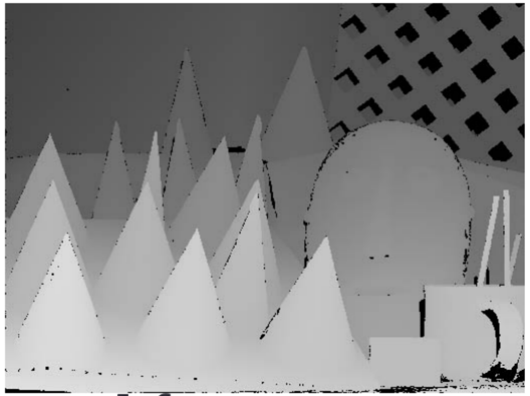
- Comparison: 2D vs 3D



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

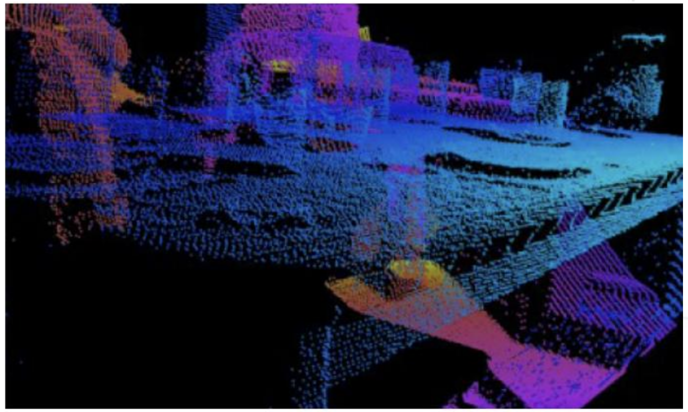
Depth Images



- 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

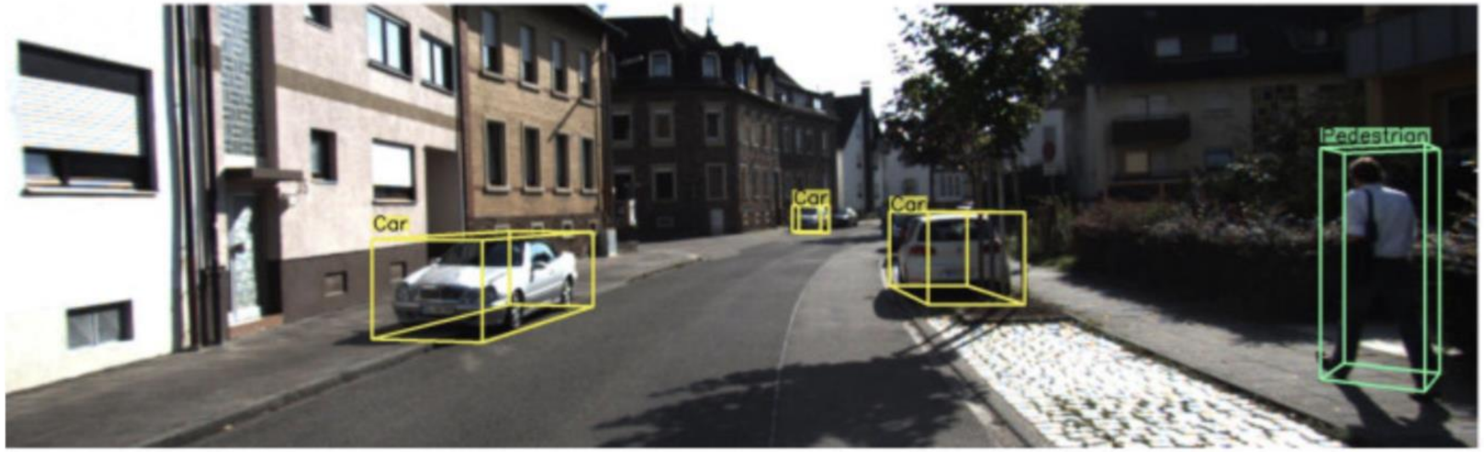
Part 1 3D Vision Basics

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.*



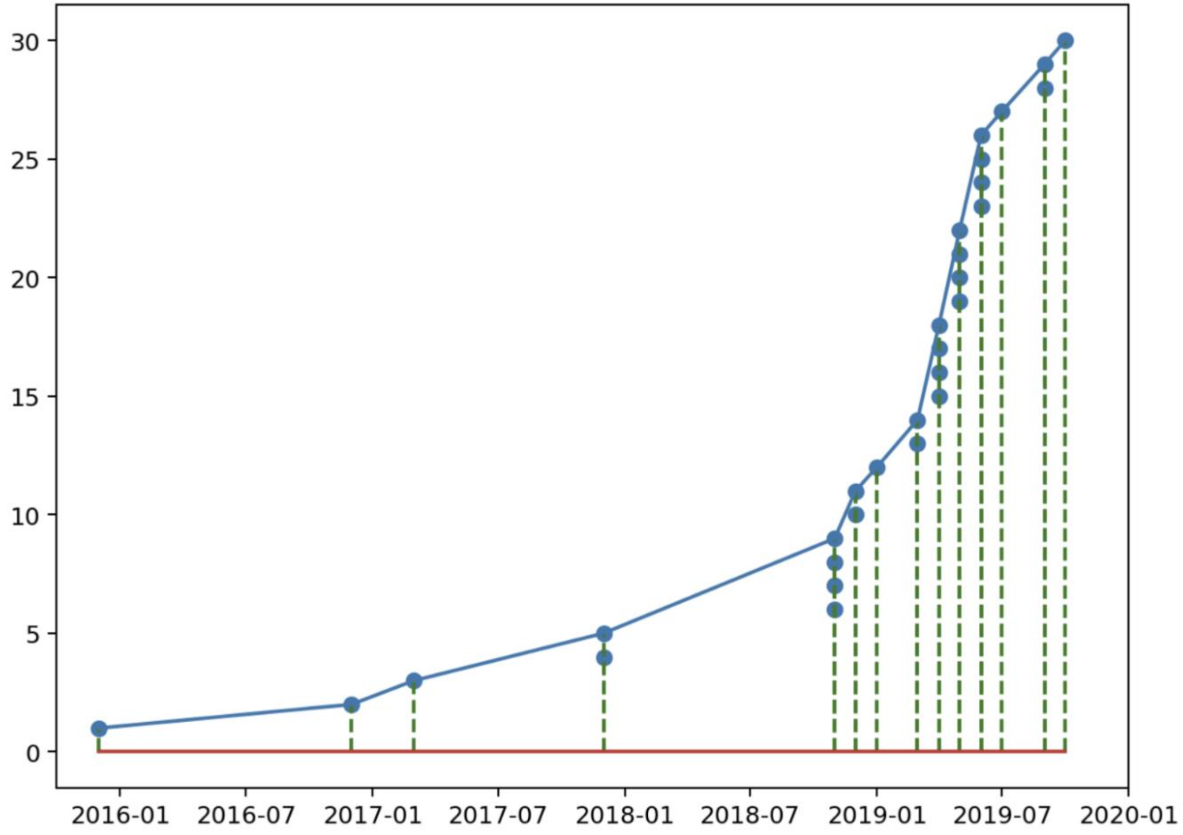
Monocular 3D Object Detection draws 3D bounding boxes on RGB images (source: M3D-RPN)

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

3D Object Detection

- Monocular 3D object detection with RGB image

Publication trend on Mono3DOD in Autonomous Driving



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.

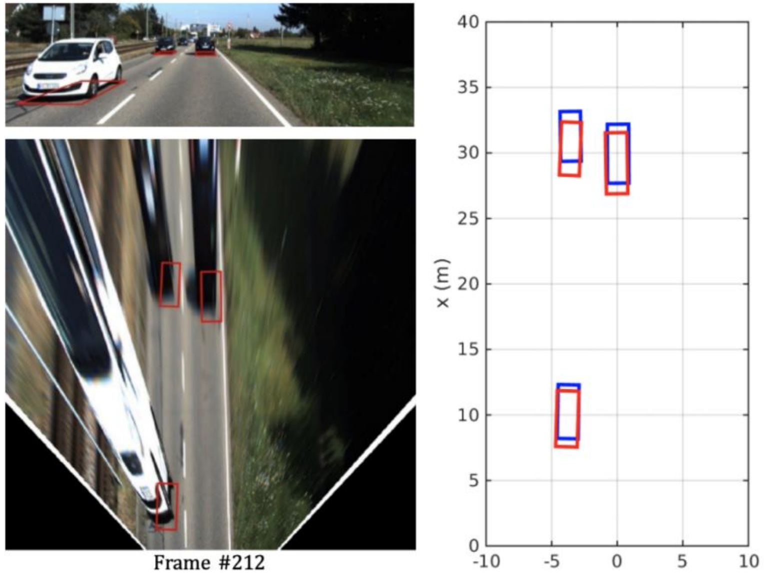
3D Object Detection

- **Representation transformation: BEV/pseudo-lidar**

Why BEV? Scale or occlusion

BEV method: IPM, inverse perspective mapping

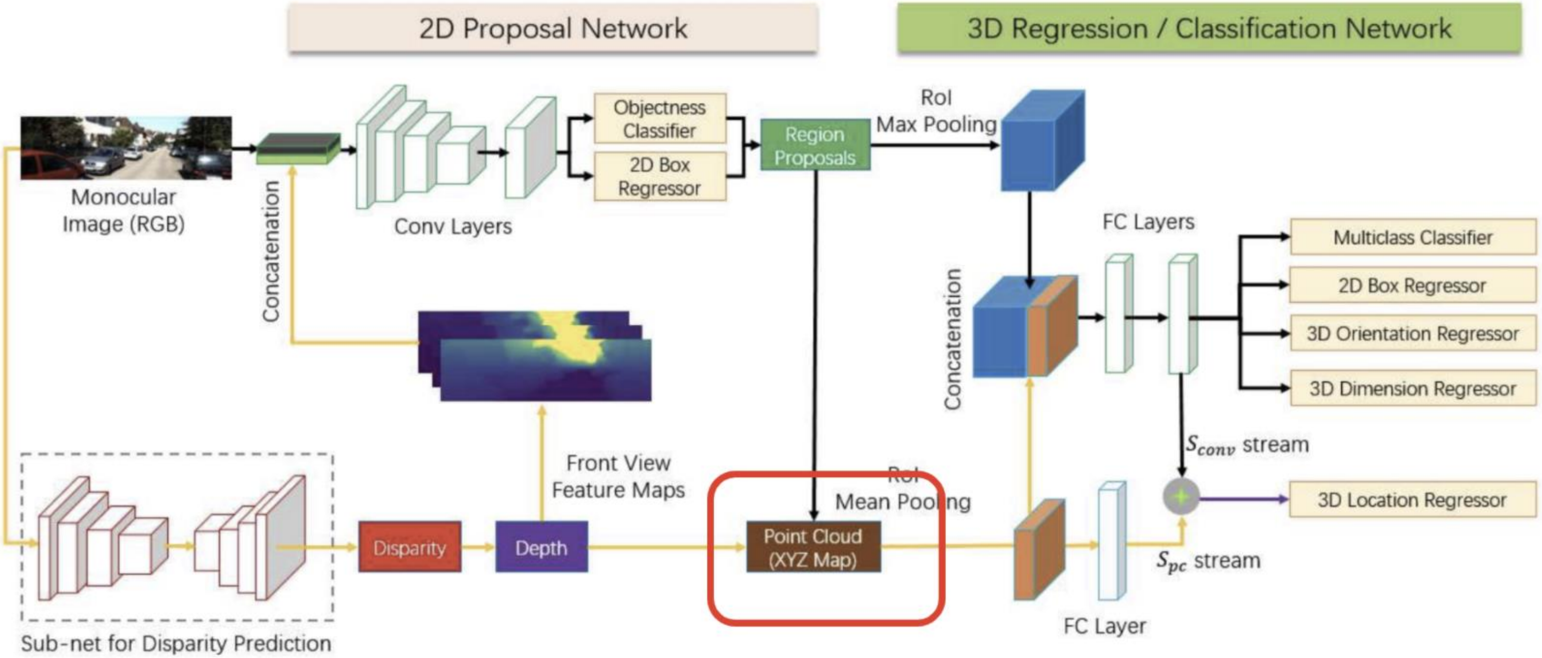
- Pseudo-lidar
  - Idea: to generate point cloud based on the estimated depth from the image
  - Details: using RGBD images as depth, **as the fourth channel** and apply the normal networks on this input, with minimal changes to the first layer.



Convert perspective image to BEV (from BEV-IPM)

3D Object Detection

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



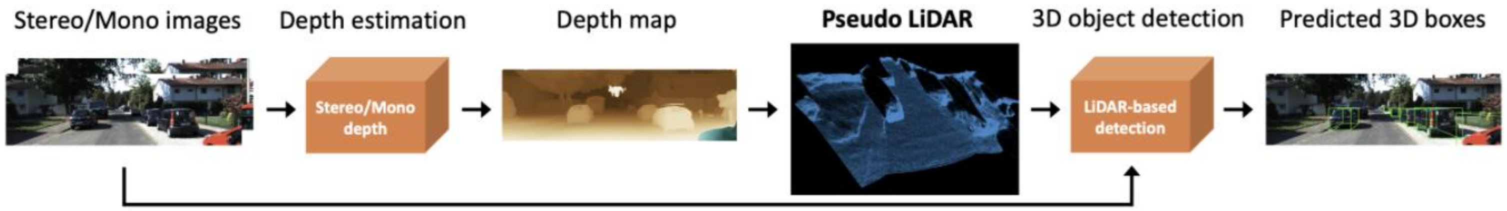
Architecture of Multi-level fusion (MLF), with pseudo-lidar generation circled in red (source)

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.

3D Object Detection

- **Representation transformation: BEV/pseudo-lidar**
- Pseudo-lidar



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

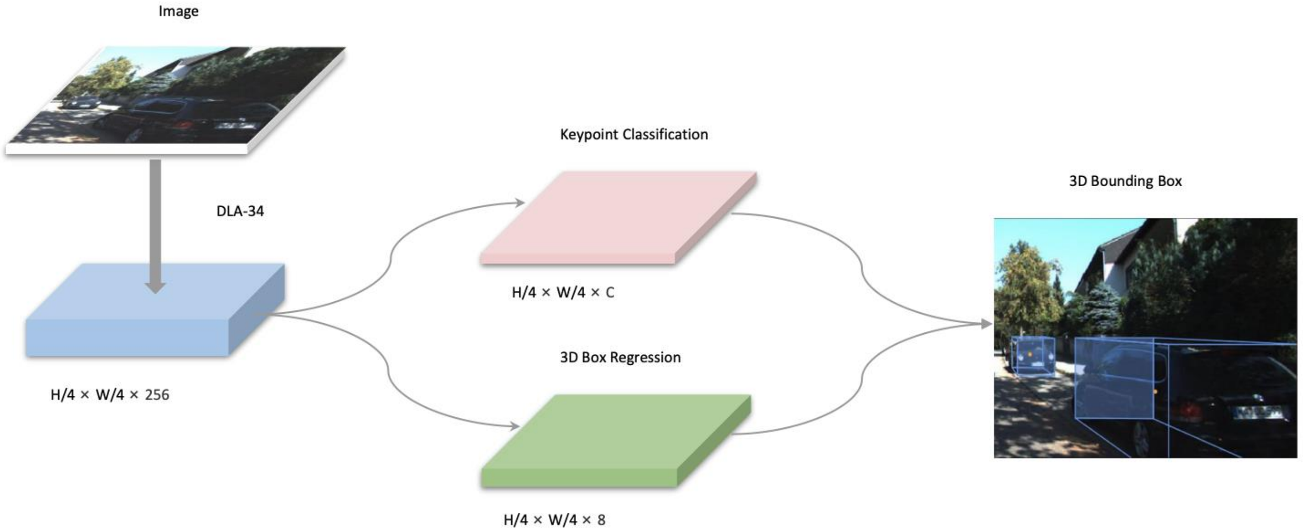
Caveat note:
there is some **overlap** between the training data of DORN, the off-the-shelf depth estimator, and the validation data of pseudo-lidar 3DOD.

3D Object Detection

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.



3D Object Detection

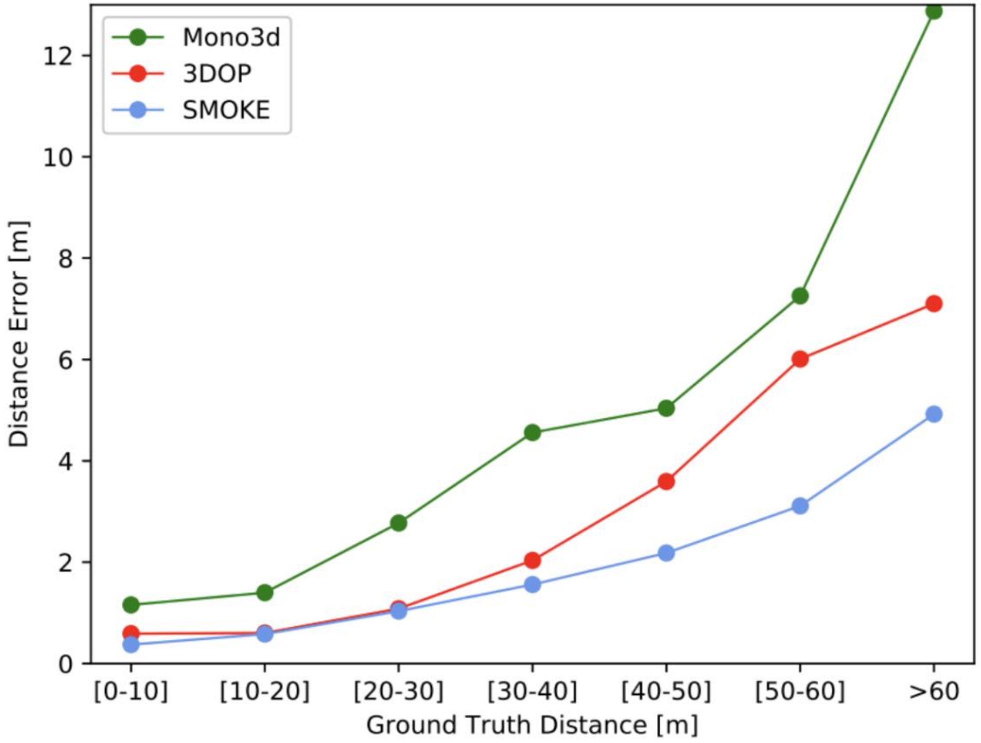
2. Keypoints and shapes

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.



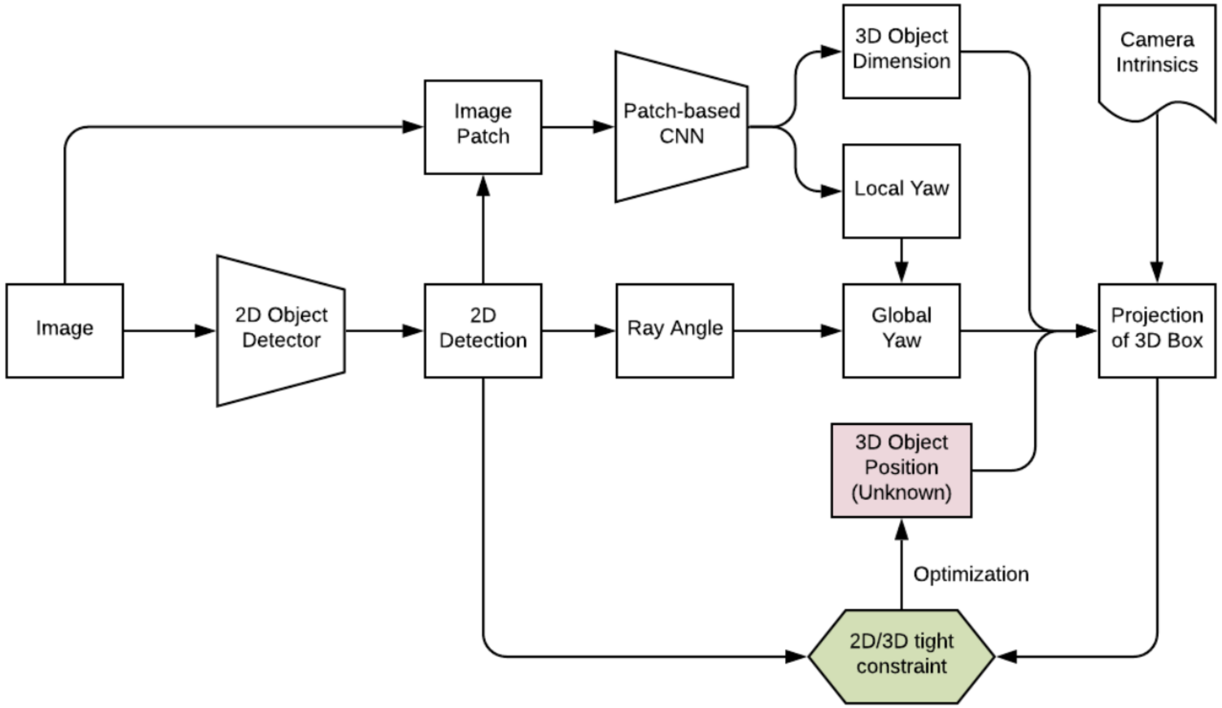
Average depth estimation error by SMOKE

3D Object Detection

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)

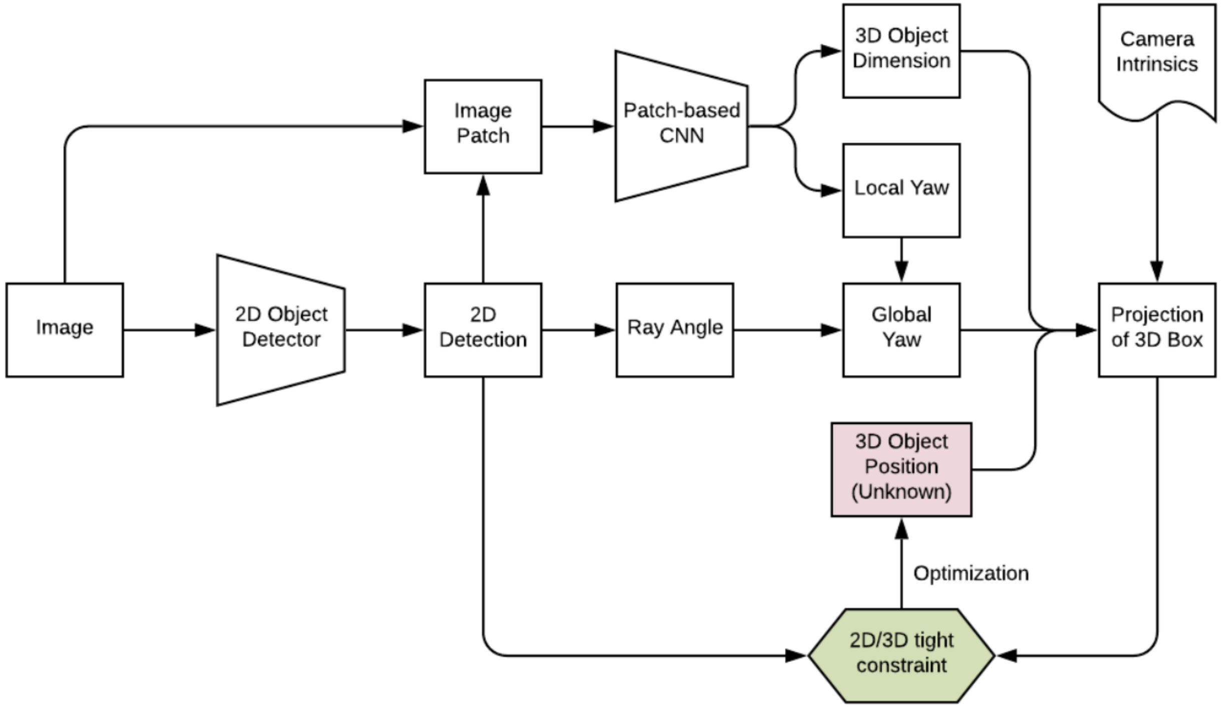
3D Object Detection

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.*



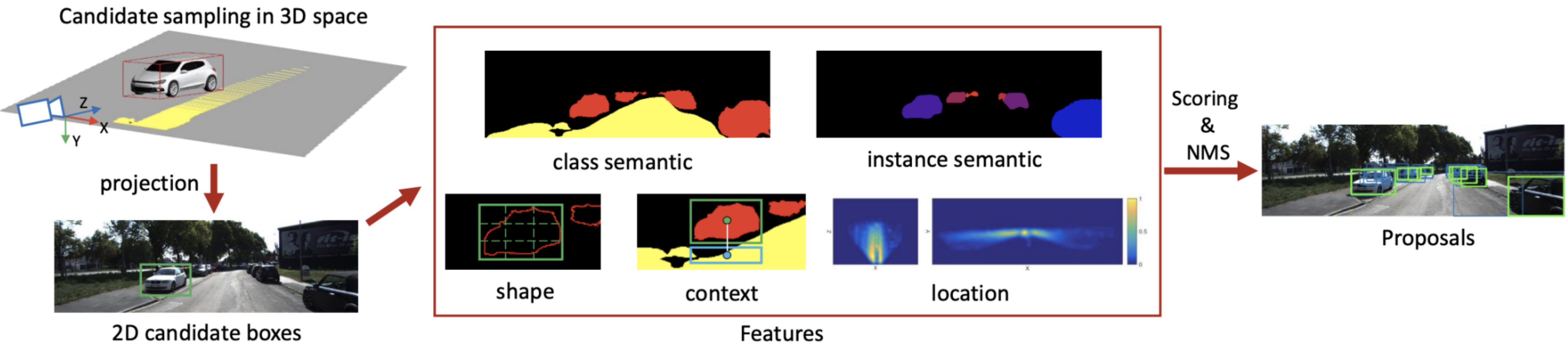
The architecture of deep3DBox, representative of many other similar works (source)

3D Object Detection

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)

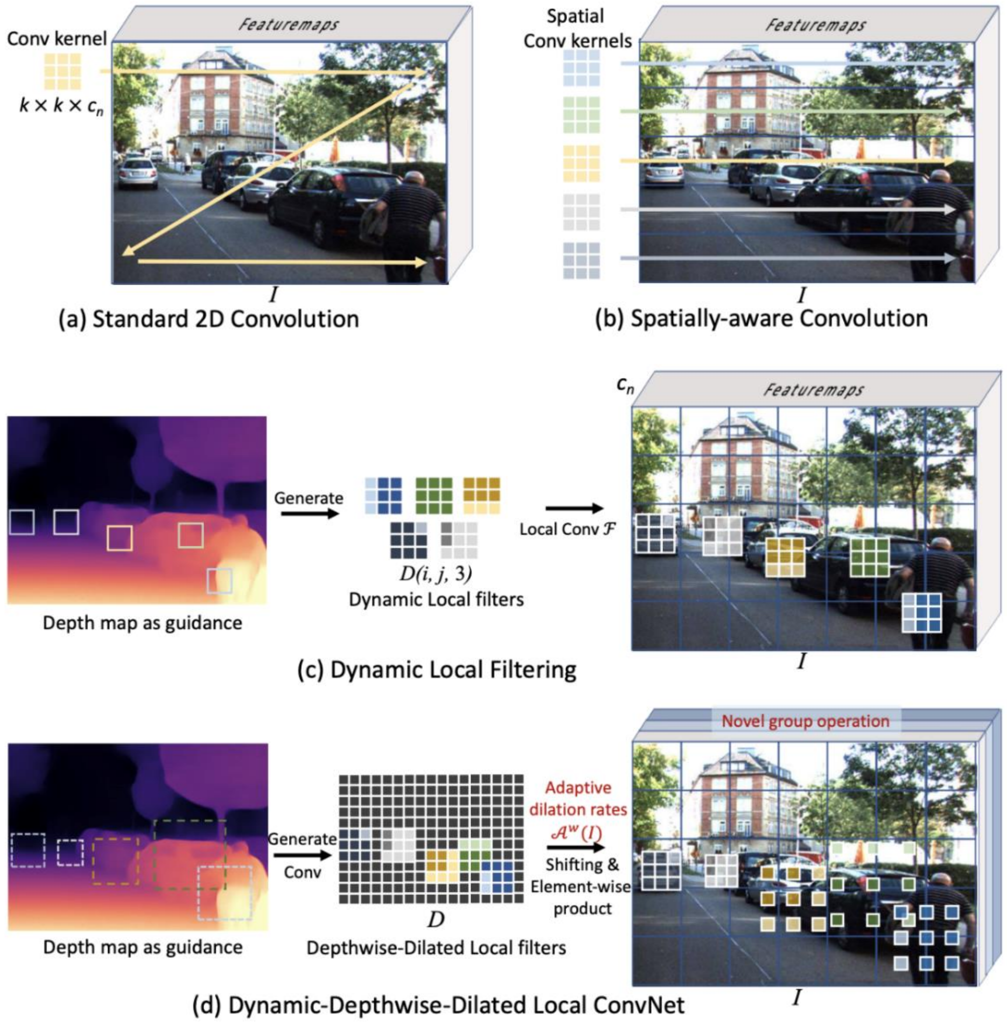
3D Object Detection

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

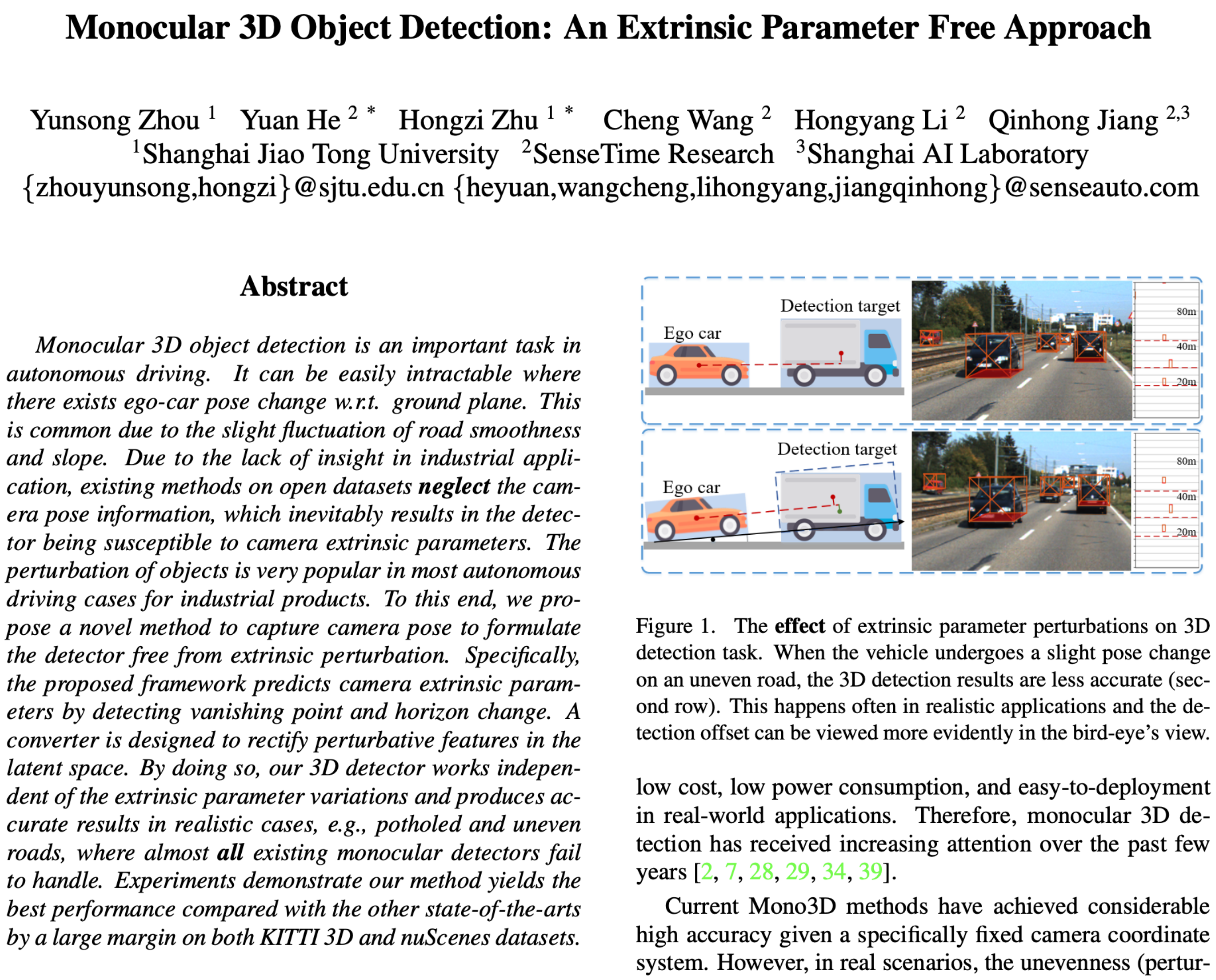
2D and 3D consistency can help regularize joint 2D and 3D training and can help 3D reasoning as a postprocessing step after the prediction of 2D bounding box and geometric hints.

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 Object Detection: takeaways



Monocular 3D Object Detection: An Extrinsic Parameter Free Approach

Yunsong Zhou [1] Yuan He [2] * Hongzi Zhu [1] * Cheng Wang [2] Hongyang Li [2] Qinhong Jiang [2,3]
[1]Shanghai Jiao Tong University [2]SenseTime Research [3]Shanghai AI Laboratory
{zhouyunsong,hongzi}@sjtu.edu.cn {heyuan,wangcheng,lihongyang,jiangqinhong}@senseauto.com

Abstract

*Monocular 3D object detection is an important task in autonomous driving. It can be easily intractable where there exists ego-car pose change w.r.t. ground plane. This is common due to the slight fluctuation of road smoothness and slope. Due to the lack of insight in industrial application, existing methods on open datasets **neglect** the camera pose information, which inevitably results in the detector being susceptible to camera extrinsic parameters. The perturbation of objects is very popular in most autonomous driving cases for industrial products. To this end, we propose a novel method to capture camera pose to formulate the detector free from extrinsic perturbation. Specifically, the proposed framework predicts camera extrinsic parameters by detecting vanishing point and horizon change. A converter is designed to rectify perturbative features in the latent space. By doing so, our 3D detector works independent of the extrinsic parameter variations and produces accurate results in realistic cases, e.g., potholed and uneven roads, where almost **all** existing monocular detectors fail to handle. Experiments demonstrate our method yields the best performance compared with the other state-of-the-arts by a large margin on both KITTI 3D and nuScenes datasets.*

Figure 1. The **effect** of extrinsic parameter perturbations on 3D detection task. When the vehicle undergoes a slight pose change on an uneven road, the 3D detection results are less accurate (second row). This happens often in realistic applications and the detection offset can be viewed more evidently in the bird-eye's view.

low cost, low power consumption, and easy-to-deployment in real-world applications. Therefore, monocular 3D detection has received increasing attention over the past few years [2, 7, 28, 29, 34, 39].

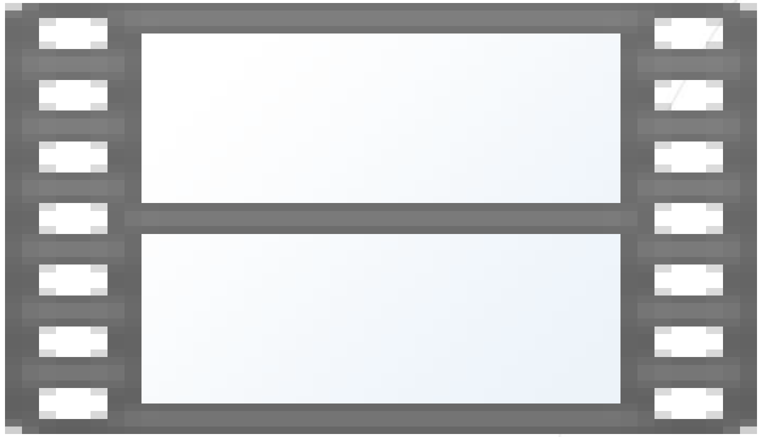
Current Mono3D methods have achieved considerable high accuracy given a specifically fixed camera coordinate system. However, in real scenarios, the unevenness (pertursensetime

Tsinghua University

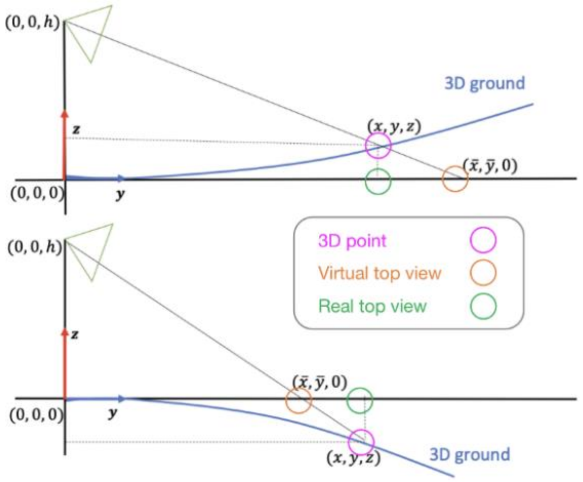
Outline

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

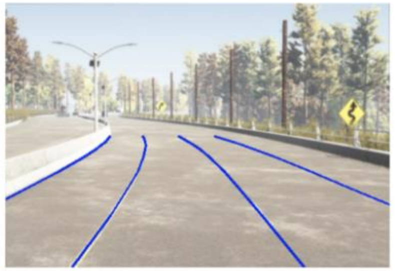
3D Lane Detection: why doing 3D?



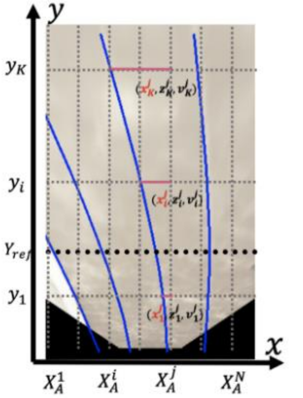
3D Lane Geometry



Perspective image of an uphill road



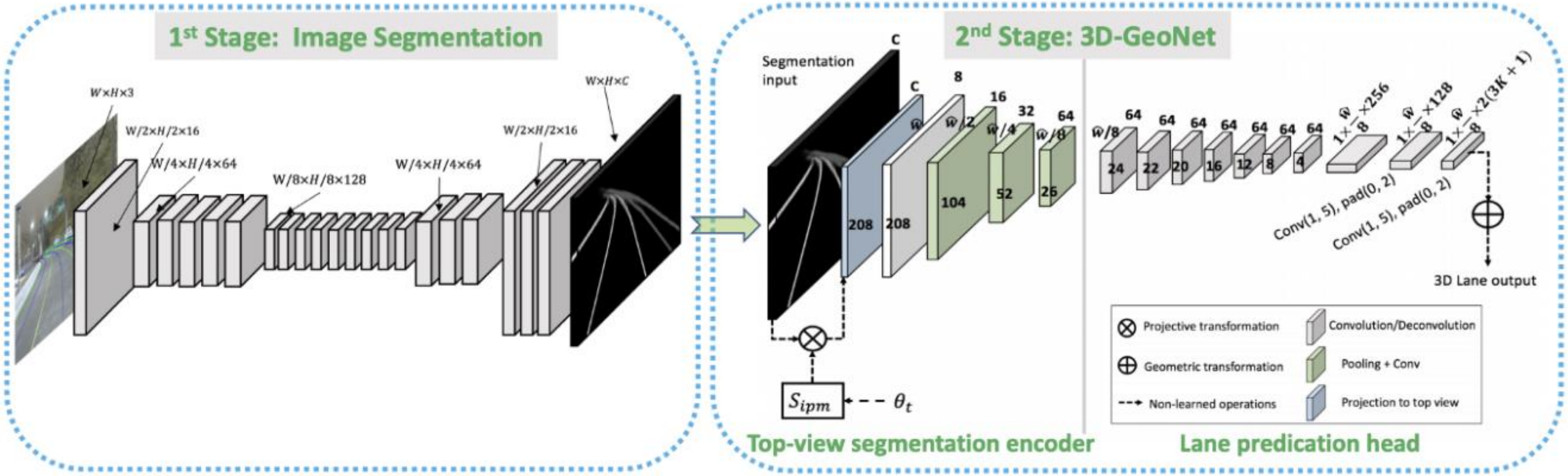
Virtual top view showing diverging lane lines



Gen-LaneNet:

- the top view projection does not align with the IPM transformed features in the presence of a non-zero slope.

3D Lane Detection: pipeline



Network architecture for Gen-LaneNet (source: Gen-LaneNet)

Gen-LaneNet:

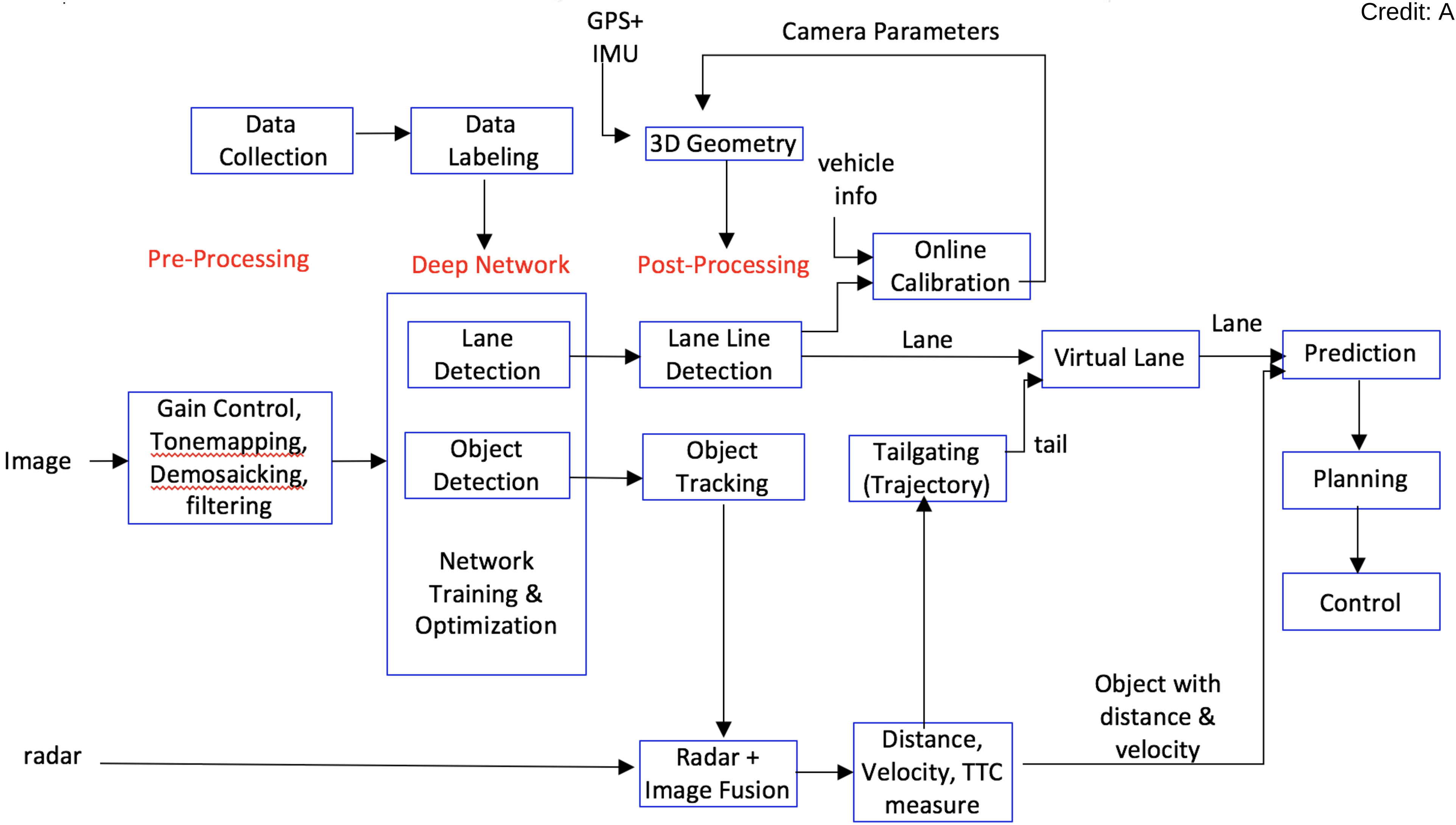
- the top view projection does not align with the IPM transformed features in the presence of a non-zero slope.

One last drop(s) for me

Reach me at lihongyang@senseauto.com

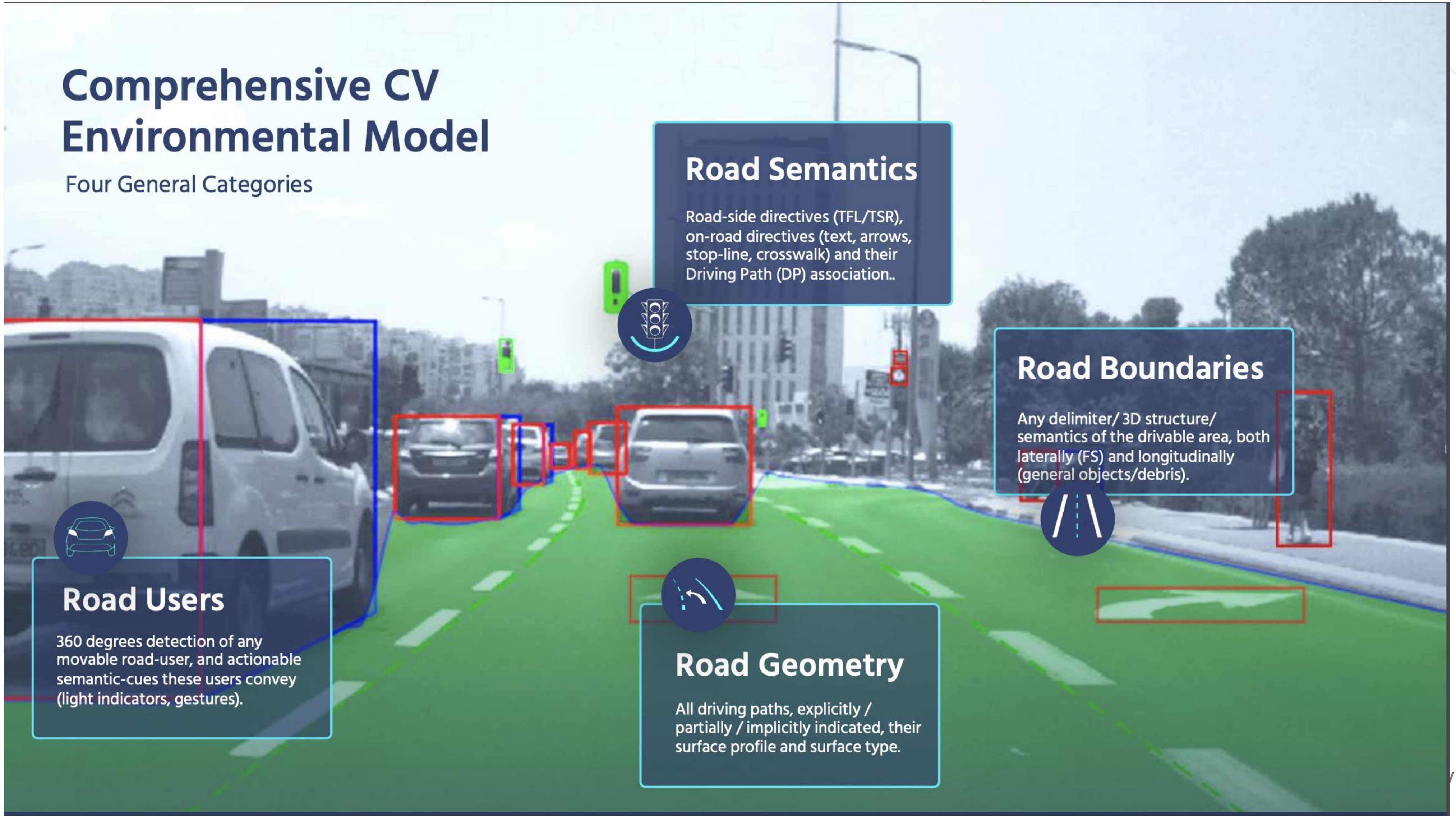
Autonomous driving algorithm pipeline

Credit: Apollo 2.5 version



SOTA: Mobileye Solution

https://s21.q4cdn.com/600692695/files/doc_presentations/2020/1/Mobileye-CES-2020-presentation.pdf



Comprehensive CV Environmental Model

Four General Categories

Road Semantics

Road-side directives (TFL/TSR), on-road directives (text, arrows, stop-line, crosswalk) and their Driving Path (DP) association..

Road Boundaries

Any delimiter/ 3D structure/ semantics of the drivable area, both laterally (FS) and longitudinally (general objects/debris).

Road Users

360 degrees detection of any movable road-user, and actionable semantic-cues these users convey (light indicators, gestures).

Road Geometry

All driving paths, explicitly / partially / implicitly indicated, their surface profile and surface type.

Building up your research career

- How to select research topics?
  - Impactful research
- How to judge the value/contribution of a paper?
  - Solid experiments
  - Issue-driven motivation
  - Other properties
- How to contribute contributions? ideas/novelty
  - How to come up with ideas
- Right ways to conduct a full-round research project
  - Choose topic, paper survey, build baseline, run experiments/formulate idea, write paper, rebuttal, camera-ready, poster/oral session, blog/open-source repo
- Academia vs industry