Review of

Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving

by: Fitria Wulandari

Table of Content

To achieve robust and accurate scene understanding, autonomous vehicles are usually **equipped with different sensors**, such as:

- Camera
- LiDARs (Light Detection and Racing)
- Radars -> used to detect the speed and range ofobjects in the vicinity (around) of the car

Many methods have been proposed for deep multi-modal perception problems

However, **there is no general guideline for network architecture design**, and questions of:

- What to fuse?
- When to fuse?
- How to fuse?

1^{st} autonomous driving in 1980s and DARPA in 2007

It offers high potential to:

- Decrease traffic congestion,
- Improve road safety, and
- Reduce carbon emissions

What needed in the driverless cars:

- Perceive, predict, decide, plan, excuse their decisions in the real world.
- Often in uncontrolled or complex environments
- Cause a small error in the system can cause fatal accidents

Perception systems in driverless cars need to be:

- Accurate: they need to give precise information of driving environments
- Robust: they should work properly in adverse weather, in situations that are not covered during training (open-set conditions), and when some sensors are degraded or even defective
- Real-time: especially when the cars are driving at high speed

Towards these goals, autonomous cars are usually equipped with multi-modal sensors: camera, LiDARs, Radars; and different sensing modalities are fused so that their complementary properties are exploited

And here are the summarizes of the methods that have been proposed for deep multi-modal perception problems on the Object Detection and Semantic Segmentation for Autonomous Driving

DATASETS

Datasets (1)

Name	Sensing Modalities	Year (published)	Labelled (benchmark)	Recording area	Size	Categories / Remarks	Link
Ford AV Dataset	Visual camera (7), 3D LiDAR (4)	2020	6 DoF Pose	Michigan	1.6 TB (amount of frames not given)	; Seasonal variation in weather, lighting, construction and traffic conditions	Dataset Website
Toyota Research Institute DDAD	Visual camera (6), 3D LiDAR	2020	Depth	San Francisco, Bay Area, Cambridge, Detroit, Ann Arbor, Tokyo, Odaiba	Labeled: 99k frames (camera); 200 scenes	Long-range depth (~250m)	Dataset Website
PandaSet	3D LiDAR (2), Visual cameras (6), GNSS and inertial sensors	2020	3D bounding box	San Francisco, El Camino Real	48k frames (camera), 16k frames (LiDAR), 100+ scenes	28 classes, 37 semantic segmentation labels; Solid state LiDAR	Dataset Website
CADC	Visual camera (8), 3D LiDAR	2020	3D bounding boxes	Waterloo (Canada)	Labeled: 56k frames (camera), 7k frames (LiDAR); Raw: 263k frames (camera), 32k frames (LiDAR)	Car, Pedestrian, Truck, Bus, Garbage Containers on Wheels, Traffic Guidance Objects, Bicycle, Pedestrian With Object, Horse and Buggy, Animals; Adverse Weather conditions, different intensities of snowfall	Dataset Website

Datasets (2)

Name	Sensing Modalities	Year (published)	Labelled (benchmark)	Recording area	Size	Categories / Remarks	Link
Astyx HiRes2019	Radar, Visual camera, 3D LiDAR	2019	3D bounding boxes	n.a.	500 frames (5000 annotated objects)	Car, Bus, Cyclist, Motorcyclist, Person, Trailer, Truck	Dataset Website
A2D2	Visual cameras (6); 3D LiDAR (5); Bus data	2019	2D/3D bounding boxes, 2D/3D instance segmentation	Gaimersheim, Ingolstadt, Munich	40k frames (semantics), 12k frames (3D objects), 390k frames unlabeled	Car, Bicycle, Pedestrian, Truck, Small vehicles, Traffic signal, Utility vehicle, Sidebars, Speed bumper, Curbstone, Solid line, Irrelevant signs, Road blocks, Tractor, Non-drivable street, Zebra crossing, Obstacles / trash, Poles, RD restricted area, Animals, Grid structure, Signal corpus, Drivable cobbleston, Electronic traffic, Slow drive area, Nature object, Parking area, Sidewalk, Ego car, Painted driv. instr., Traffic guide obj., Dashed line, RD normal street, Sky, Buildings, Blurred area, Rain dirt	Dataset Website

Datasets (3)

Name	Sensing Modalities	Year (published)	Labelled (benchmark)	Recording area	Size	Categories / Remarks	Link
A*3D Dataset	Visual cameras (2); 3D LiDAR	2019	3D bounding boxes	Singapore	39k frames, 230k objects	Car, Van, Bus, Truck, Pedestrians, Cyclists, and Motorcyclists; Afternoon and night, wet and dry	Dataset Website
EuroCity Persons	Visual camera; Announced: stereo, LiDAR, GNSS and intertial sensors	2019	2D bounding boxes	12 countries in Europe, 27 cities	47k frames, 258k objects	Pedestrian, Rider, Bicycle, Motorbike, Scooter, Tricycle, Wheelchair, Buggy, Co-Rider; Highly diverse: 4 seasons, day and night, wet and dry	Dataset Website
Oxford RobotCar	2016: Visual cameras (fisheye & stereo), 2D & 3D LiDAR, GNSS, and inertial sensors; 2019: Radar, 3D Lidar (2), 2D LiDAR (2), visual cameras (6), GNSS and inertial sensors	2016, 2019	no	Oxford	2016: 11,070,651 frames (stereo), 3,226,183 frames (3D LiDAR); 2019: 240k scans (Radar), 2.4M frames (LiDAR)	Long-term autonomous driving. Various weather conditions, including heavy rain, night, direct sunlight and snow.	Dataset Website 2016, Dataset Website 2019
Waymo Open Dataset	3D LiDAR (5), Visual cameras (5)	2019	3D bounding box, Tracking	n.a.	200k frames, 12M objects (3D LiDAR), 1.2M objects (2D camera)	Vehicles, Pedestrians, Cyclists, Signs	Dataset Website

Datasets (4)

Name	Sensing Modalities	Year (published)	Labelled (benchmark)	Recording area	Size	Categories / Remarks	Link
Lyft Level 5 AV Dataset 2019	3D LiDAR (5), Visual cameras (6)	2019	3D bounding box	n.a.	55k frames	Semantic HD map included	Dataset Website
Argoverse	3D LiDAR (2), Visual cameras (9, 2 stereo)	2019	3D bounding box, Tracking, Forecasting	Pittsburgh, Pennsylvania, Miami, Florida	113 scenes, 300k trajectories	Vehicle, Pedestrian, Other Static, Large Vehicle, Bicycle, Bicyclist, Bus, Other Mover, Trailer, Motorcyclist, Moped, Motorcycle, Stroller, Emergency Vehicle, Animal, Wheelchair, School Bus; Semantic HD maps (2) included	Dataset Website
nuScenes dataset	Visual cameras (6), 3D LiDAR, Radars (5)	2019	3D bounding box	Boston, Singapore	1000 scenes, 1.4M frames (camera, Radar), 390k frames (3D LiDAR)	Car or Van or SUV, Truck, Pickup Truck, Front Of Semi Truck, Bendy Bus, Rigid Bus, Construction Vehicle, Motorcycle, Bicycle, Bicycle Rack, Trailer, Police Vehicle, Ambulance, Train, Adult Pedestrian, Child Pedestrian, Construction Worker, Stroller, Wheelchair, Portable Personal Mobility Vehicle, Traffic Police, Other Police, Animal, Traffic Cone, Temporary Traffic Barrier, Pushable Pullable Object, Debris	Dataset Website

Datasets (5)

Name	Sensing Modalities	Year (published)	Labelled (benchmark)	Recording area	Size	Categories / Remarks	Link
BLVD	Visual (Stereo) camera, 3D LiDAR	2019	3D bounding box, Tracking, Interaction, Intention	Changshu	120k frames, 249,129 objects	Vehicle, Pedestrian, Rider during day and night	Dataset Website
H3D dataset	Visual cameras (3), 3D LiDAR	2019	3D bounding box	San Francisco	27,721 frames, 1,071,302 objects	Car, Pedestrian, Cyclist, Truck, Misc, Animal, Motorcyclist, Bus	Dataset Website
ApolloScape	Visual (Stereo) camera, 3D LiDAR, GNSS and inertial sensors	2018, 2019	2D/3D pixel-level segmentation, lane marking, instance segmentation, Depth	n.a.	143,906 frames, 89,430 objects	Rover, Sky, Car, Motobicycle, Bicycle, Person, Rider, Truck, Bus, Tricycle, Road, Sidewalk, Traffic Cone, Road Pile, Fence, Traffic Light, Pole, Traffic Sign, Wall, Dustbin, Billboard, Building, Bridge, Tunnel, Overpass, Vegetation	Dataset Website
DBNet dataset	3D LiDAR, Dashboard visual camera, GNSS	2018	Driving behaviours (Vehicle speed and wheel angles)	Multiple areas in China	Over 10k frames	In total seven datasets with different test scenarios, such as seaside roads, school areas, mountain roads	Dataset Website

Datasets (6)

Name	Sensing Modalities	Year (published)	Labelled (benchmark)	Recording area	Size	Categories / Remarks	Link
KAIST multispectral dataset	Visual (Stereo) and thermal camera, 3D LiDAR, GNSS and inertial sensors	2018	2D bounding box, drivable region, image enhancement, depth, colorization	Seoul	7,512 frames, 308,913 objects	Person, Cyclist, Car during day and night, fine time slots (sunrise, afternoon,...)	Dataset Website
Multi-spectral Object Detection dataset	Visual and thermal cameras	2017	2D bounding box	University environment in Japan	7,512 frames, 5,833 objects	Bike, Car, Car Stop, Color Cone, Person during day and night	Dataset Website
Multi-spectral Semantic Segmentation dataset	Visual and thermal camera	2017	2D pixel-level segmentation	n.a.	1569 frames	Bike, Car, Person, Curve, Guardrail, Color Cone, Bump during day and night	Dataset Website
Multi-modal Panoramic 3D Outdoor (MPO) dataset	Visual camera, LiDAR and GNSS	2016	Place categorization	Fukuoka	650 scans (dense), 34200 scans (sparse)	No dynamic objects	Dataset Website
KAIST multispectral pedestrian	Visual and thermal camera	2015	2D bounding box	Seoul	95,328 frames, 103,128 objects	Person, People, Cyclist during day and night	Dataset Website

Datasets (7)

Name	Sensing Modalities	Year (published)	Labelled (benchmark)	Recording area	Size	Categories / Remarks	Link
KITTI	Visual (Stereo) camera, 3D LiDAR, GNSS and inertial sensors	2012, 2013, 2015	2D, 3D bounding box, visual odometry, road detection, optical flow, tracking, depth, 2D instance and pixel-level segmentation	Karlsruhe	7481 frames (training) 80.256 objects	Car, Van, Truck, Pedestrian, Person (sitting), Cyclist, Tram, Misc	Dataset Website
The Málaga Stereo and Laser Urban dataset	Visual (Stereo) camera, 5x 2D LiDAR (yielding 3D information), GNSS and inertial sensors	2014	no	Málaga	113,082 frames, 5,654.6 s (camera); >220,000 frames, ~5,000 s (LiDARs)	n.a.	Dataset Website

DETECTIONS

Detection 2D (1)

Reference	Sensors	Object Type	Sensing Modality Representations and Processing	Network Pipeline	How to generate Region Proposals (RP)	When to fuse	Fusion Operation and Method	Fusion Level	Dataset(s) used
Nabati *et al.*, 2019 [pdf]	Radar, visual camera	2D Vehicle	Radar object, RGB image. Radar projected to image frame.	Fast R-CNN	Radar used to generate region proposal	Implicit at RP	Region proposal	Middle	nuScenes
Bijelic *et al.*, 2019 [pdf]	LiDAR, visual camera	2D Car in foggy weather	Lidar front view images (depth, intensity, height), RGB image. Each processed by VGG16	SSD	Predictions with fused features	Before RP	Feature concatenation	From early to middle layers	Self-recorded datasets focused on foggy weather, simulated foggy images from KITTI
Chadwick *et al.*, 2019 [pdf]	Radar, visual camera	2D Vehicle	Radar range and velocity maps, RGB image. Each processed by ResNet	One stage detector	Predictions with fused features	Before RP	Addition, feature concatenation	Middle	Self-recorded
Pfeuffer *et al.*, 2018 [pdf]	LiDAR, vision camera	Multiple 2D objects	LiDAR spherical, and front-view sparse depth, dense depth image, RGB image. Each processed by VGG16	Faster-RCNN	RPN from fused features	Before RP	Feature concatenation	Early, Middle, Late	KITTI

Detection 2D (2)

Reference	Sensors	Object Type	Sensing Modality Representations and Processing	Network Pipeline	How to generate Region Proposals (RP)	When to fuse	Fusion Operation and Method	Fusion Level	Dataset(s) used
Kim *et al.*, 2018 [pdf]	LiDAR, vision camera	2D Car	LiDAR front-view depth image, RGB image. Each input processed by VGG16	SSD	SSD with fused features	Before RP	Feature concatenation, Mixture of Experts	Middle	KITTI
Guan *et al.*, 2018 [pdf]	Vision camera, thermal camera	2D Pedestrian	RGB image, thermal image. Each processed by a base network built on VGG16	Faster-RCNN	RPN with fused features	Before and after RP	Feature concatenation, Mixture of Experts	Early, Middle, Late	KAIST Pedestrian Dataset
Asvadi *et al.*, 2017 [pdf]	LiDAR, vision camera	2D Car	LiDAR front-view dense-depth (DM) and reflectance maps (RM), RGB image. Each processed through a YOLO net	YOLO	YOLO outputs for LiDAR DM and RM maps, and RGB image	After RP	Ensemble: feed engineered features from ensembled bounding boxes to a network to predict scores for NMS	Late	KITTI

Detection 2D (3)

Reference	Sensors	Object Type	Sensing Modality Representations and Processing	Network Pipeline	How to generate Region Proposals (RP)	When to fuse	Fusion Operation and Method	Fusion Level	Dataset(s) used
Oh *et al.*, 2017 [pdf]	LiDAR, vision camera	2D Car, Pedestrian, Cyclist	LiDAR front-view dense-depth map (for fusion: processed by VGG16), LiDAR voxel (for ROIs: segmentation and region growing), RGB image (for fusion: processed by VGG16; for ROIs: segmentation and grouping)	R-CNN	LiDAR voxel and RGB image separately	After RP	Association matrix using basic belief assignment	Late	KITTI
Du *et al.*, 2017 [pdf]	LiDAR, vision camera	2D Car	LiDAR voxel (processed by RANSAC and model fitting), RGB image (processed by VGG16 and GoogLeNet)	Faster-RCNN	First clustered by LiDAR point clouds, then fine-tuned by a RPN of RGB image	Before RP	Ensemble: feed LiDAR RP to RGB image-based CNN for final prediction	Late	KITTI
Schneider *et al.*, 2017 [pdf]	Vision camera	Multiple 2D objects	RGB image (processed by GoogLeNet), depth image from stereo camera (processed by NiN net)	SSD	SSD predictions.	Before RP	Feature concatenation	Early, Middle, Late	Cityscape

Detection 2D (4)

Reference	Sensors	Object Type	Sensing Modality Representations and Processing	Network Pipeline	How to generate Region Proposals (RP)	When to fuse	Fusion Operation and Method	Fusion Level	Dataset(s) used
Takumi *et al.*, 2017 [pdf]	Vision camera, thermal camera	Multiple 2D objects	RGB image, NIR, FIR, FIR image. Each processed by YOLO	YOLO	YOLO predictions for each spectral image	After RP	Ensemble: ensemble final predictions for each YOLO detector	Late	self-recorded data
Matti *et al.*, 2017 [pdf]	LiDAR, vision camera	2D Pedestrian	LiDAR points (clustering with DBSCAN) and RGB image (processed by ResNet)	R-CNN	Clustered by LiDAR point clouds, then size and ratio corrected on RGB image.	Before and at RP	Ensemble: feed LiDAR RP to RGB image-based CNN for final prediction	Late	KITTI
Schlosser *et al.*, 2016 [pdf]	LiDAR, vision camera	2D Pedestrian	LiDAR HHA image, RGB image. Each processed by a small ConvNet	R-CNN	Deformable Parts Model with RGB image	After RP	Feature concatenation	Early, Middle, Late	KITTI
Kim *et al.*, 2016 [pdf]	LiDAR, vision camera	2D Pedestrian, Cyclist	LiDAR front-view depth image, RGB image. Each processed by Fast-RCNN network	Fast-RCNN	Selective search for LiDAR and RGB image separately.	At RP	Ensemble: joint RP are fed to RGB image based CNN.	Late	KITTI

Detection 2D (5)

Reference	Sensors	Object Type	Sensing Modality Representations and Processing	Network Pipeline	How to generate Region Proposals (RP)	When to fuse	Fusion Operation and Method	Fusion Level	Dataset(s) used
Mees *et al.*, 2016 [pdf]	RGB-D camera	2D Pedestrian	RGB image, depth image from depth camera, optical flow. Each processed by GoogLeNet	Fast-RCNN	Dense multi-scale sliding window for RGB image	After RP	Mixture of Experts	Late	RGB-D People Unihall Dataset, InOutDoor RGB-D People Dataset.
Wagner *et al.*, 2016 [pdf]	Vision camera, thermal camera	2D Pedestrian	RGB image, thermal image. Each processed by CaffeeNet	R-CNN	ACF+T+THOG detector	After RP	Feature concatenation	Early, Late	KAIST Pedestrian Dataset
Liu *et al.*, 2016 [pdf]	Vision camera, thermal camera	2D Pedestrian	RGB image, thermal image. Each processed by NiN network	Faster-RCNN	RPN with fused (or separate) features	Before and after RP	Feature concatenation, average mean, Score fusion (Cascaded CNN)	Early, Middle, Late	KAIST Pedestrian Dataset

Detection 3D (1)

Reference	Sensors	Object Type	Sensing Modality Representations and Processing	Network Pipeline	How to generate Region Proposals (RP)	When to fuse	Fusion Operation and Method	Fusion Level	Dataset(s) used
Meyer and Kuschk, 2019 [pdf]	Radar, visual camera	3D Vehicle	Radar pointcloud, RGB image. Fused features extracted from CNN.	Faster R-CNN	Before and after RP	Average mean	Region proposal	Early, Middle	Astyx HiRes2019
Liang *et al.*, 2019 [pdf]	LiDAR, visual camera	3D Car, Pedestrian, Cyclist	LiDAR BEV maps, RGB image. Each processed by a ResNet with auxiliary tasks: depth estimation and ground segmentation	Faster R-CNN	Predictions with fused features	Before RP	Addition, continuous fusion layer	Middle	KITTI, self-recorded
Wang *et al.*, 2019 [pdf]	LiDAR, visual camera	3D Car, Pedestrian, Cyclist, Indoor objects	LiDAR voxelized frustum (each frustum processed by the PointNet), RGB image (using a pre-trained detector).	R-CNN	Pre-trained RGB image detector	After RP	Using RP from RGB image detector to build LiDAR frustums	Late	KITTI, SUN-RGBD
Dou *et al.*, 2019 [pdf]	LiDAR, visual camera	3D Car	LiDAR voxel (processed by VoxelNet), RGB image (processed by a FCN to get semantic features)	Two stage detector	Predictions with fused features	Before RP	Feature concatenation	Middle	KITTI

Detection 3D (2)

Reference	Sensors	Object Type	Sensing Modality Representations and Processing	Network Pipeline	How to generate Region Proposals (RP)	When to fuse	Fusion Operation and Method	Fusion Level	Dataset(s) used
Sindagi *et al.*, 2019 [pdf]	LiDAR, visual camera	3D Car	LiDAR voxel (processed by VoxelNet), RGB image (processed by a pre-trained 2D image detector).	One stage detector	Predictions with fused features	Before RP	Feature concatenation	**Early**, Middle	KITTI
Liang *et al.*, 2018 [pdf]	LiDAR, vision camera	3D Car, Pedestrian, Cyclist	LiDAR BEV maps, RGB image. Each processed by ResNet	One stage detector	Predictions with fused features.	Before RP	Addition, continuous fusion layer	Middle	KITTI, self-recorded
Du *et al.*, 2018 [pdf]	LiDAR, vision camera	3D Car	LiDAR voxel (processed by RANSAC and model fitting), RGB image (processed by VGG16 and GoogLeNet)	R-CNN	Pre-trained RGB image detector produces 2D bounding boxes to crop LiDAR points, which are then clustered	Before and at RP	Ensemble: use RGB image detector to regress car dimensions for a model fitting algorithm.	Late	KITTI, self-recorded data

Detection 3D (3)

Reference	Sensors	Object Type	Sensing Modality Representations and Processing	Network Pipeline	How to generate Region Proposals (RP)	When to fuse	Fusion Operation and Method	Fusion Level	Dataset(s) used
Yang *et al.*, 2018 [pdf]	LiDAR, HD-map	3D Car	LiDAR BEV maps, Road mask image from HD map. Inputs processed by PIXOR++ with the backbone similar to FPN	One stage detector	Detector predictions	Before RP	Feature concatenation	Early	KITTI, TOR4D Dataset
Casas *et al.*, 2018 [pdf]	LiDAR, HD-map	3D Car	sequential LiDAR BEV maps, sequential several road topology mask images from HD map. Each input processed by a base network with residual blocks	One stage detector	Detector predictions	Before RP	Feature concatenation	Middle	self-recorded data
Shin *et al.*, 2018 [pdf]	LiDAR, vision camera	3D Car	LiDAR point clouds, (processed by PointNet); RGB image (processed by a 2D CNN)	R-CNN	A 3D object detector for RGB image	After RP	Using RP from RGB image detector to search LiDAR point clouds	Late	KITTI
Chen *et al.*, 2017 [pdf]	LiDAR, vision camera	3D Car	LiDAR BEV and spherical maps, RGB image. Each processed by a base network built on VGG16	Faster-RCNN	A RPN from LiDAR BEV map	After RP	average mean, deep fusion	Early, Middle, Late	KITTI

Detection 3D (4)

Reference	Sensors	Object Type	Sensing Modality Representations and Processing	Network Pipeline	How to generate Region Proposals (RP)	When to fuse	Fusion Operation and Method	Fusion Level	Dataset(s) used
Wang *et al.*, 2017 [pdf]	LiDAR, vision camera	3D Car, Pedestrian	LiDAR BEV map, RGB image. Each processed by a RetinaNet	One stage detector	Fused LiDAR and RGB image features extracted from CNN	Before RP	Sparse mean manipulation	Middle	KITTI
Ku *et al.*, 2017 [pdf]	LiDAR, vision camera	3D Car, Pedestrian, Cyclist	LiDAR BEV map, RGB image. Each processed by VGG16	Faster-RCNN	Fused LiDAR and RGB image features extracted from CNN	Before and after RP	Average mean	Early, Middle, Late	KITTI
Xu *et al.*, 2017 [pdf]	LiDAR, vision camera	3D Car, Pedestrian, Cyclist, Indoor objects	LiDAR points (processed by PointNet), RGB image (processed by ResNet)	R-CNN	Pre-trained RGB image detector	After RP	Feature concatenation for local and global features	Middle	KITTI, SUN-RGBD
Qi *et al.*, 2017 [pdf]	LiDAR, vision camera	3D Car, Pedestrian, Cyclist, Indoor objects	LiDAR points (processed by PointNet), RGB image (using a pre-trained detector)	R-CNN	Pre-trained RGB image detector	After RP	Feature concatenation	Middle, Late	KITTI, SUN-RGBD

Detection Thermal

Reference	Sensors	Object Type	Sensing Modality Representations and Processing	Network Pipeline	How to generate Region Proposals (RP)	When to fuse	Fusion Operation and Method	Fusion Level	Dataset(s) used
Guan *et al.*, 2018 [pdf]	Vision camera, thermal camera	2D Pedestrian	RGB image, thermal image. Each processed by a base network built on VGG16	Faster-RCNN	RPN with fused features	Before and after RP	Feature concatenation, Mixture of Experts	Early, Middle, Late	KAIST Pedestrian Dataset
Takumi *et al.*, 2017 [pdf]	Vision camera, thermal camera	Multiple 2D objects	RGB image, NIR, FIR, FIR image. Each processed by YOLO	YOLO	YOLO predictions for each spectral image	After RP	Ensemble: ensemble final predictions for each YOLO detector	Late	self-recorded data
Wagner *et al.*, 2016 [pdf]	Vision camera, thermal camera	2D Pedestrian	RGB image, thermal image. Each processed by CaffeeNet	R-CNN	ACF+T+THOG detector	After RP	Feature concatenation	Early, Late	KAIST Pedestrian Dataset
Liu *et al.*, 2016 [pdf]	Vision camera, thermal camera	2D Pedestrian	RGB image, thermal image. Each processed by NiN network	Faster-RCNN	RPN with fused (or separate) features	Before and after RP	Feature concatenation, average mean, Score fusion (Cascaded CNN)	Early, Middle, Late	KAIST Pedestrian Dataset

Detection LiDAR (1)

Reference	Sensors	Object Type	Sensing Modality Representations and Processing	Network Pipeline	How to generate Region Proposals (RP)	When to fuse	Fusion Operation and Method	Fusion Level	Dataset(s) used
Liang *et al.*, 2019 [pdf]	LiDAR, visual camera	3D Car, Pedestrian, Cyclist	LiDAR BEV maps, RGB image. Each processed by a ResNet with auxiliary tasks: depth estimation and ground segmentation	Faster R-CNN	Predictions with fused features	Before RP	Addition, continuous fusion layer	Middle	KITTI, self-recorded
Wang *et al.*, 2019 [pdf]	LiDAR, visual camera	3D Car, Pedestrian, Cyclist, Indoor objects	LiDAR voxelized frustum (each frustum processed by the PointNet), RGB image (using a pre-trained detector).	R-CNN	Pre-trained RGB image detector	After RP	Using RP from RGB image detector to build LiDAR frustums	Late	KITTI, SUN-RGBD
Dou *et al.*, 2019 [pdf]	LiDAR, visual camera	3D Car	LiDAR voxel (processed by VoxelNet), RGB image (processed by a FCN to get semantic features)	Two stage detector	Predictions with fused features	Before RP	Feature concatenation	Middle	KITTI
Sindagi *et al.*, 2019 [pdf]	LiDAR, visual camera	3D Car	LiDAR voxel (processed by VoxelNet), RGB image (processed by a pre-trained 2D image detector).	One stage detector	Predictions with fused features	Before RP	Feature concatenation	**Early**, Middle	KITTI

Detection LiDAR (2)

Reference	Sensors	Object Type	Sensing Modality Representations and Processing	Network Pipeline	How to generate Region Proposals (RP)	When to fuse	Fusion Operation and Method	Fusion Level	Dataset(s) used
Bijelic *et al.*, 2019 [pdf]	LiDAR, visual camera	2D Car in foggy weather	Lidar front view images (depth, intensity, height), RGB image. Each processed by VGG16	SSD	Predictions with fused features	Before RP	Feature concatenation	From early to middle layers	Self-recorded datasets focused on foggy weather, simulated foggy images from KITTI
Pfeuffer *et al.*, 2018 [pdf]	LiDAR, vision camera	Multiple 2D objects	LiDAR spherical, and front-view sparse depth, dense depth image, RGB image. Each processed by VGG16	Faster-RCNN	RPN from fused features	Before RP	Feature concatenation	Early, Middle, Late	KITTI
Liang *et al.*, 2018 [pdf]	LiDAR, vision camera	3D Car, Pedestrian, Cyclist	LiDAR BEV maps, RGB image. Each processed by ResNet	One stage detector	Predictions with fused features.	Before RP	Addition, continuous fusion layer	Middle	KITTI, self-recorded

Detection LiDAR (3)

Reference	Sensors	Object Type	Sensing Modality Representations and Processing	Network Pipeline	How to generate Region Proposals (RP)	When to fuse	Fusion Operation and Method	Fusion Level	Dataset(s) used
Du *et al.*, 2018 [pdf]	LiDAR, vision camera	3D Car	LiDAR voxel (processed by RANSAC and model fitting), RGB image (processed by VGG16 and GoogLeNet)	R-CNN	Pre-trained RGB image detector produces 2D bounding boxes to crop LiDAR points, which are then clustered	Before and at RP	Ensemble: use RGB image detector to regress car dimensions for a model fitting algorithm.	Late	KITTI, self-recorded data
Kim *et al.*, 2018 [pdf]	LiDAR, vision camera	2D Car	LiDAR front-view depth image, RGB image. Each input processed by VGG16	SSD	SSD with fused features	Before RP	Feature concatenation, Mixture of Experts	Middle	KITTI
Yang *et al.*, 2018 [pdf]	LiDAR, HD-map	3D Car	LiDAR BEV maps, Road mask image from HD map. Inputs processed by PIXOR++ with the backbone similar to FPN	One stage detector	Detector predictions	Before RP	Feature concatenation	Early	KITTI, TOR4D Dataset

Detection LiDAR (4)

Reference	Sensors	Object Type	Sensing Modality Representations and Processing	Network Pipeline	How to generate Region Proposals (RP)	When to fuse	Fusion Operation and Method	Fusion Level	Dataset(s) used
Casas *et al.*, 2018 [pdf]	LiDAR, HD-map	3D Car	sequential LiDAR BEV maps, sequential several road topology mask images from HD map. Each input processed by a base network with residual blocks	One stage detector	Detector predictions	Before RP	Feature concatenation	Middle	self-recorded data
Shin *et al.*, 2018 [pdf]	LiDAR, vision camera	3D Car	LiDAR point clouds, (processed by PointNet); RGB image (processed by a 2D CNN)	R-CNN	A 3D object detector for RGB image	After RP	Using RP from RGB image detector to search LiDAR point clouds	Late	KITTI
Chen *et al.*, 2017 [pdf]	LiDAR, vision camera	3D Car	LiDAR BEV and spherical maps, RGB image. Each processed by a base network built on VGG16	Faster-RCNN	A RPN from LiDAR BEV map	After RP	average mean, deep fusion	Early, Middle, Late	KITTI

Detection LiDAR (5)

Reference	Sensors	Object Type	Sensing Modality Representations and Processing	Network Pipeline	How to generate Region Proposals (RP)	When to fuse	Fusion Operation and Method	Fusion Level	Dataset(s) used
Asvadi *et al.*, 2017 [pdf]	LiDAR, vision camera	2D Car	LiDAR front-view dense-depth (DM) and reflectance maps (RM), RGB image. Each processed through a YOLO net	YOLO	YOLO outputs for LiDAR DM and RM maps, and RGB image	After RP	Ensemble: feed engineered features from ensembled bounding boxes to a network to predict scores for NMS	Late	KITTI
Oh *et al.*, 2017 [pdf]	LiDAR, vision camera	2D Car, Pedestrian, Cyclist	LiDAR front-view dense-depth map (for fusion: processed by VGG16), LiDAR voxel (for ROIs: segmentation and region growing), RGB image (for fusion: processed by VGG16; for ROIs: segmentation and grouping)	R-CNN	LiDAR voxel and RGB image separately	After RP	Association matrix using basic belief assignment	Late	KITTI
Wang *et al.*, 2017 [pdf]	LiDAR, vision camera	3D Car, Pedestrian	LiDAR BEV map, RGB image. Each processed by a RetinaNet	One stage detector	Fused LiDAR and RGB image features extracted from CNN	Before RP	Sparse mean manipulation	Middle	KITTI

Detection LiDAR (6)

Reference	Sensors	Object Type	Sensing Modality Representations and Processing	Network Pipeline	How to generate Region Proposals (RP)	When to fuse	Fusion Operation and Method	Fusion Level	Dataset(s) used
Ku *et al.*, 2017 [pdf]	LiDAR, vision camera	3D Car, Pedestrian, Cyclist	LiDAR BEV map, RGB image. Each processed by VGG16	Faster-RCNN	Fused LiDAR and RGB image features extracted from CNN	Before and after RP	Average mean	Early, Middle, Late	KITTI
Xu *et al.*, 2017 [pdf]	LiDAR, vision camera	3D Car, Pedestrian, Cyclist, Indoor objects	LiDAR points (processed by PointNet), RGB image (processed by ResNet)	R-CNN	Pre-trained RGB image detector	After RP	Feature concatenation for local and global features	Middle	KITTI, SUN-RGBD
Qi *et al.*, 2017 [pdf]	LiDAR, vision camera	3D Car, Pedestrian, Cyclist, Indoor objects	LiDAR points (processed by PointNet), RGB image (using a pre-trained detector)	R-CNN	Pre-trained RGB image detector	After RP	Feature concatenation	Middle, Late	KITTI, SUN-RGBD
Du *et al.*, 2017 [pdf]	LiDAR, vision camera	2D Car	LiDAR voxel (processed by RANSAC and model fitting), RGB image (processed by VGG16 and GoogLeNet)	Faster-RCNN	First clustered by LiDAR point clouds, then fine-tuned by a RPN of RGB image	Before RP	Ensemble: feed LiDAR RP to RGB image-based CNN for final prediction	Late	KITTI

Detection LiDAR (7)

Reference	Sensors	Object Type	Sensing Modality Representations and Processing	Network Pipeline	How to generate Region Proposals (RP)	When to fuse	Fusion Operation and Method	Fusion Level	Dataset(s) used
Matti *et al.*, 2017 [pdf]	LiDAR, vision camera	2D Pedestrian	LiDAR points (clustering with DBSCAN) and RGB image (processed by ResNet)	R-CNN	Clustered by LiDAR point clouds, then size and ratio corrected on RGB image.	Before and at RP	Ensemble: feed LiDAR RP to RGB image-based CNN for final prediction	Late	KITTI
Schlosser *et al.*, 2016 [pdf]	LiDAR, vision camera	2D Pedestrian	LiDAR HHA image, RGB image. Each processed by a small ConvNet	R-CNN	Deformable Parts Model with RGB image	After RP	Feature concatenation	Early, Middle, Late	KITTI
Kim *et al.*, 2016 [pdf]	LiDAR, vision camera	2D Pedestrian, Cyclist	LiDAR front-view depth image, RGB image. Each processed by Fast-RCNN network	Fast-RCNN	Selective search for LiDAR and RGB image separately.	At RP	Ensemble: joint RP are fed to RGB image based CNN.	Late	KITTI

Detection Radar

Reference	Sensors	Object Type	Sensing Modality Representations and Processing	Network Pipeline	How to generate Region Proposals (RP)	When to fuse	Fusion Operation and Method	Fusion Level	Dataset(s) used
Meyer and Kuschk, 2019 [pdf]	Radar, visual camera	3D Vehicle	Radar pointcloud, RGB image. Fused features extracted from CNN.	Faster R-CNN	Before and after RP	Average mean	Region proposal	Early, Middle	Astyx HiRes2019
Nabati *et al.*, 2019 [pdf]	Radar, visual camera	2D Vehicle	Radar object, RGB image. Radar projected to image frame.	Fast R-CNN	Radar used to generate region proposal	Implicit at RP	Region proposal	Middle	nuScenes
Chadwick *et al.*, 2019 [pdf]	Radar, visual camera	2D Vehicle	Radar range and velocity maps, RGB image. Each processed by ResNet	One stage detector	Predictions with fused features	Before RP	Addition, feature concatenation	Middle	Self-recorded

SEGMENTATIONS

Segmentation 2D (1)

Reference	Sensors	Semantics	Sensing Modality Representations	Fusion Operation and Method	Fusion Level	Dataset(s) used
Chen *et al.*, 2019 [pdf]	LiDAR, visual camera	Road segmentation	RGB image, altitude difference image. Each processed by a CNN	Feature adaptation module, modified concatenation.	Middle	KITTI
Valada *et al.*, 2019 [pdf]	Visual camera, depth camera, thermal camera	Multiple 2D objects	RGB image, thermal image, depth image. Each processed by FCN with ResNet backbone (Adapnet++ architecture)	Extension of Mixture of Experts	Middle	Six datasets, including Cityscape, Sun RGB-D, etc.
Sun *et al.*, 2019 [pdf]	Visual camera, thermal camera	Multiple 2D objects in campus environments	RGB image, thermal image. Each processed by a base network built on ResNet	Element-wise summation in the encoder networks	Middle	Datasets published by [pdf]
Caltagirone *et al.*, 2019 [pdf]	LiDAR, vision camera	Road segmentation	LiDAR front-view depth images, RGB image. Each input processed by a FCN	Feature concatenation (For early and late fusion), weighted addition similar to gating network (for middle-level cross fusion)	Early, Middle, Late	KITTI
Erkent *et al.*, 2018 [pdf]	LiDAR, visual camera	Multiple 2D objects	LiDAR BEV occupancy grids (processed based on Bayesian filtering and tracking), RGB image (processed by a FCN with VGG16 backbone)	Feature concatenation	Middle	KITTI, self-recorded

Segmentation 2D (2)

Reference	Sensors	Semantics	Sensing Modality Representations	Fusion Operation and Method	Fusion Level	Dataset(s) used
Lv *et al.*, 2018 [pdf]	LiDAR, vision camera	Road segmentation	LiDAR BEV maps, RGB image. Each input processed by a FCN with dilated convolution operator. RGB image features are alo projected onto LiDAR BEV plane before fusion	Feature concatenation	Middle	KITTI
Wulff *et al.*, 2018 [pdf]	LiDAR, vision camera	Road segmentation. Alternatives: freespace, ego-lane detection	LiDAR BEV maps, RGB image projected onto BEV plane. Inputs processed by a FCN with UNet	Feature concatenation	Early	KITTI
Kim *et al.*, 2018 [pdf]	LiDAR, vision camera	2D Off-road terrains	LiDAR voxel (processed by 3D convolution), RGB image (processed by ENet)	Addition	Early, Middle, Late	self-recorded
Guan *et al.*, 2018 [pdf]	Vision camera, thermal camera	2D Pedestrian	RGB image, thermal image. Each processed by a base network built on VGG16	Feature concatenation, Mixture of Experts	Early, Middle, Late	KAIST Pedestrian Dataset
Yang *et al.*, 2018 [pdf]	LiDAR, vision camera	Road segmentation	LiDAR points (processed by PointNet++), RGB image (processed by FCN with VGG16 backbone)	Optimizing Conditional Random Field (CRF)	Late	KITTI

Segmentation 2D (3)

Reference	Sensors	Semantics	Sensing Modality Representations	Fusion Operation and Method	Fusion Level	Dataset(s) used
Gu *et al.*, 2018 [pdf]	LiDAR, visual camera	Road segmentation	LiDAR front-view depth and height maps (processed by a inverse-depth histogram based line scanning strategy), RGB image (processed by a FCN).	Optimizing Conditional Random Field	Late	KITTI
Cai *et al.*, 2018 [pdf]	Satellite map with route information, visual camera	Road segmentation	Route map image, RGB image. Images are fused and processed by a FCN	Overlaying the line and curve segments in the route map onto the RGB image to generate the Map Fusion Image (MFI)	Early	self-recorded data
Ha *et al.*, 2017 [pdf]	Vision camera, thermal camera	Multiple 2D objects in campus environments	RGB image, thermal image. Each processed by a FCN and mini-inception block	Feature concatenation, addition (``short-cut fusion'')	Middle	self-recorded data
Valada *et al.*, 2017 [pdf]	Vision camera, thermal camera	Multiple 2D objects	RGB image, thermal image, depth image. Each processed by FCN with ResNet backbone	Mixture of Experts	Late	Cityscape, Freiburg Multispectral Dataset, Synthia

Segmentation 2D (4)

Reference	Sensors	Semantics	Sensing Modality Representations	Fusion Operation and Method	Fusion Level	Dataset(s) used
Schneider *et al.*, 2017 [pdf]	Vision camera	Multiple 2D Objects	RGB image, depth image RGB image (processed by GoogLeNet), depth image from stereo camera (processed by NiN net)	Feature concatenation	Early, Middle, Late	Cityscape
Valada *et al.*, 2016 [pdf]	Vision camera, thermal camera	Multiple 2D objects in forested environments	RGB image, thermal image, depth image. Each processed by the UpNet (built on VGG16 and up-convolution)	Feature concatenation, addition	Early, Late	self-recorded data

Segmentation Thermal (1)

Reference	Sensors	Semantics	Sensing Modality Representations	Fusion Operation and Method	Fusion Level	Dataset(s) used
Valada *et al.*, 2019 [pdf]	Visual camera, depth camera, thermal camera	Multiple 2D objects	RGB image, thermal image, depth image. Each processed by FCN with ResNet backbone (Adapnet++ architecture)	Extension of Mixture of Experts	Middle	Six datasets, including Cityscape, Sun RGB-D, etc.
Sun *et al.*, 2019 [pdf]	Visual camera, thermal camera	Multiple 2D objects in campus environments	RGB image, thermal image. Each processed by a base network built on ResNet	Element-wise summation in the encoder networks	Middle	Datasets published by [pdf]
Guan *et al.*, 2018 [pdf]	Vision camera, thermal camera	2D Pedestrian	RGB image, thermal image. Each processed by a base network built on VGG16	Feature concatenation, Mixture of Experts	Early, Middle, Late	KAIST Pedestrian Dataset
Ha *et al.*, 2017 [pdf]	Vision camera, thermal camera	Multiple 2D objects in campus environments	RGB image, thermal image. Each processed by a FCN and mini-inception block	Feature concatenation, addition (“short-cut fusion”)	Middle	self-recorded data

Segmentation Thermal (2)

Reference	Sensors	Semantics	Sensing Modality Representations	Fusion Operation and Method	Fusion Level	Dataset(s) used
Valada *et al.*, 2017 [pdf]	Vision camera, thermal camera	Multiple 2D objects	RGB image, thermal image, depth image. Each processed by FCN with ResNet backbone	Mixture of Experts	Late	Cityscape, Freiburg Multispectral Dataset, Synthia
Valada *et al.*, 2016 [pdf]	Vision camera, thermal camera	Multiple 2D objects in forested environments	RGB image, thermal image, depth image. Each processed by the UpNet (built on VGG16 and up-convolution)	Feature concatenation, addition	Early, Late	self-recorded data

Segmentation LiDAR (1)

Reference	Sensors	Semantics	Sensing Modality Representations	Fusion Operation and Method	Fusion Level	Dataset(s) used
Chen *et al.*, 2019 [pdf]	LiDAR, visual camera	Road segmentation	RGB image, altitude difference image. Each processed by a CNN	Feature adaptation module, modified concatenation.	Middle	KITTI
Caltagirone *et al.*, 2019 [pdf]	LiDAR, vision camera	Road segmentation	LiDAR front-view depth images, RGB image. Each input processed by a FCN	Feature concatenation (For early and late fusion), weighted addition similar to gating network (for middle-level cross fusion)	Early, Middle, Late	KITTI
Erkent *et al.*, 2018 [pdf]	LiDAR, visual camera	Multiple 2D objects	LiDAR BEV occupancy grids (processed based on Bayesian filtering and tracking), RGB image (processed by a FCN with VGG16 backbone)	Feature concatenation	Middle	KITTI, self-recorded
Lv *et al.*, 2018 [pdf]	LiDAR, vision camera	Road segmentation	LiDAR BEV maps, RGB image. Each input processed by a FCN with dilated convolution operator. RGB image features are alo projected onto LiDAR BEV plane before fusion	Feature concatenation	Middle	KITTI

Segmentation LiDAR (2)

Reference	Sensors	Semantics	Sensing Modality Representations	Fusion Operation and Method	Fusion Level	Dataset(s) used
Wulff *et al.*, 2018 [pdf]	LiDAR, vision camera	Road segmentation. Alternatives: freespace, ego-lane detection	LiDAR BEV maps, RGB image projected onto BEV plane. Inputs processed by a FCN with UNet	Feature concatenation	Early	KITTI
Kim *et al.*, 2018 [pdf]	LiDAR, vision camera	2D Off-road terrains	LiDAR voxel (processed by 3D convolution), RGB image (processed by ENet)	Addition	Early, Middle, Late	self-recorded
Yang *et al.*, 2018 [pdf]	LiDAR, vision camera	Road segmentation	LiDAR points (processed by PointNet++), RGB image (processed by FCN with VGG16 backbone)	Optimizing Conditional Random Field (CRF)	Late	KITTI
Gu *et al.*, 2018 [pdf]	LiDAR, visual camera	Road segmentation	LiDAR front-view depth and height maps (processed by a inverse-depth histogram based line scanning strategy), RGB image (processed by a FCN).	Optimizing Conditional Random Field	Late	KITTI

VEHICLES

Vehicles (1)

Demonstration	Year	Route	Vehicle	Sensor Setup	Perception Range
Daimler2013	2013	Bertha Benz Memorial Route from Mannheim to Pforzheim (approx. 65 miles). The route comprises rural roads, urban areas, small villages and various traffic situations (e.g. intersections)	Mercedes Benz S class	Stereo camera on the front, two mono cameras (front and back), 4 short-range radars, 4 long-range radars	360 view± 200 meter for radar, ±130m for camera, ± 80m for stereo camera, ±40 m short-range radar
BMW2015	2015	Firstly, the system was thoroughly tested on a test track. "The first successful automated trip between Munich and Ingolstadt, without driver intervention, occurred on June 16th, 2011. Since then, thousands of kilometers of automated driving experience on highways have been achieved."	BMW 5	Differential GPS, four laser scanner (two 4-layer, two single-layer), three radar, four ultrasonic and a mono camera	Values for views and ranges not published. "The laser scanner sensors provide a complete surround view of the vehicle's environment without any gaps. The radar sensors in the front and the rear enable long range detection of vehicles and obstacles. The ultrasonic sensors on the side provide a redundant source for detecting close vehicles directly to the side. The mono camera in the front is able to reliably classify obstacles, such as vehicles, and detect lane markings for localization."

Vehicles (2)

Demonstration	Year	Route	Vehicle	Sensor Setup	Perception Range
Ulm2015	2015	5 km route around the campus of Ulm University. It includes traffic lights, crosswalks, and roundabouts. Half of the track does not contain any lane markings. Huge amount of pedestrians and other vulnerable road users. Vehicles suddenly turning into the ego vehicle's driving path (especially during rush hour). The speed limit on the course varies between 50 and 70 km/h.	Mercedes-Benz E-Class	Three IBEO LUX laser scanners in the front, a forward-facing monochrome camera (Baumer TXG14f), a long-range front radar (Continental ARS 310), two mid-range rear radar (Bosch MRR), a long-range rear radar (Bosch LRR 3 FMCW), two rearward facing cameras and a real-time kinematic (RTK) system in combination with a differential GPS	Laser scanners: 210° view and maximum range of up to 200 m. Monochrome front camera: 56° view. long-range front radar: range from 0.25 m to 200 m. Mid-range rear radar: 150° view and up to 90 m range. Long-range rear radar: 30° view and 250 m range. Rear camera (Baumer TXG14f): 20° view. Rear camera (Baumer TXG06): 56° view.
Stanford 2008	2008	DARPA Urban Challenge:97 km urban environment including a variety of roads, intersections, and parking lots. Maneuvers: passing parked or slow-moving vehicles, precedence handling at intersections with multiple stop signs, merging into fast- moving traffic, left turns across oncoming traffic, parking in a parking lot, and the execution of U-turns in situations where a road is completely blocked. Vehicle speeds were generally limited to 30mph, with lower speed limits in many places.	Modified 2006 Volkswagen Passat Wagon	Five laser rangefinders (manufactured by IBEO, Riegl, SICK, and Velodyne), five BOSCH radars and an Applanix GPS-aided inertial navigation system.	The vehicle has an obstacle detection range of up to 120 meters. Velodyne laser scanner: 360° horizontal FOV, 30° vertical FOV and 60 m range. IBEO laser scanner: capable of detecting large vertical obstacles, such as cars and signposts

Vehicles (3)

Demonstration	Year	Route	Vehicle	Sensor Setup	Perception Range
CMU2008	2008	DARPA Urban Challenge (see above)	Modified 2007 Chevrolet Tahoe	GPS/IMU, 10 lasers (manufactured by IBEO, SICK, Continental, and Velodyne), two front-radars, one rear-radar, and 2 cameras	SICK LMS 291-S05/S14 LIDAR: 180/90 deg × 0.9 deg FOV with 1/0.5-deg angular resolution & 80m range.Velodyne HDL-64 LIDAR: 360×26-deg FOV with 0.1-deg angular resolution & 70m range. Continental ISF 172 LIDAR: 12×3.2 deg FOV & 150m range. IBEO Alasca XT LIDAR: 240×3.2 deg FOV & 300 m range. Continental ARS 300 Radar: 60/17 deg×3.2 deg FOV & 60m/200m range. Point Grey Firefly High-dynamic-range camera: 45° FOV
Baidu link: 1, 2				5 cameras (2 front, 2 on either side and 1 rear) and 2 radars (front and rear) along with 3 16-line LiDARs (2 rear and 1 front) and 1 128-line LiDAR	Velodyne HDL-64 LIDAR: 360° FOV

Vehicles (4)

Demonstration	Year	Route	Vehicle	Sensor Setup	Perception Range
UBER		Since September 2016: Self-driving taxis with a safety driver in Pittsburgh Since December 2016: Self-driving Volvo XC90s in San Francisco	Volvo XC90	Multiple LiDAR sensors (including one, top-mounted Velodyne LIDAR) Multiple cameras that provide high resolution, near-, medium-, and long-range imagery. There are cameras mounted in the sensor pod on top of the vehicle and around the vehicle. Some of these cameras have a wide field of view and some have a narrow field of view. Forward-facing radars are mounted below the headlamps, side-facing radars are mounted in the front and rear corners of the vehicle, and rear-facing radars are mounted near the ends of the bumper beam. GPS	LiDAR: 360° FOV, range over 100 m All cameras together enable a 360° FOV A system of cameras provides imagery to support near-range sensing of people and objects within 5m from vehicle.

Vehicles (5)

Demonstration	Year	Route	Vehicle	Sensor Setup	Perception Range
Waymo2017	2017	Over the last eight years, Waymo has tested its vehicles in four U.S. states and self-driven in more than 20 cities—from sunny Phoenix, AZ to rainy Kirkland, WA—accumulating more than 3.5 million autonomous miles in the process. Waymo has set up a private, 91-acre, closed-course testing facility in California specially designed and built for our own unique testing needs. This private facility, nicknamed "Castle," is set up like a mock city, including everything from high- speed roads to suburban driveways to a railroad crossing. Waymo has developed more than 20,000 simulation scenarios at Castle. Each recreates a driving situation for practicing. Waymo's system is designed so each vehicle does not operate outside of its approved operational design domain (It does not travel outside of a "geo-fenced" area, which has been mapped in detail)	Different vehicles	Typically short-range, mid-range and long-range LiDAR, a camera system and a radar system. Waymo's vision system is comprised of several sets of high-resolution cameras, designed to work well at long range, in daylight and low-light conditions. Waymo vehicles also have a number of additional sensors, including an audio detection system that can hear police and emergency vehicle sirens up to hundreds of feet away, and GPS.	Long-range LiDAR: 360° FOV, range: approx. 320 m 360° camera view 360° radar view

Vehicles (6)

Demonstration	Year	Route	Vehicle	Sensor Setup	Perception Range
Tesla			Tesla	A forward-facing radar Twelve ultrasonic sensors Eight surround cameras	Radar: range up to 160 m. Ultrasonics: range up to 8 m. All eight cameras together: 360° FOV, range up to 250 m . Narrow front camera: range up to 250 m. Main front camera: range up to 150 m. Wide front camera: 120° fisheye lens, range up to 60 m. Two forward looking side cameras: 90° FOV, range up to 80 m. Two rearward looking side cameras: range up to 100 m. Rear camera: range up to 50 m.
GM + Cruise		"Our driverless cars are on the road in California, Arizona, and Michigan navigating some of the most challenging and unpredictable driving environments." "In our controlled deployment, our self-driving vehicles will drive only in known geo-fenced boundaries, and only on roads for which we have developed high-definition map data. They will also drive only under known operational conditions and constraints that apply to the entire fleet."	GM	5 LiDARs16 cameras 21 radars	All sensors together scan both long and short range with views 360 degrees around the vehicle. Field of view overlaps enable 360-degree vision even if a sensor fails.

Vehicles (7)

Demonstration	Year	Route	Vehicle	Sensor Setup	Perception Range
Zoox		Zoox has a California DMV permit to test these vehicles on public roads. Zoox's vehicle testing team performs daily drives around the San Francisco Bay Area. They test in various weather and road conditions on private roads, test tracks, and public roads. Today (December 2018), Zoox's system can drive autonomously in a range of conditions, from suburbs, to freeways at higher speeds, and dense urban environments.	Toyota Highlanders and Prius C's	Symmetric sensor configuration using multiple cameras, lidar, radar, and proprietary sensors.	All sensors together provide a 360° view
Nvidia Link: 1, 2	2016 / 2018	2016: "For a typical drive in Monmouth County NJ from our office in Holmdel to Atlantic Highlands, we are autonomous approximately 98% of the time. We also drove 10 miles on the Garden State Parkway (a multi-lane divided highway with on and off ramps) with zero intercepts." 2018: ?	2016 Lincoln MKZ, 2013 Ford Focus	Three front-facing cameras mounted behind the windshield. For testing only one front-facing camera was used. 2018: Multiple cameras, radar and LiDAR sensors	2016: ? 2018: All sensors together provide a 360° view

Vehicles (8)

Demonstration	Year	Route	Vehicle	Sensor Setup	Perception Range
Aptiv	2015 - 2018	First coast-to-coast autonomous drive in Apr 2015: "Our team and technology helped complete the longest automated vehicle drive ever – traveling nearly 3,400 miles from San Francisco to New York City, with 99 percent of the drive in fully automated mode. The vehicle successfully navigated through complex driving situations collecting data essential to advancing the emerging active safety technology sector." "At CES 2018 in Las Vegas, our self-driving cars performed more than 400 point-to-point rides, 99% of the miles driven in fully autonomous mode, with a 4.997 average ride rating." "In May 2018, our team announced the deployment of 30 self-driving cars, equipped with Aptiv's autonomous driving platform. These vehicles are offered to the public of Las Vegas via the Lyft app. We are proud of a significant milestone: 5,000 self-driving public rides—powered by the Aptiv autonomous driving platform."	BMW 5	4 short-range LiDARs: • One in the front • One in the rear • One on each side of the car below the side mirror 5 long-range LiDARs: • Two in the front • One in the rear • One on each side of the car 6 electronically scanning radars (ESR): • Three in the front • One in the rear • One on each side of the car 4 short-range radars (SRR): • Two in the front • Two in the rear 1 trifocal camera behind the windscreen 1 traffic light camera behind the windscreen 2 GPS antennas 1 Dedicated Short Range Communications antenna (DSRC)	360° radar technology

Vehicles (9)

Demonstration	Year	Route	Vehicle	Sensor Setup	Perception Range
[Lyft]()				4 near-angle cameras: • Two in the front • Two directed to the left and right of the vehicle respectively 4 wide-angle cameras: • One forward-facing • One backward-facing • Two directed to the left and right of the vehicle respectively. One front-facing long-range radar. One front-facing short-range radar. One top-mounted LiDAR sensor	Near-angle cameras: approx. 60° FOV each. Together, they aquire a combined view of approx. 240° FOV. Wide-angle cameras: approx. 150° FOV each. Long-range radar: approx. 35° FOV Short-range radar: approx. 80° FOV LiDAR: 360° FOV
[Ford]()		There are fleets of vehicles testing on public roads in Miami, Fla., Pittsburgh, Pa. and Dearborn, Mich.		Ford Fusion Hybrid sedan 360-degree view camera Rear-facing camera Top-mounted LiDAR Four radar sensors: • One front-facing radar • One rear-facing radar One radar on each side of the vehicle	LiDAR: 360° FOV, more than 250 m range. Cameras: 360° FOV

Vehicles (10)

Demonstration	Year	Route	Vehicle	Sensor Setup	Perception Range
PROUD-Car Test 2013 (BRAiVE, VisLab)	2013	Route from the Campus of the University of Parma to Piazza della Pace: It included two-way rural roads, two freeways with junctions, and plenty of urban areas such as pedestrian crossings, tunnels, artificial bumps, tight roundabouts, and traffic lights.		• Altogether: 10 cameras, 5 laser scanners, 1 GPS+IMU, 1 e-Stop system. Cameras:4 front-facing cameras (two color and two monochrome DragonFly2 cameras) are placed behind the windshield. • 2 cameras at the back in the boot top • 1 camera (FireFlyMV) in each side mirror • 1 lateral camera on each side of the hood Laser scanners: • Two UTM-30LX are mounted on the sides of the front bumper • One UTM-30LX is placed on the center rear bumper • One IBEO Lux is placed in the front bumper's center One Hella IDIS laser is mounted over the Lux in forward looking direction	HxV aperture of the cameras: • Stereo front (short baseline): 73.76° x 58.86° • Stereo front (long baseline): 73.76° x 58.86° • Stereo back: 130.81° x 117.32° • Side mirror: 96.93° x 71.54° • Lateral: 100.43° x 84.15° Laser scanners: • Hokuyo UTM-30LX: 270° FOV, 0.25° resolution, 0.1 –30 m • IBEO Lux: 85° FOV, 0.125 –1° resolution, 0.3 –80 m range • Hella IDIS: 16° FOV, 1° resolution, 0.7 – 110 m range

Questions need to addressed

- What sensing modalities should be fused, and how to represent and process them in an appropriate way?
- What fusion operations should be utilized?
- At which stage of feature representation in a neural network should the sensing modalities be combined?

Autonomous Driving research is facing some key challenges as below:

Topics		Challenges
Multi-modal data preparation	Data diversity	• Relative small size of training dataset. • Limited driving scenarios and conditions, limited sensor variety, object class imbalance.
	Data quality	• Labeling errors. • Spatial and temporal misalignment of different sensors.
Fusion methodology	What to fuse	• Too few sensing modalities are fused. • Lack of studies for different feature representations
	How to fuse	• Lack of uncertainty quantification for each sensor channel. • Too simple fusion operations.
	When to fuse	• Fusion architecture is often designed by empirical results. No guideline for optimal fusion architecture design. • Lack of study for accuracy/ speed or memory/ robustness trade-offs.
Others	Evaluation	Current metrics focus on comparing networks' accuracy
	More network architecture	• Current networks lack temporal cues and cannot guarantee prediction consistency over time. • They are designed mainly for modular autonomous driving.

Study & Opinion

Most deep multi-modal perception methods are based on supervised learning. Therefore, multi-modal datasets with labeled ground-truth are required for training such deep neural networks.

Based on the review, currently I do a research and work on object detection and semantic segmentation, considering the input data using deep learning.