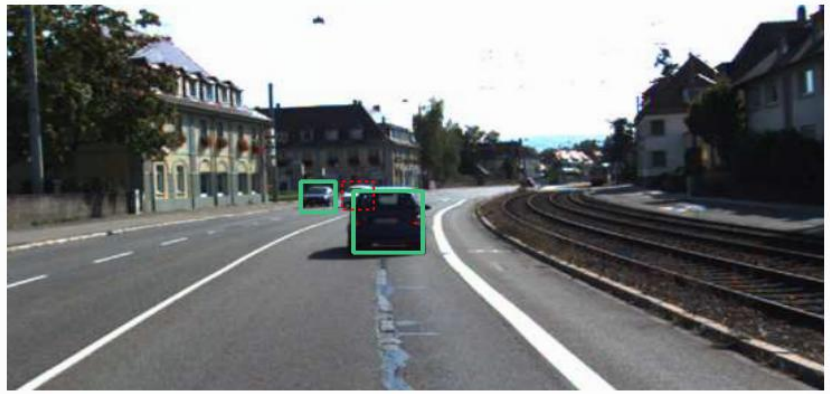
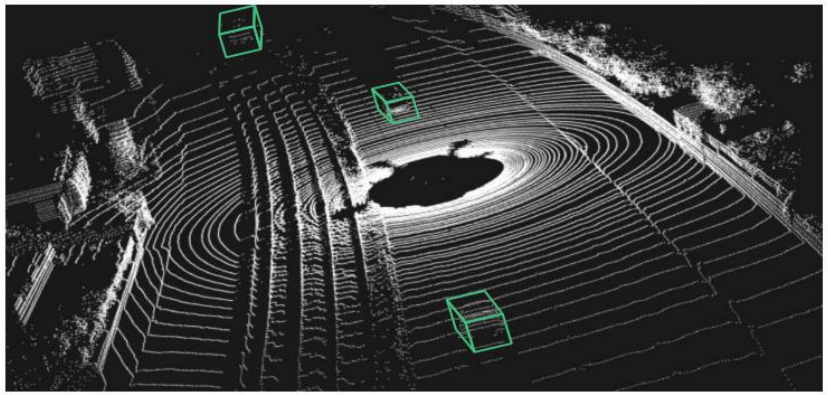
Multimodal Fusion perception in self-driving

Why Multimodal?

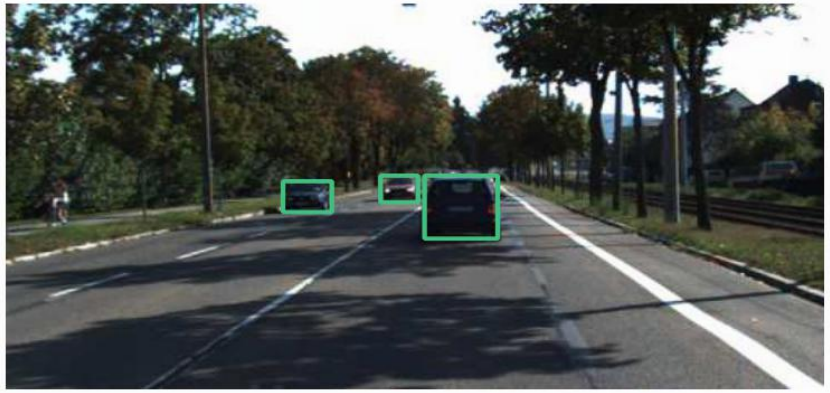
Typical problem using single modal sensors



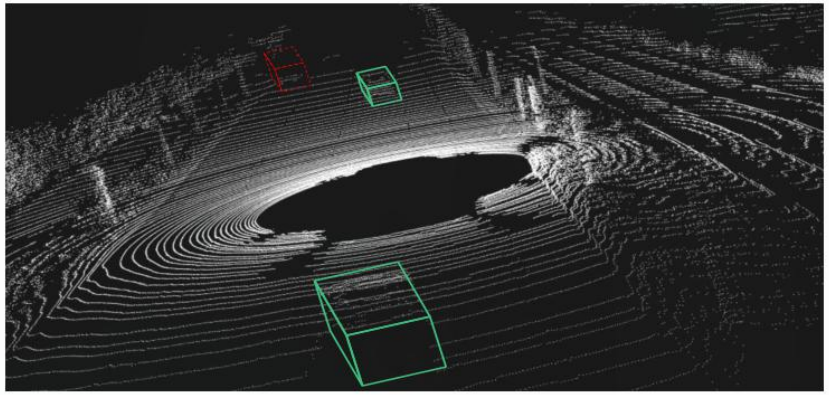
(a) Camera only detector #1



(b) LiDAR only detector #1



(c) Camera only detector #2

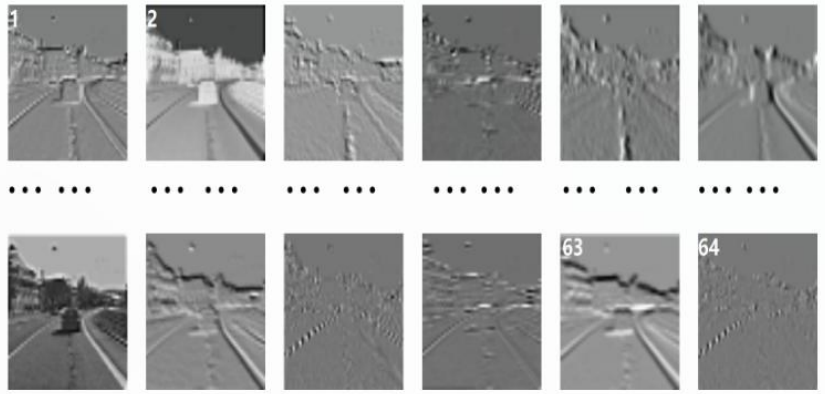


(d) LiDAR only detector #2

Representation

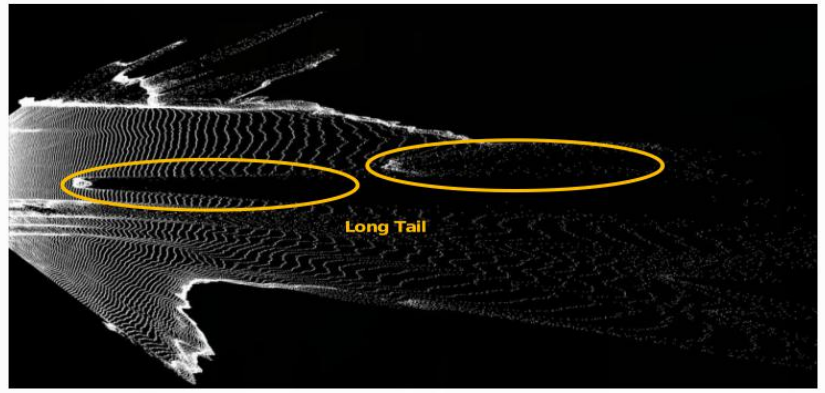
- Image
  - feature map
  - mask
  - pseudo-LiDAR

(a) raw image



(b) feature maps (64 channels)

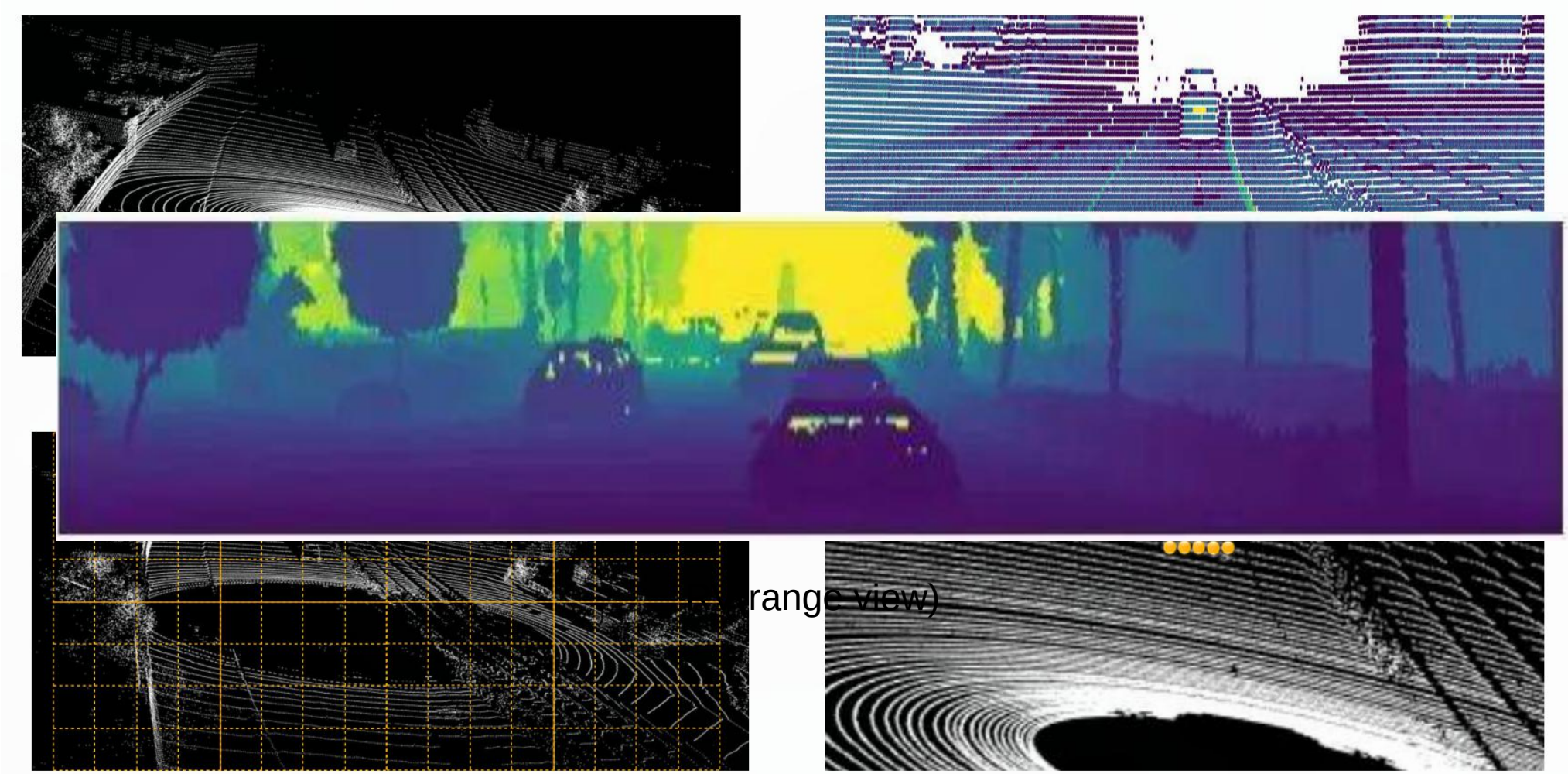
(c) mask



(d) pseudo-LiDAR point cloud (BEV)

Representation

- LiDAR
  - points
  - voxels
  - views
    - BEV
    - RV



range view)

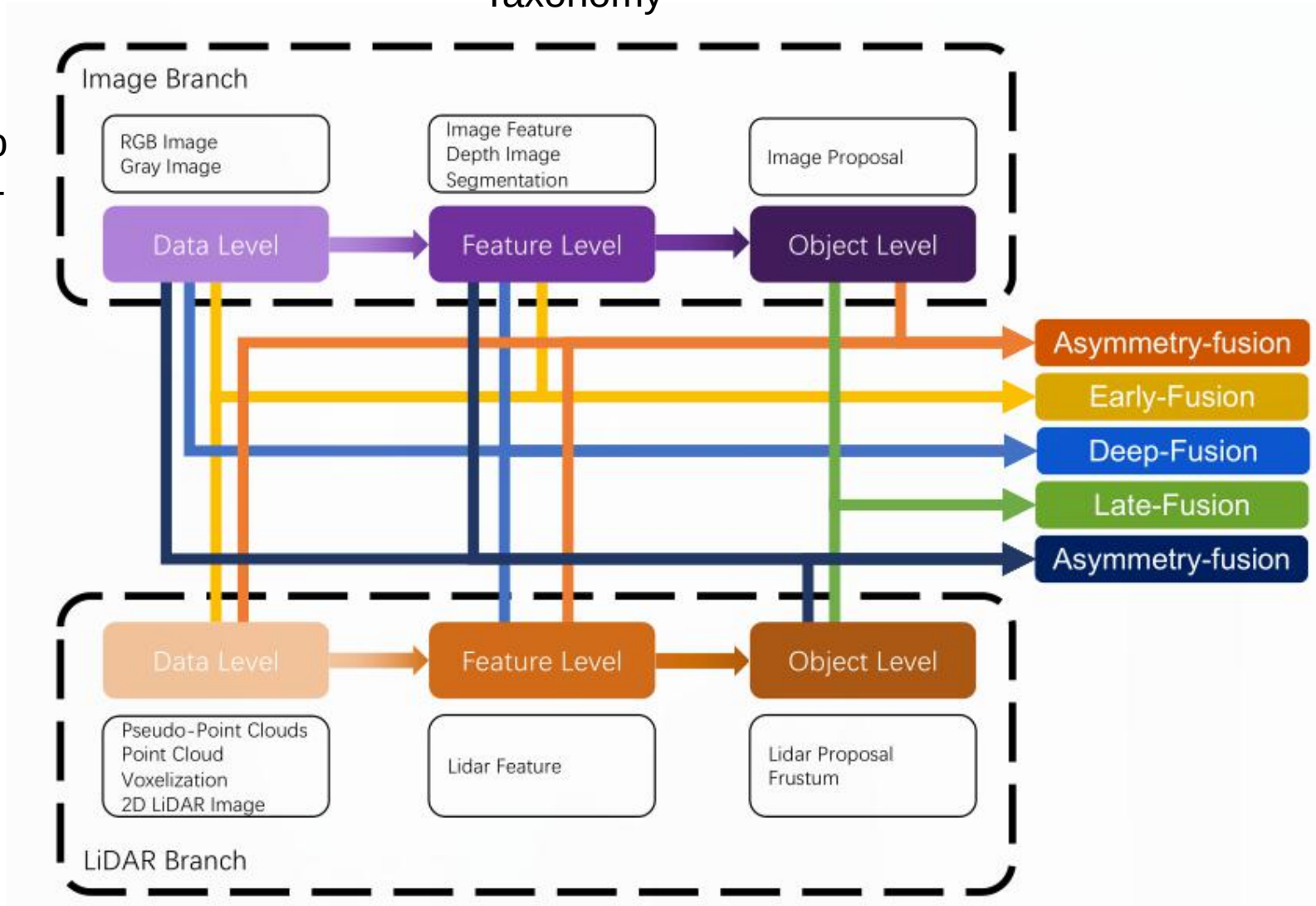
(c) point cloud voxels

(d) point cloud points

Methods

The fusion methos can be divided into four categories, as early-fusion, deep-fusion, late-fusion, and asymmetry-fusion.

Taxonomy



- Early Fusion
- Deep Fusion
- Late Fusion

Early Fusion

Early fusion fuses LiDAR data at the data level and camera data at data-level or feature-level. One example of early-fusion can be the model as follows:

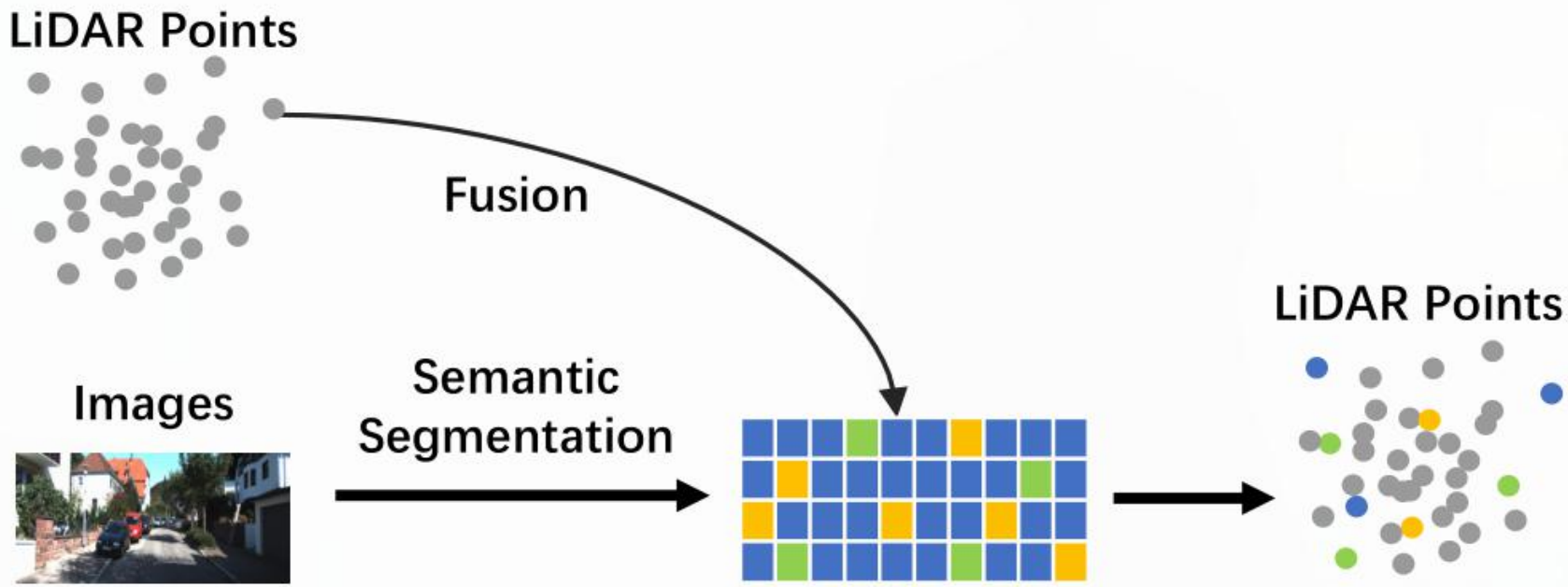
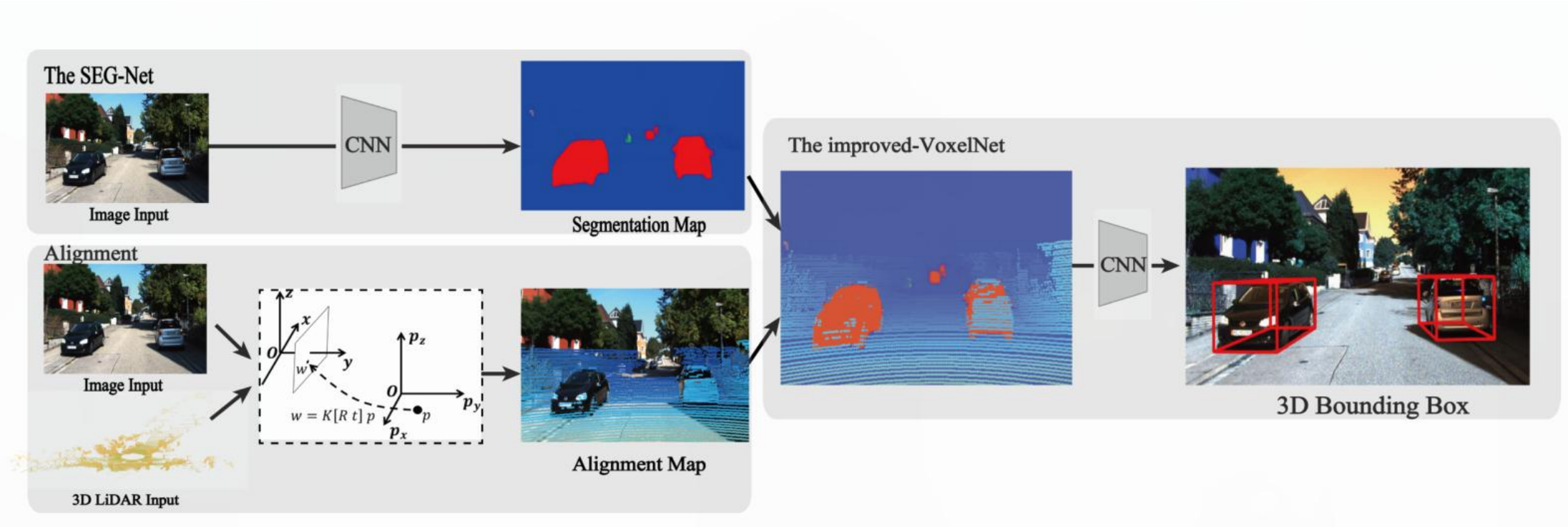


Figure 4. An Example of Early-Fusion

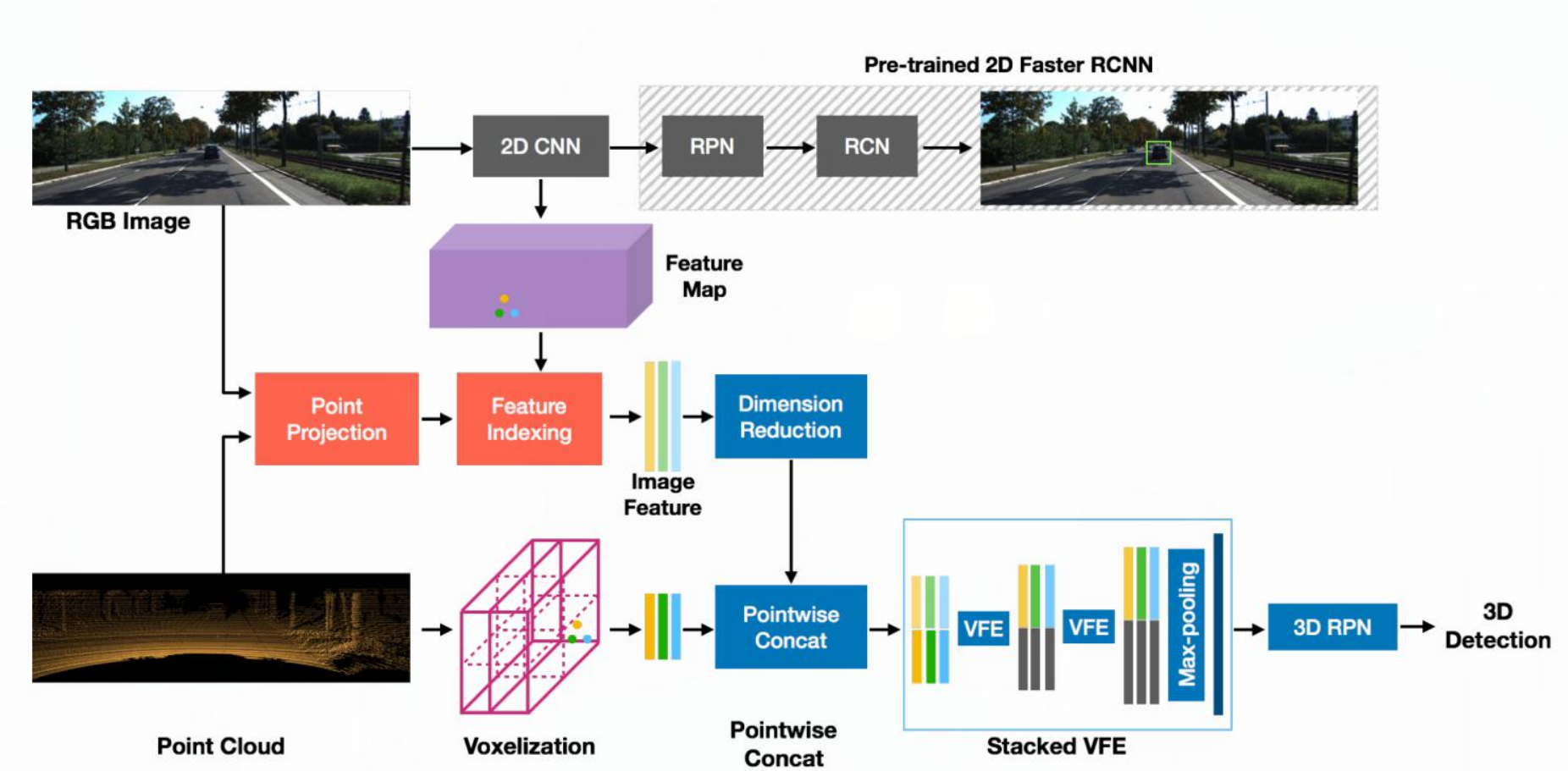
Early Fusion

* Seg-voxelnet(2019): Seg-voxelnet for 3d vehicle detection from rgb and lidar data.



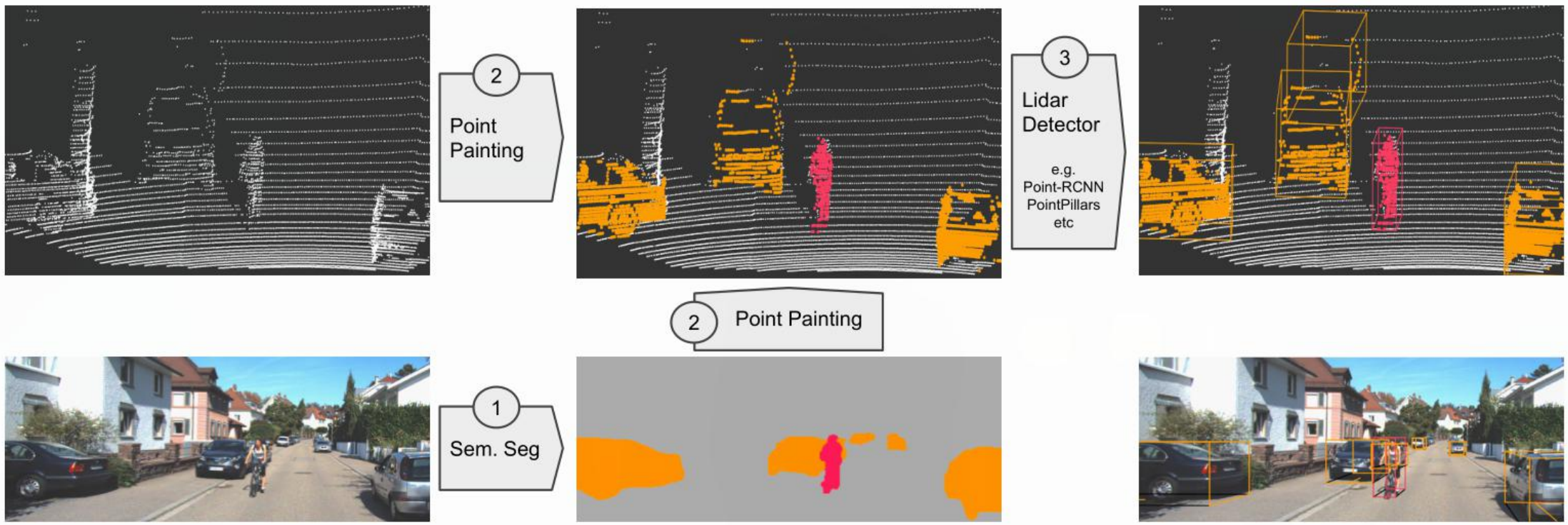
Early Fusion

* Mvx-net(2019): Multimodal voxelnet for 3d object detection.



Early Fusion

* Pointpainting(2020): Sequential fusion for 3d object detection.



- Early Fusion
- Deep Fusion
- Late Fusion

Deep Fusion

Deep-fusion methods fuse cross-modal data at the feature level for the LiDAR branch but data-level and feature-level for the image branch.

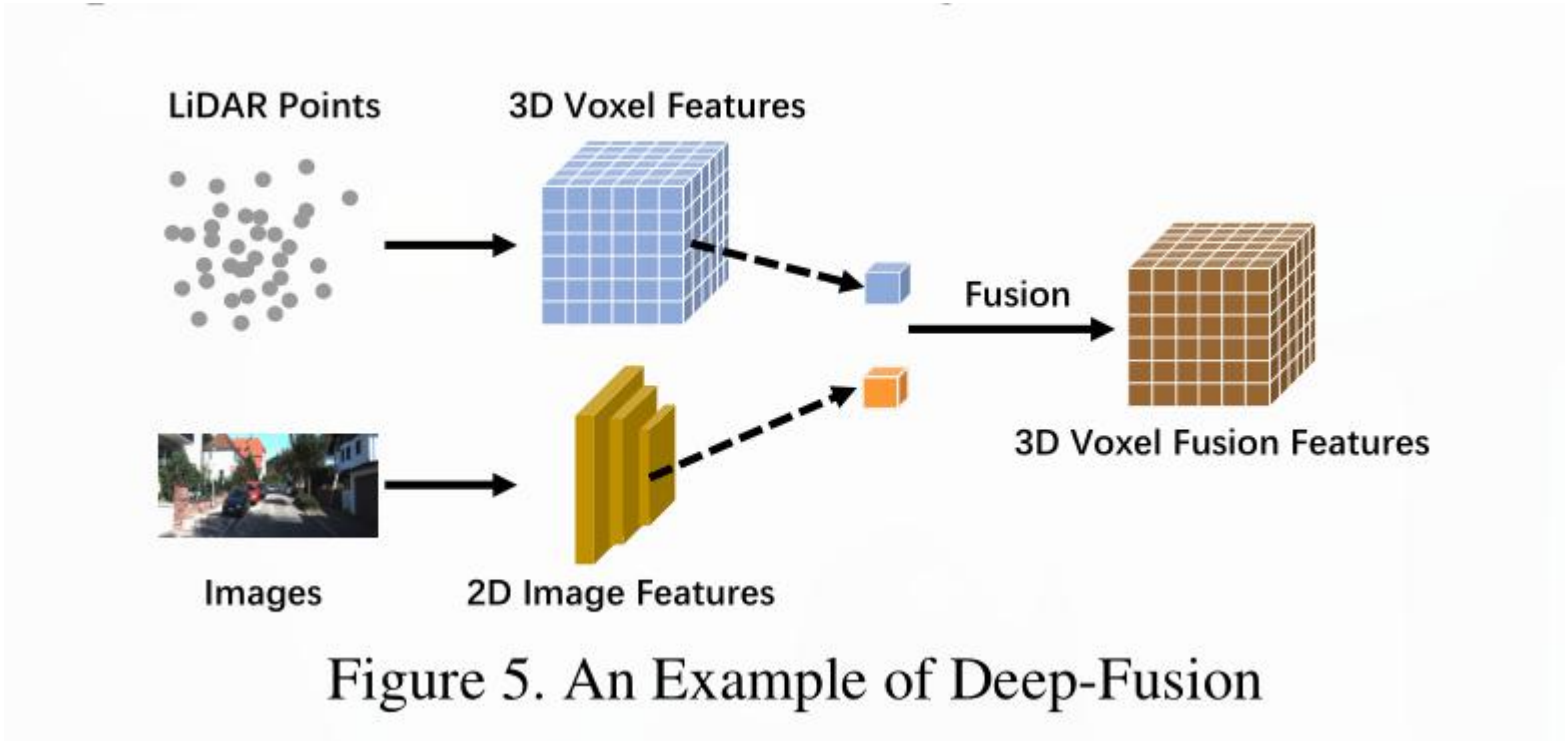
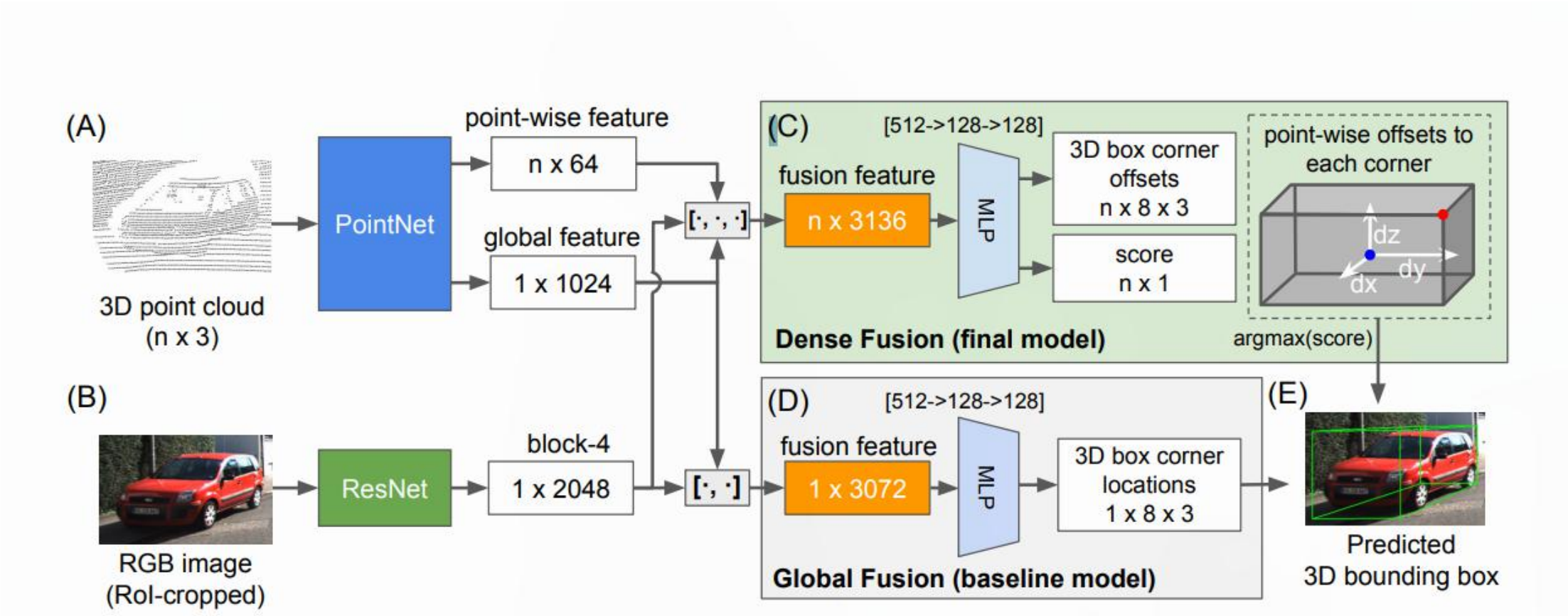


Figure 5. An Example of Deep-Fusion

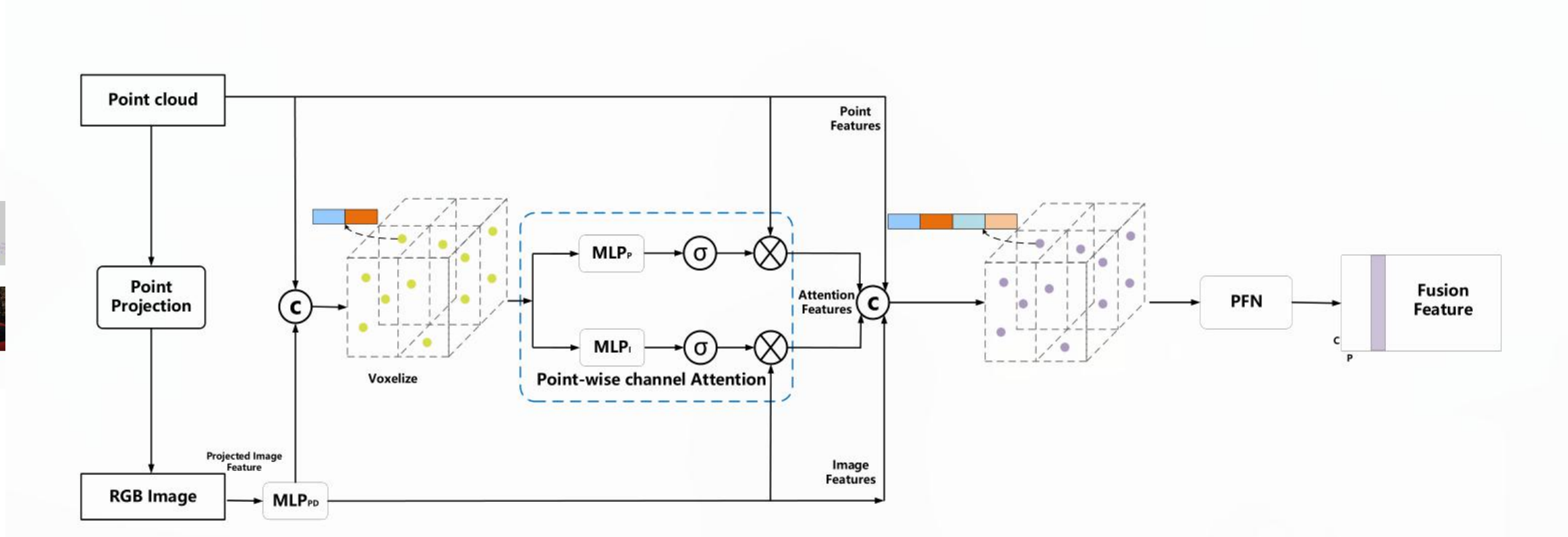
Deep Fusion

* PointFusion(CVPR 2018): Deep sensor fusion for 3d bounding box estimation.



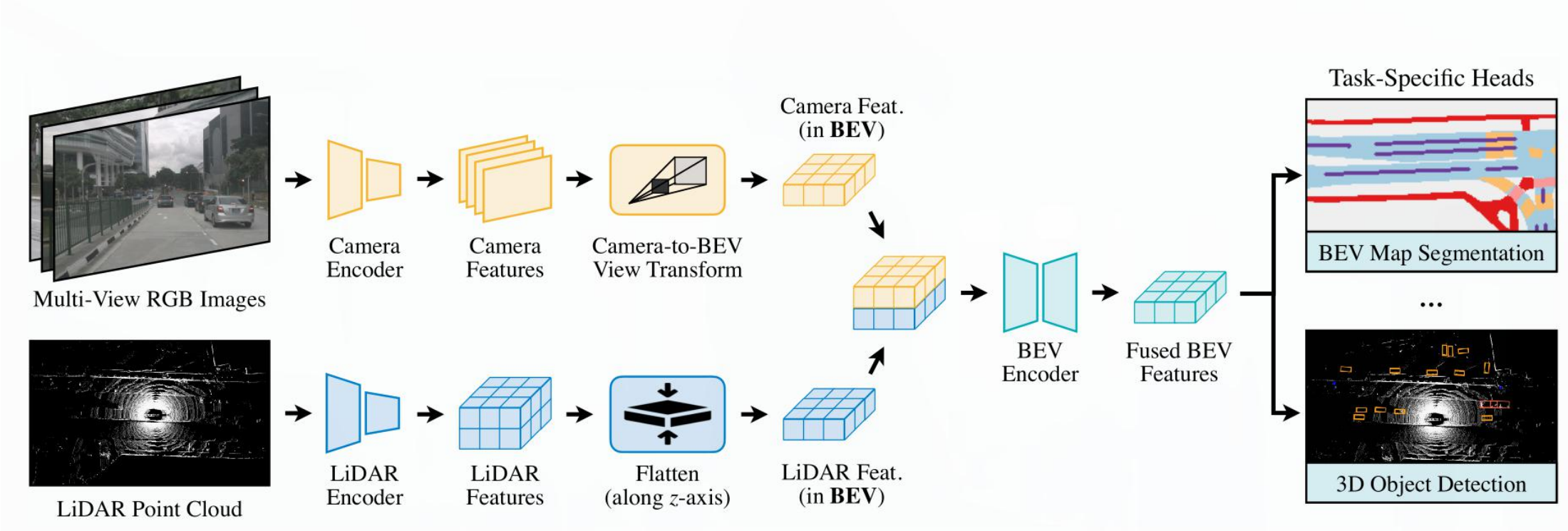
Deep Fusion

* Maff-net(2020): Filter false positive for 3d vehicle detection with multi-modal adaptive feature fusion.



Deep Fusion

* BEVFusion(2022):Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation



Deep Fusion

- **Roifusion**(2021): 3d object detection from lidar and vision.
- (CVPR 2019): Noise-aware unsupervised deep lidar-stereo fusion.
- (CVPR 2017): Fast boosting based detection using scale invariant multimodal multi-resolution filtered features.
- **Semanticvoxels**(2020): Sequential fusion for 3d pedestrian detection using lidar point cloud and semantic segmentation.
- (IV 2018): Robust camera lidar sensor fusion via deep gated information fusion network.
- (ECCV 2018): Deep continuous fusion for multi-sensor 3d object detection.
- **R-agno-rpn**(2020): A lidar-camera region deep network for resolution-agnostic detection.
- (2020): Multi-view adaptive fusion network for 3d object detection.
- (2017): Fusing bird view lidar point cloud and front view camera image for deep object detection.

- Early Fusion
- Deep Fusion
- Late Fusion

Late Fusion

Late-fusion, also known as object-level fusion, denotes the methods that fuse the result of pipelines in each modality.

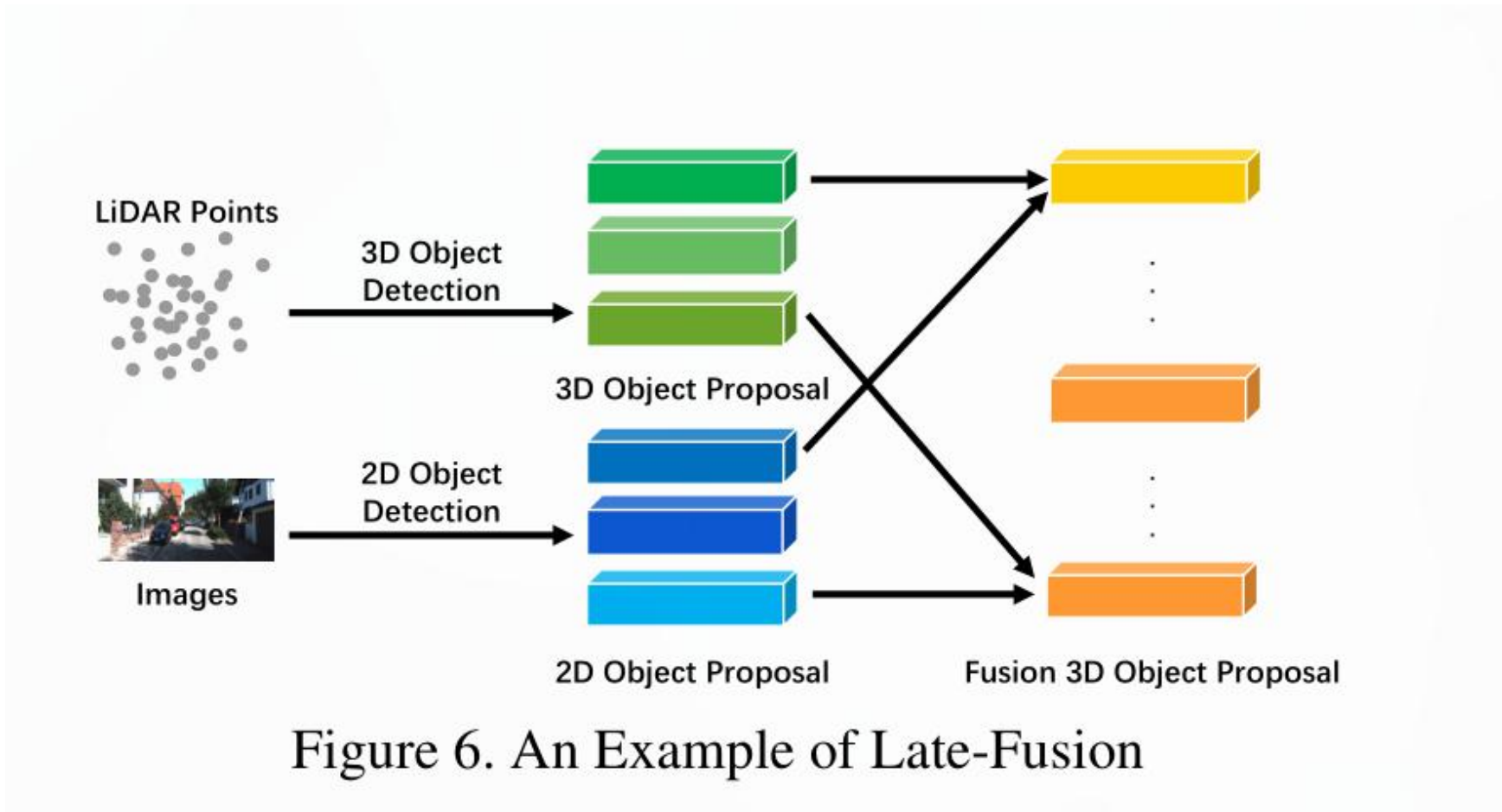


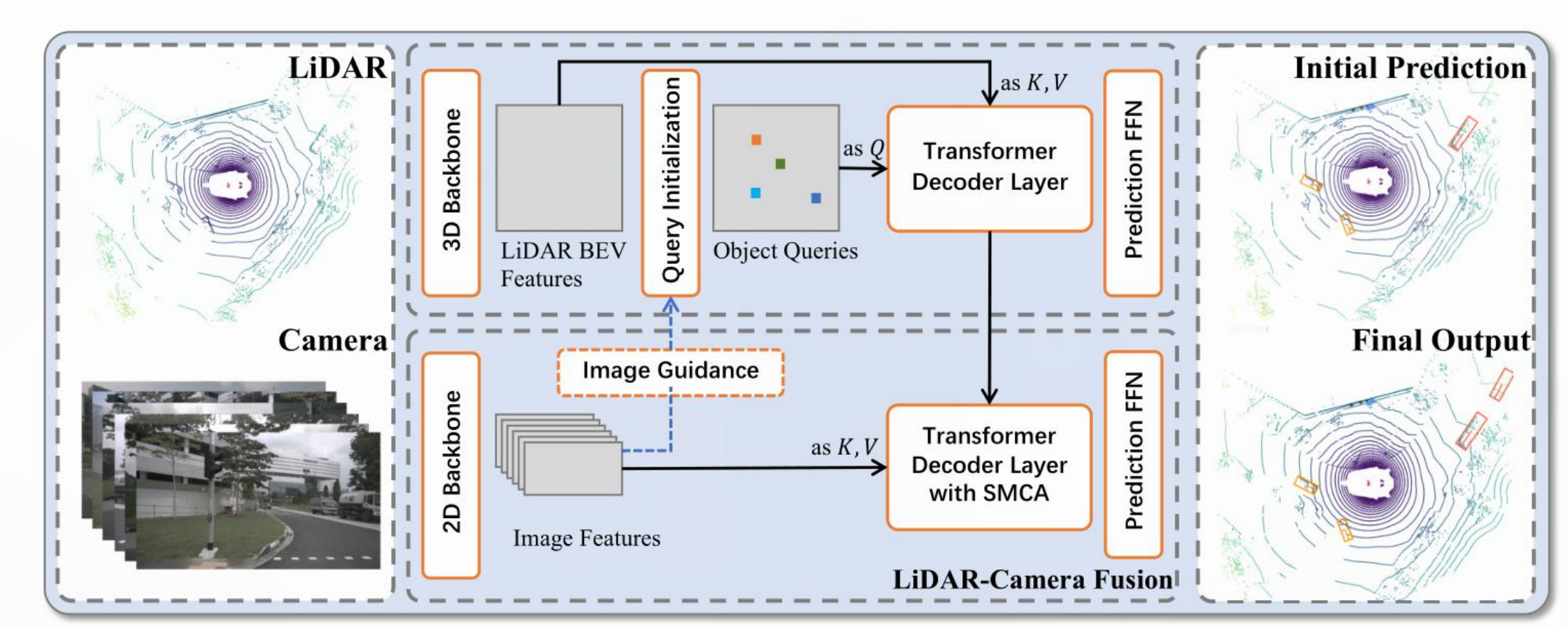
Figure 6. An Example of Late-Fusion

Opportunities in Multi-Modal Fusion

- More Advanced Fusion Methodology
  - Misalignment and Information Loss
  - More Reasonable Fusion Operations

Deep Fusion

* TransFusion(CVPR 2022): Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers



Trend

- Deep fusion
- BEV representation