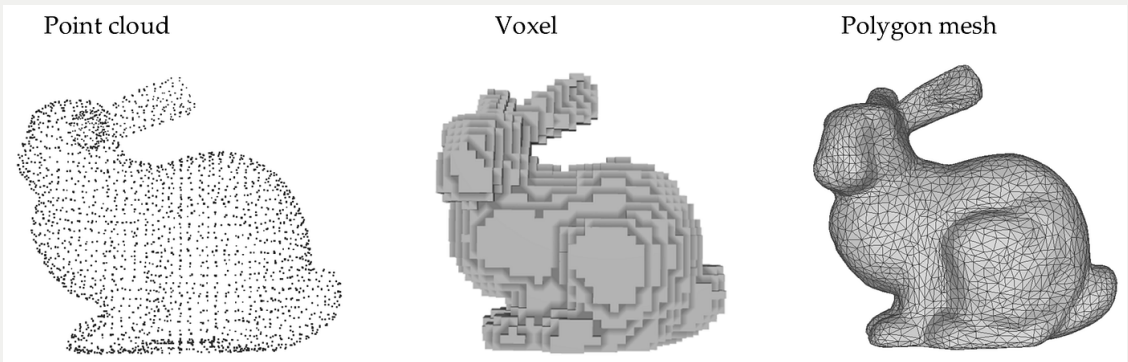
[RW] Voxel-based Scene Representation

(RW for Tri-Perspective View for Vision-Based 3D Semantic Occupancy Prediction)

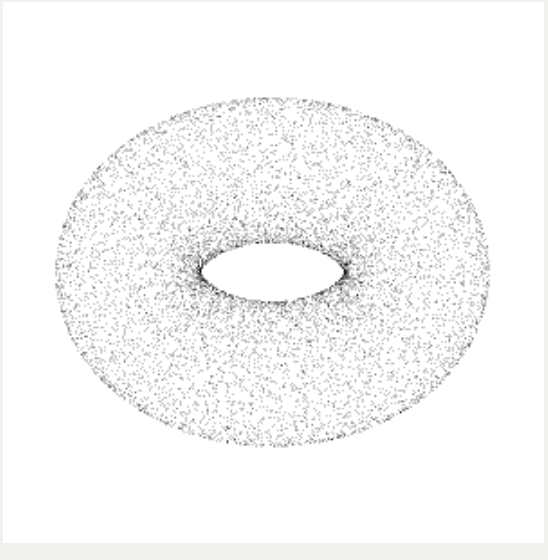
이미지처리팀 류채은

Prerequisites

📝 **Voxel이란?**
= Volume과 pixel을 조합한 혼성어로 3차원 공간에서 정규 격자 단위 값



📝 **Point Cloud의 대표적 특징**



빈 공간이 많아 **sparse**하다는 특징이 있다 (차지하는 공간에 비해 적은 정보량…)

Discretizing the 3D space into voxels and assign a vector to represent each voxel [52,24]

[52] VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection (2017, CVPR)
link: https://arxiv.org/pdf/1711.06396.pdf

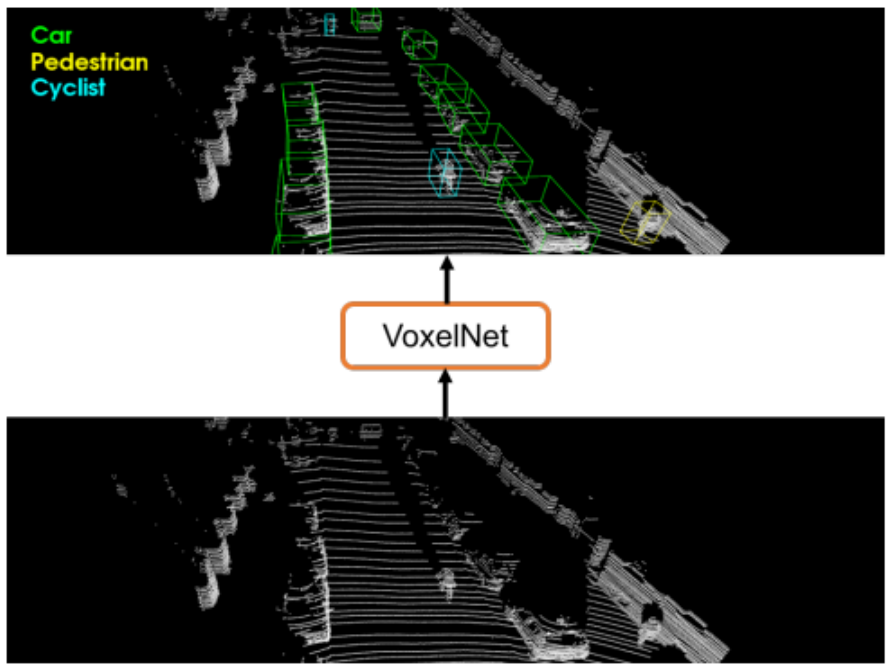
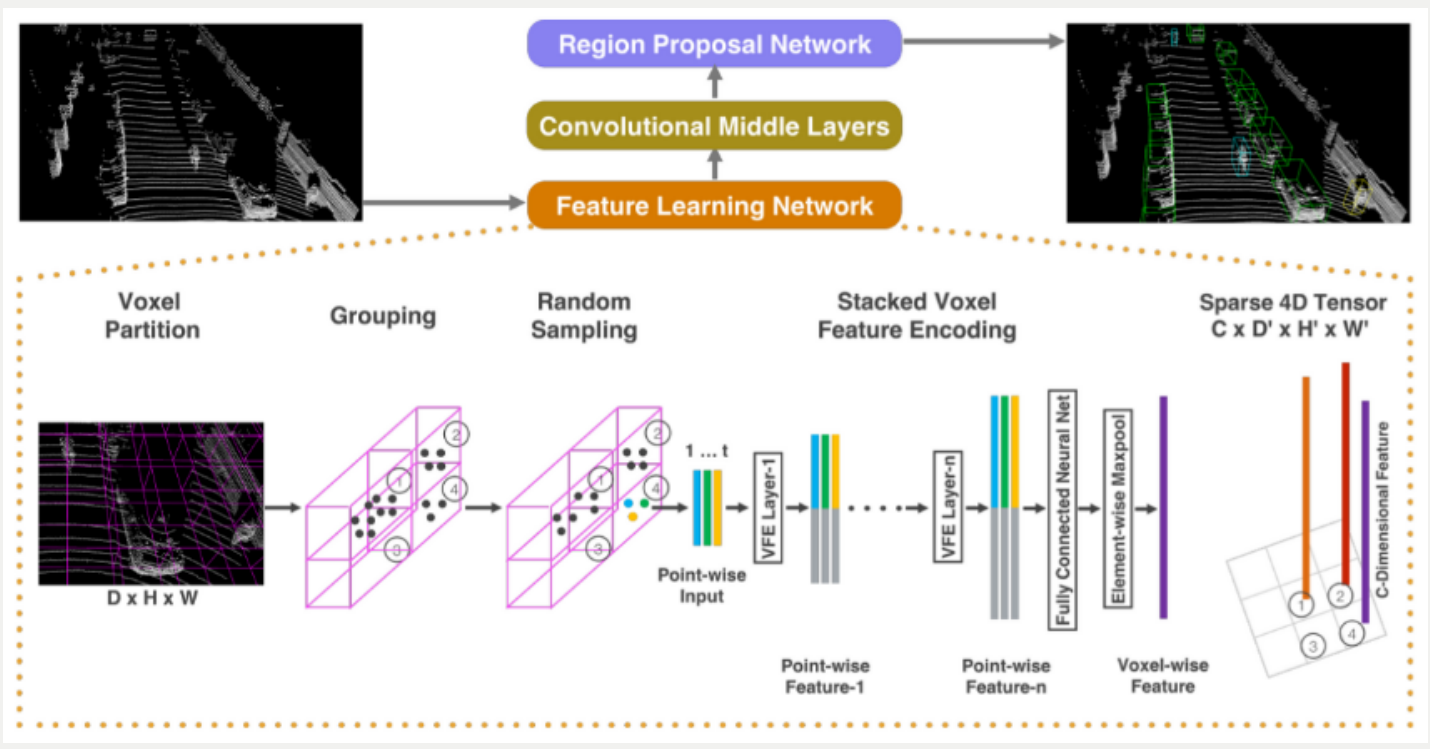


Figure 1. VoxelNet directly operates on the raw point cloud (no need for feature engineering) and produces the 3D detection results using a single end-to-end trainable network.

[**TL;DR**]
Proposes VoxelNet, a generic 3D detection network that unifies feature extraction and bounding box prediction into a single stage, end-to-end trainable deep network. (Point cloud를 바로 활용해서 object detection까지 진행하는 end-to-end를 고안했다!)

[Methodology]



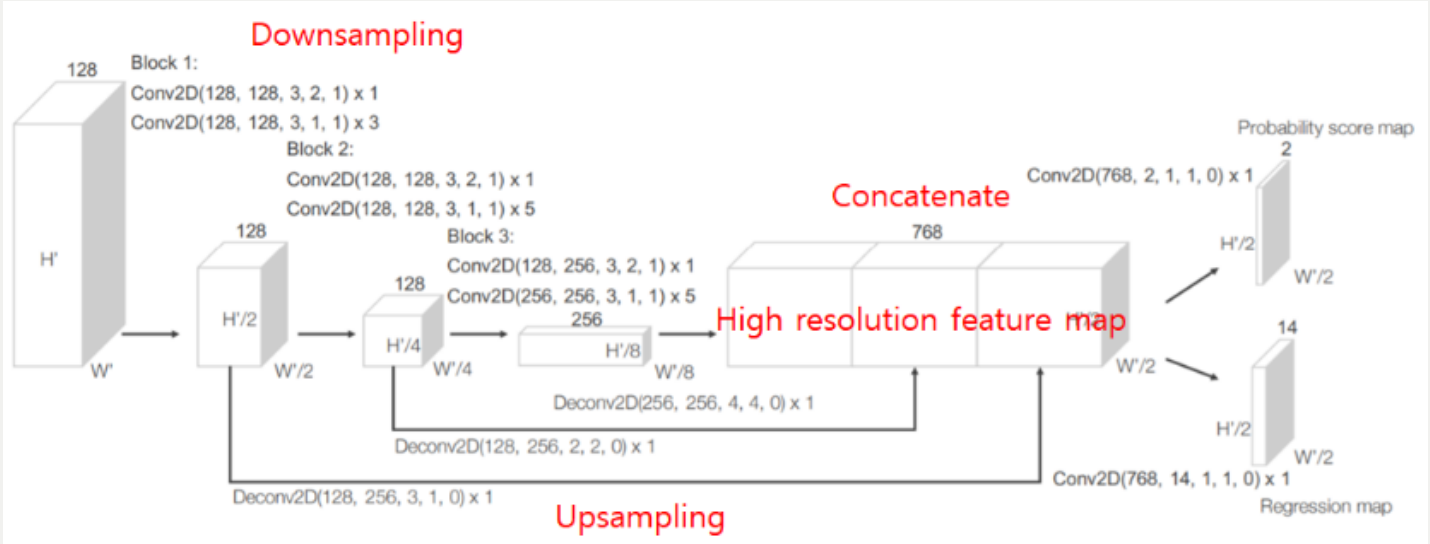
Step 1. Feature Learning Network

- Point cloud를 voxel 단위로 매핑하기 위해 일정하게 나눈다 **[voxel partition]**
- 하나의 voxel 범위 내에 있는 point를 grouping한다 **[grouping]**
- 막대한 계산 비용과 point density variance imbalance (point cloud 특성 상 sparse함이 심함)를 해결하기 위해 T개 이상의 point를 가진 voxel에서 정확히 T개의 point만 random하게 sampling 한다 **[Random sampling]**

- 연속적인 Voxel Feature Encoding Layer을 쌓아서 point-wise feature을 얻고 이후 FCN과 Pooling 연산을 통해 voxel-wise feature을 최종적으로 얻게 된다. **[Stacked Voxel Feature Encoding]**
- non-empty voxel만 처리함으로써 voxel features에 관한 리스트(non-empty voxel의 위치 정보를 담음)를 얻는다. **[Sparse Tensor Representation]**

  Step 2. Convolutional Middle Layers
- `3D Conv - BatchNorm - ReLU`를 연속적으로 통과시켜 receptive field를 점차적으로 키워서 context 정보를 배우도록 함

  Step 3. Region Proposal Network



Region Proposal Network

- downsampling과 upsampling을 거친 후 probability score map과 regression map을 도출한다.

The ability to describe fine-grained 3D structures makes voxel-based representation favorable for 3D semantic occupancy prediction tasks including lidar segmentation [12,30,43,47,48,55]

[12] AF)2-S3Net: Attentive Feature Fusion with Adaptive Feature Selection for Sparse Semantic Segmentation Network

link: https://arxiv.org/pdf/2102.04530.pdf

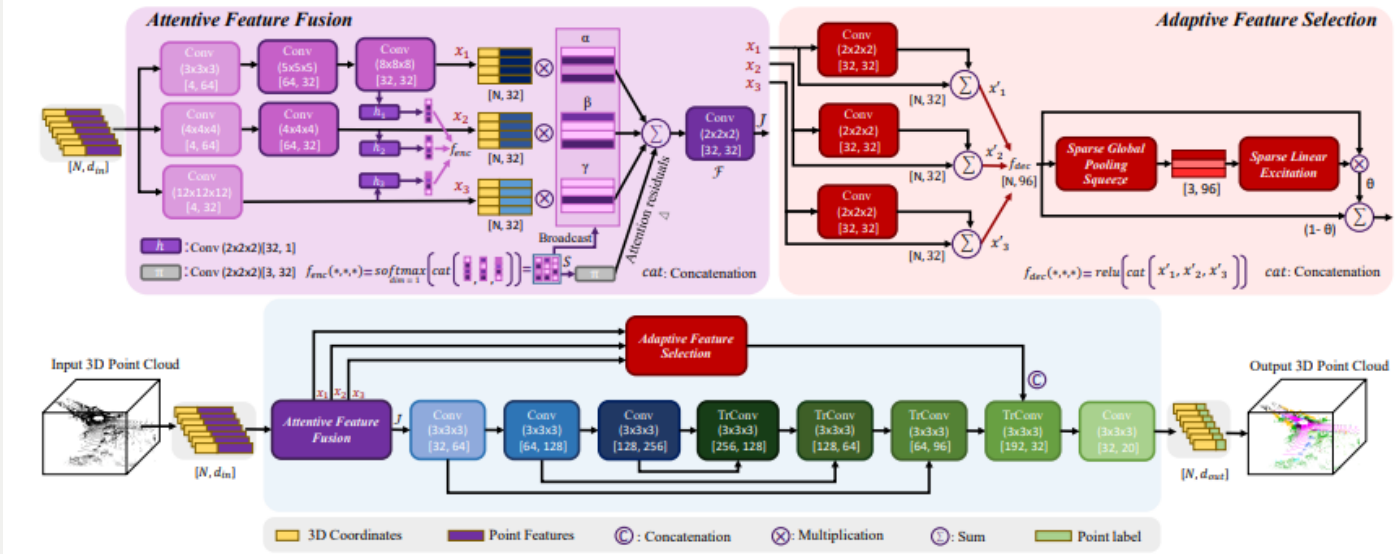


Figure 2: Overview of $(AF)^2$-S3Net. The top left block is Attentive Feature Fusion Module (AF2M) that aggregates Local and global context using a weighted combination of mutually exclusive learnable masks, α, β, and γ. The top right block illustrates how Adaptive Feature Selection Module (AFSM) uses shared parameters to learn inter relationship between channels across multi-scale feature maps from AF2M. (best viewed on display)

TL;DR

Better usage of sparse 3D convolution by the two essential modules: Attentive Feature Fusion Module (= AF2M) and Adaptive Feature Selection Module (=AFSM)

[AF2M (보라색 부분)]

Attention feature fusion으로 구성된 multi-branch structure의 모듈이다. global context와 local detail을 동시에 배우기 위해 활용한다.

[AFSM (빨간색 부분)]

Squeeze excitation module로 **AF2M이 multi-branch structure을 가짐으로써 생기는 중복(redundancy)를 filter하기 위한 모듈**이다.

[전체 파이프라인 (하단 긴 그림)]

3D point cloud를 sparse tensor로 변환 → AF2M에 input으로 활용과 동시 AFSM에 입력 → convolution blocks와 transpose convolution blocks를 통과 → 마지막 conv 레이어가 sub-manifold features를 semantic class probabilities로 변환 → semantic class로 변환

3D scene completion [5,10,24,41,46]

[5] MonoScene: Monocular 3D Semantic Scene Completion

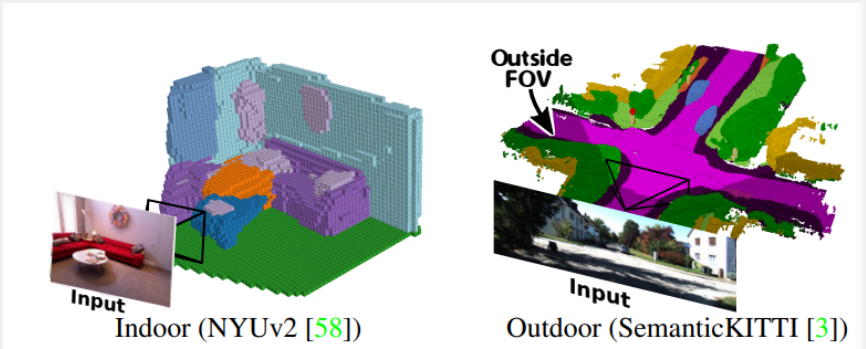
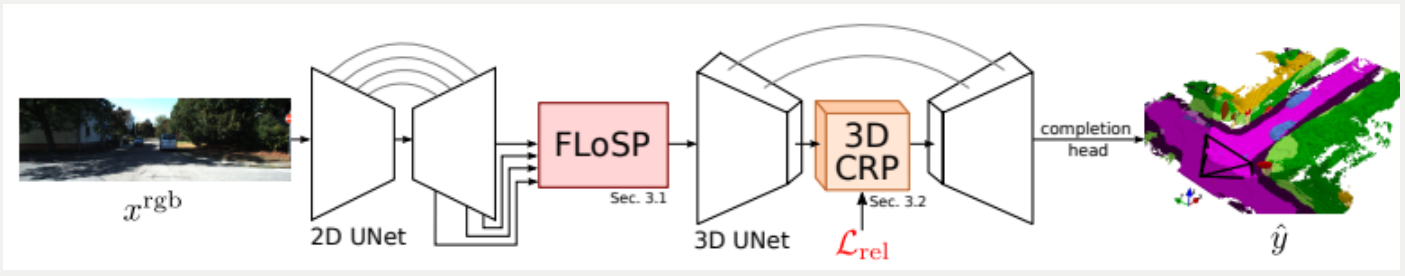


Figure 1. **RGB Semantic Scene Completion with MonoScene.** Our framework infers dense semantic scenes, hallucinating scenery outside the field of view of the image (dark voxels, right).

we solve the complex problem of 2D to 3D scene reconstruction while jointly inferring its semantics

TL;DR

Single 2D image를 활용하여 3D reconstruction을 수행한다. (point cloud 활용하는 방법론 아님!)

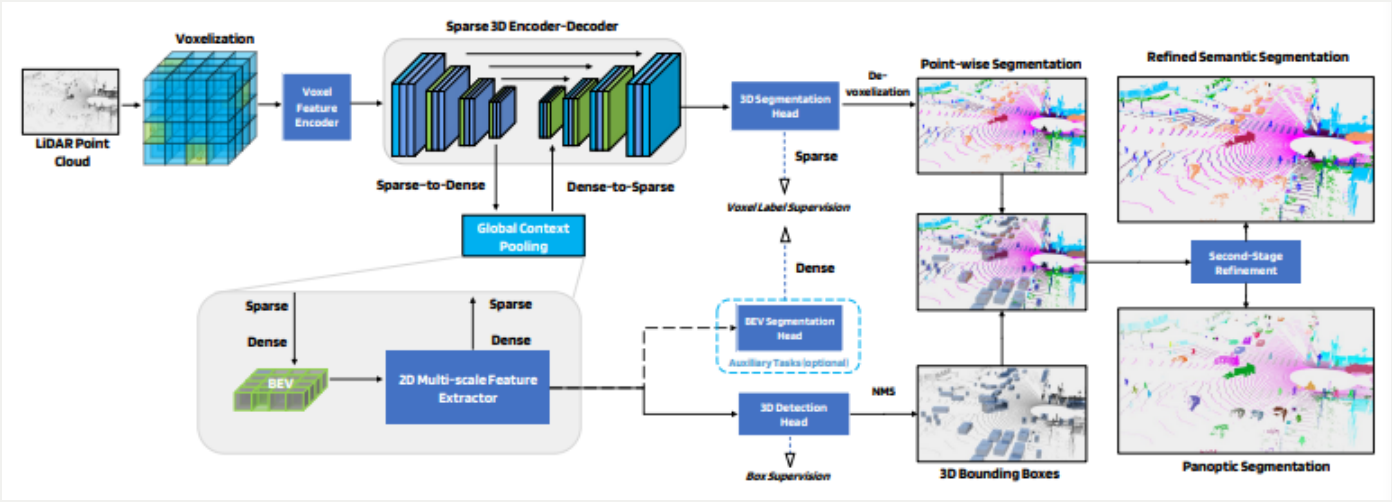


We infer 3D Semantic Scene Completion from a single RGB image, leveraging 2D and 3D UNets, bridged by our Features Line of Sight Projection and a 3D Context Relation Prior (3D CRP) to enhance spatio-semantic awareness.

2D Network과 3D Network를 모두 활용함으로써 2D Features를 효과적으로 projection하고 voxel 그룹들의 semantic distribution을 광범위하고(global) 국소하게(local) 학습시킨다.

[47] dominated the 3D segmentation task

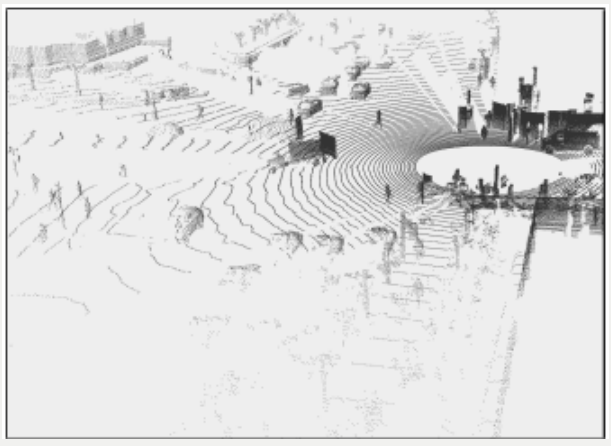
[47] Lidarmultinet: Towards a unified multi-task network for lidar perception



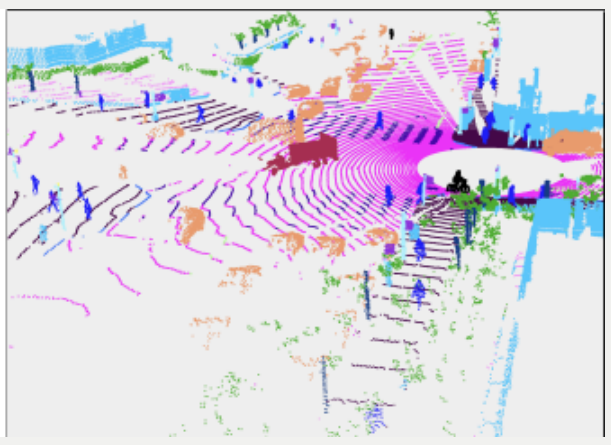
TL;DR

LiDAR Point Cloud를 input으로 활용하여 **하나의 네트워크를** 통해 1) 3D semantic segmentation 2) 3D object detection 3) panoptic segmentation을 동시에 진행한다.

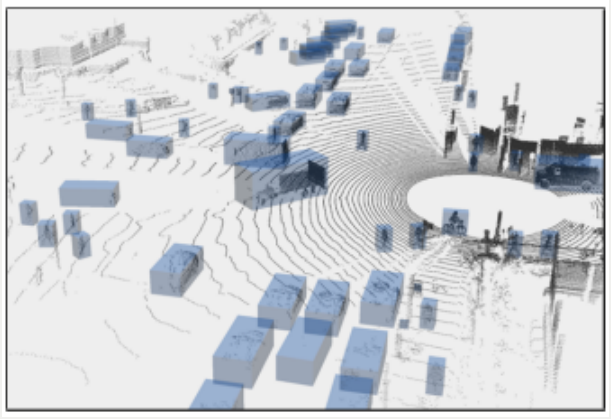
[Input: LiDAR Point Cloud]



[Output #1. 3D semantic segmentation]



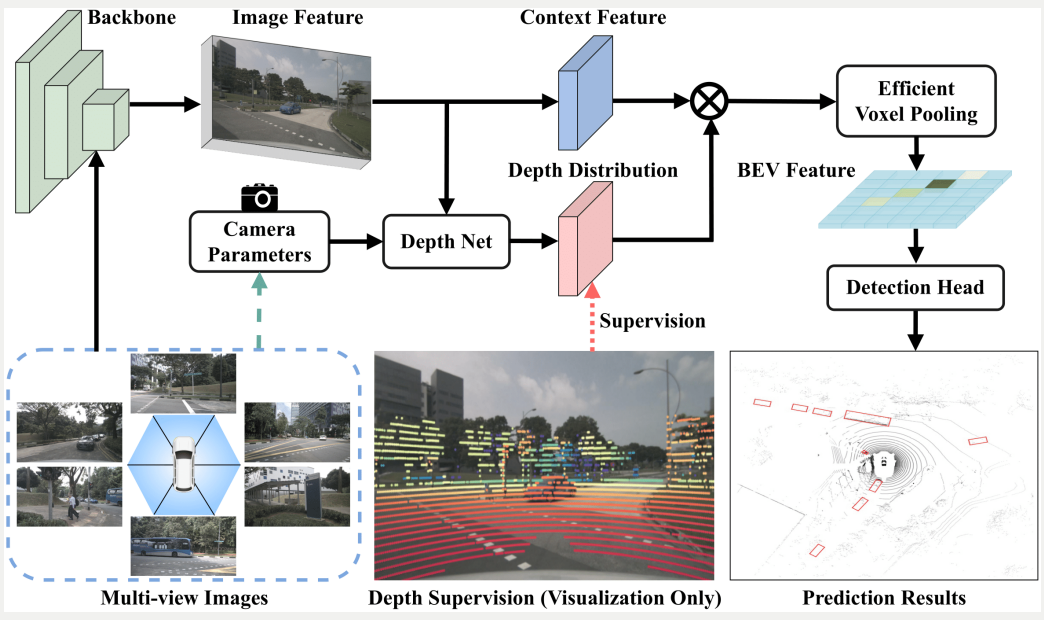
[Output #2. 3D object detection]



[Output #3. panoptic segmentation]

Still lag behind BEV-based methods on the 3D detection performance [27]

📝 **[27] Bevdepth: Acquisition of reliable depth for multi-view 3d object detection**
link: https://arxiv.org/pdf/2206.10092.pdf

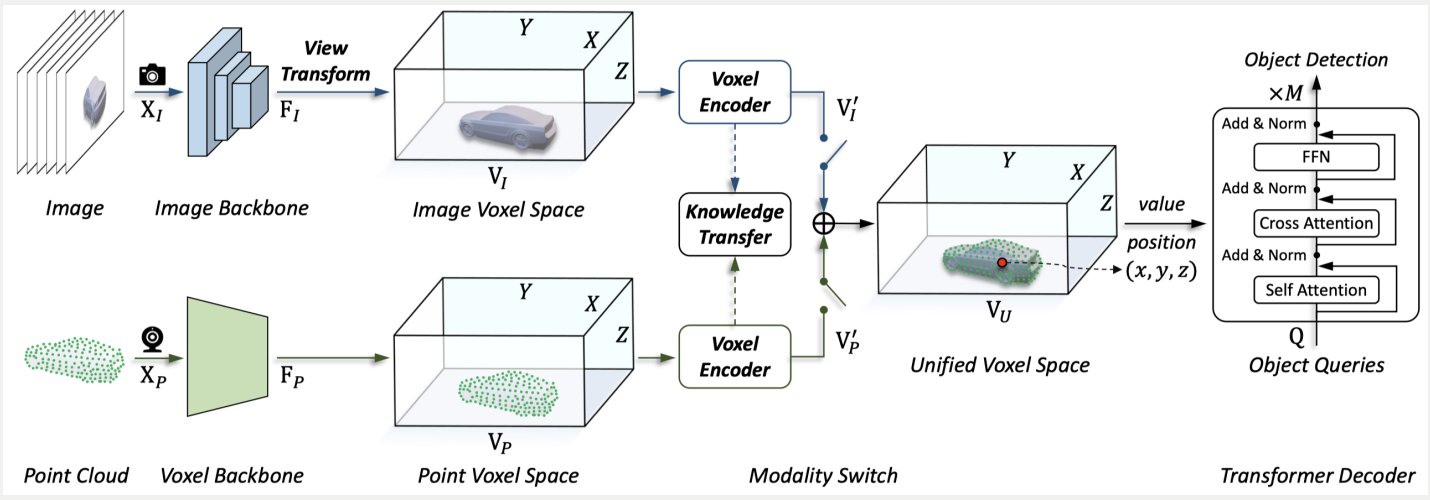


TL;DR

To overcome the deficiencies in the previous depth learning mechanism, we propose a new 3D object detector with a trustworthy depth estimation, dubbed BEVDepth, for camera-based Bird’s-Eye-View (BEV) 3D object detection.

Despite the success of voxel-based representations in LiDAR-centric surrounding perception, only a few works have explored voxel-based representations for vision-centric autonomous driving [5,26]

[26] Unifying voxel-based representation with transformer for 3d object detection
link: https://arxiv.org/pdf/2206.00630.pdf



TL;DR
Point clouds와 single (multi frame도 가능) image를 함께 활용하여 object detection을 진행한다.

[Framework Overview]
각각의 backboe(F_I, F_p)을 활용하여 각각의 Voxel Space에 mapping한다. 이후, voxel Encoder을 활용하여 하나의 통일된 Voxel Space에 mapping하고 transformer decoder을 통해서 object detection을 수행한다.