Representation for multi-modality 3D detection with transformer

Yanwei Li
CUHK

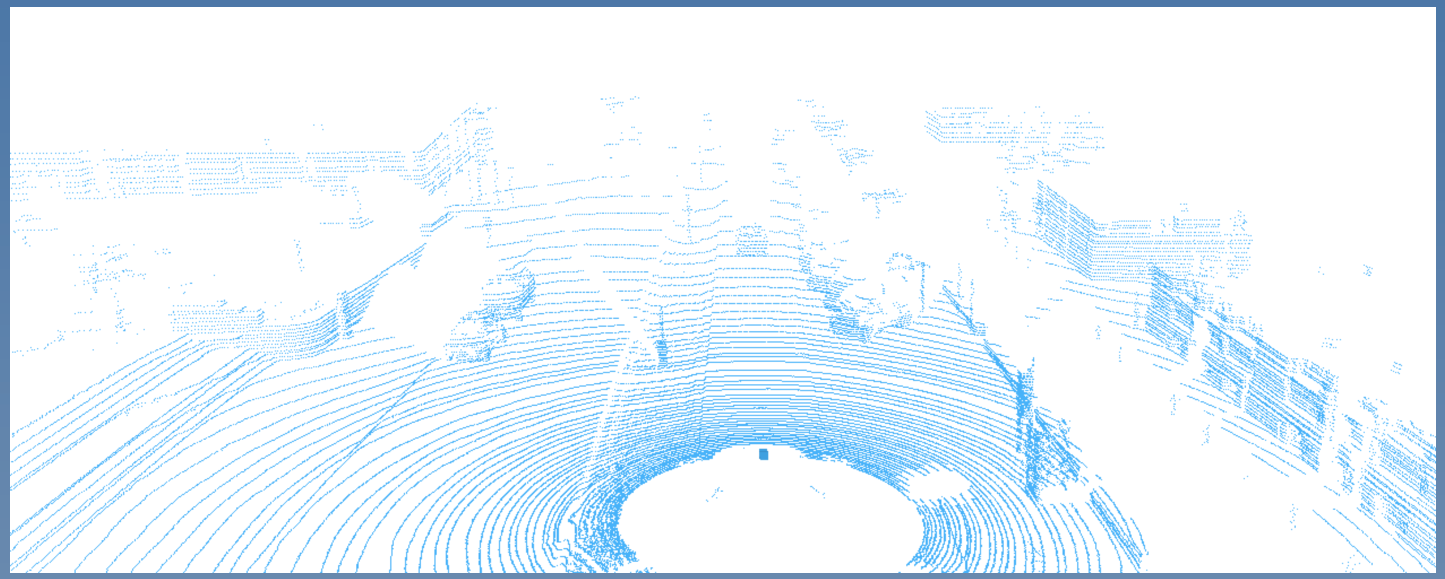
Contents

Introduction

Definition of 3D Object Detection

Locate and classify 3D objects from the given points or images.

Input with image



Input with LiDAR

3D object detection with different input.

Difficulties in 3D Object Detection

- *Input with image: lack accurate depth to establish structural representation for each object.*
- *Input with LiDAR: lack sufficient context to classify different categories for each object.*
- *Input with cross-modality is needed for accurate 3D Object Detection*

Introduction

Overview of multi-modal 3D Object Detection

Recent methods are roughly divided into Early-Fusion, Intermediate-Fusion, and Late-Fusion according to the fusion position.

Overview of multi-modality method for 3D detection [1]

[1] P3D Object Detection for Autonomous Driving: A Review and New Outlooks. *arXiv*, 2022.

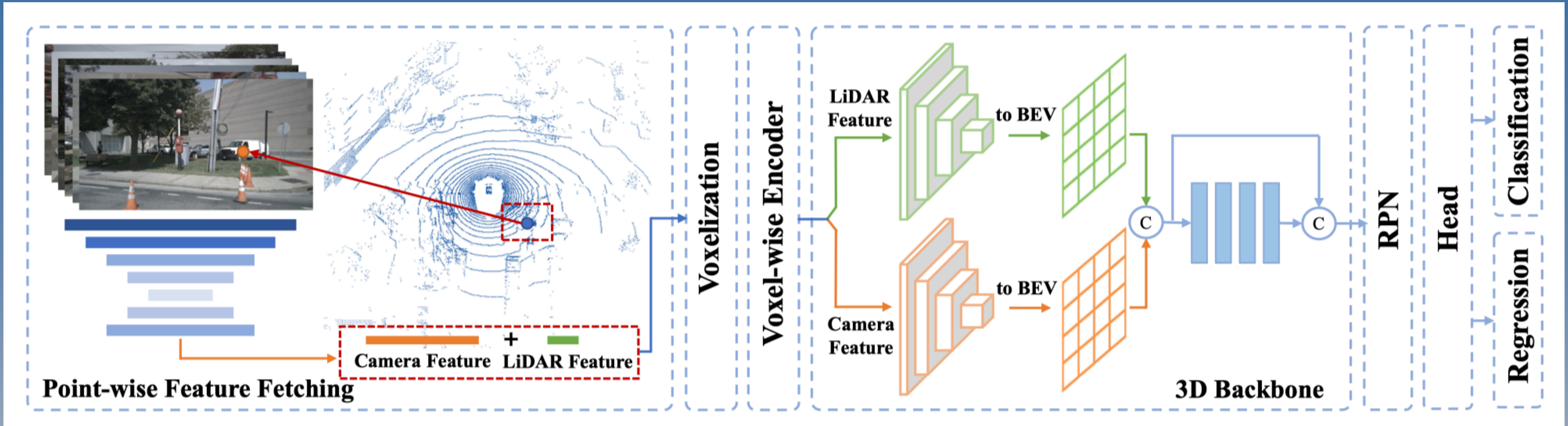
Introduction

General pipeline for cross-modality fusion

Image and point cloud are respectively processed in each network. Then, features are fused for prediction.

Key difficulties

- *How to find cross-modality features?*
- *How to align cross-modality augmentation?*



A general pipeline for cross-modality fusion in 3D detection [2]

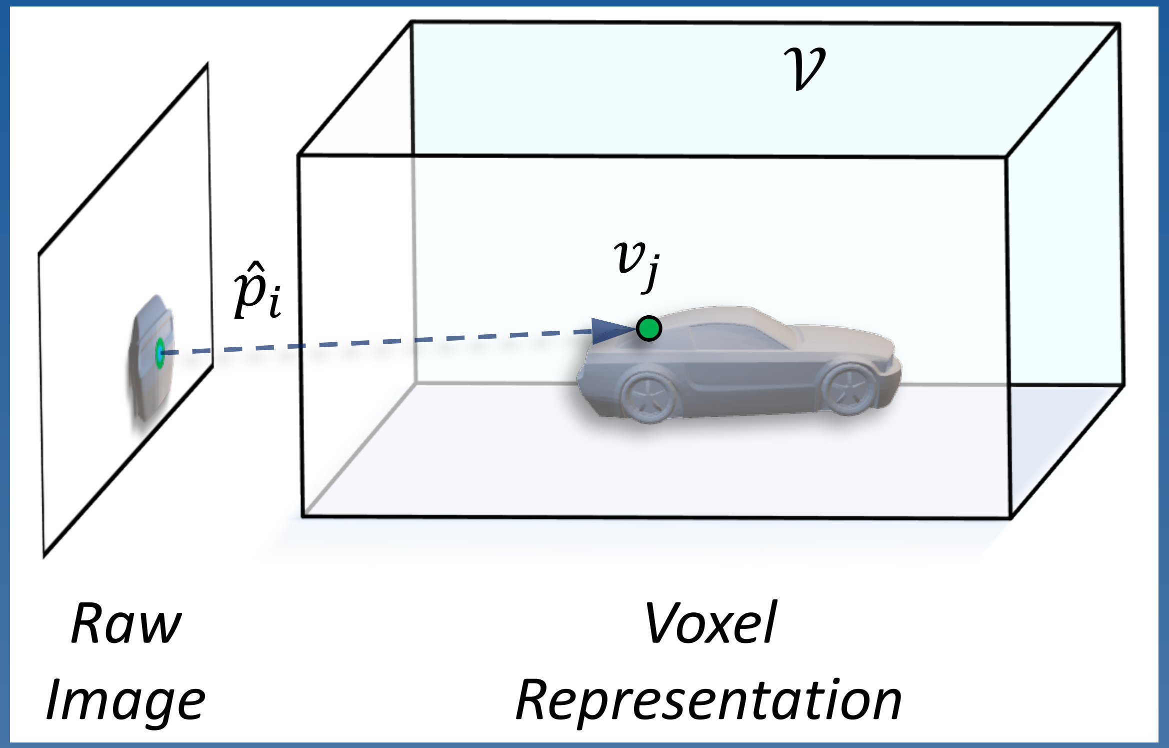
[2] PointAugmenting: Cross-Modal Augmentation for 3D Object Detection. In *CVPR*, 2021.

Introduction

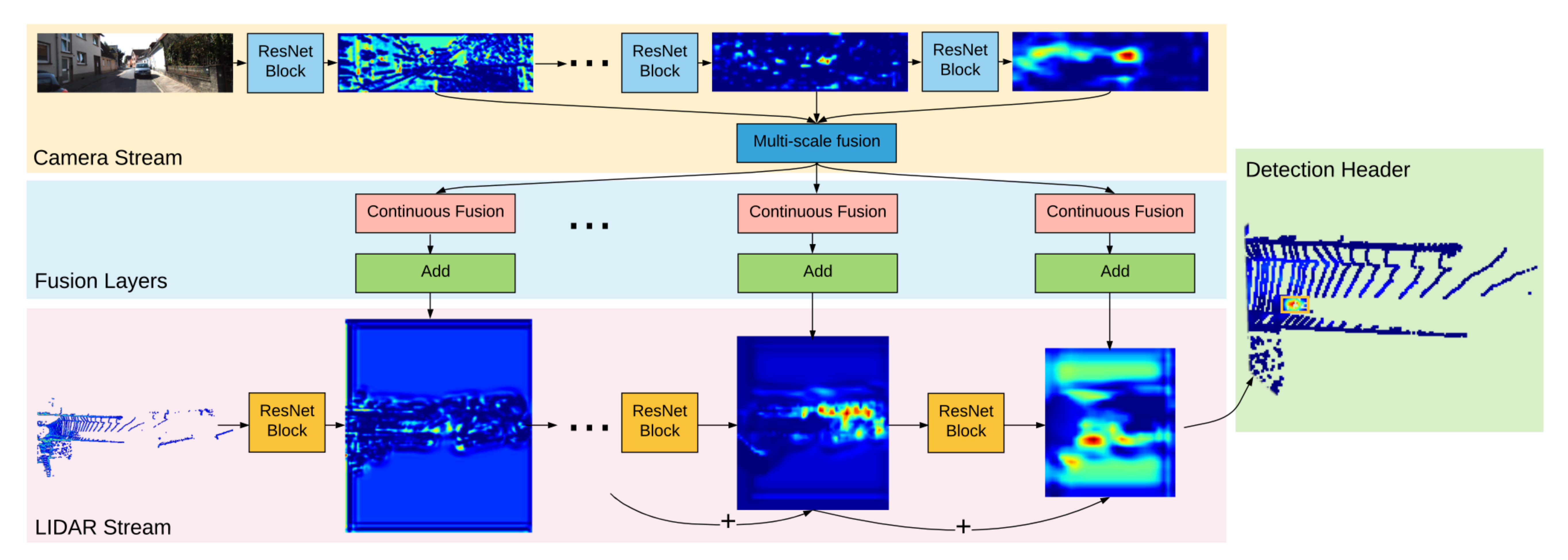
Previous methods fuse in a point-to-point manner

- *Find one-one correspondence across modality according to projection matrix.*
- *Fuse point feature and retrieved image features directly in a point-to-point manner.*

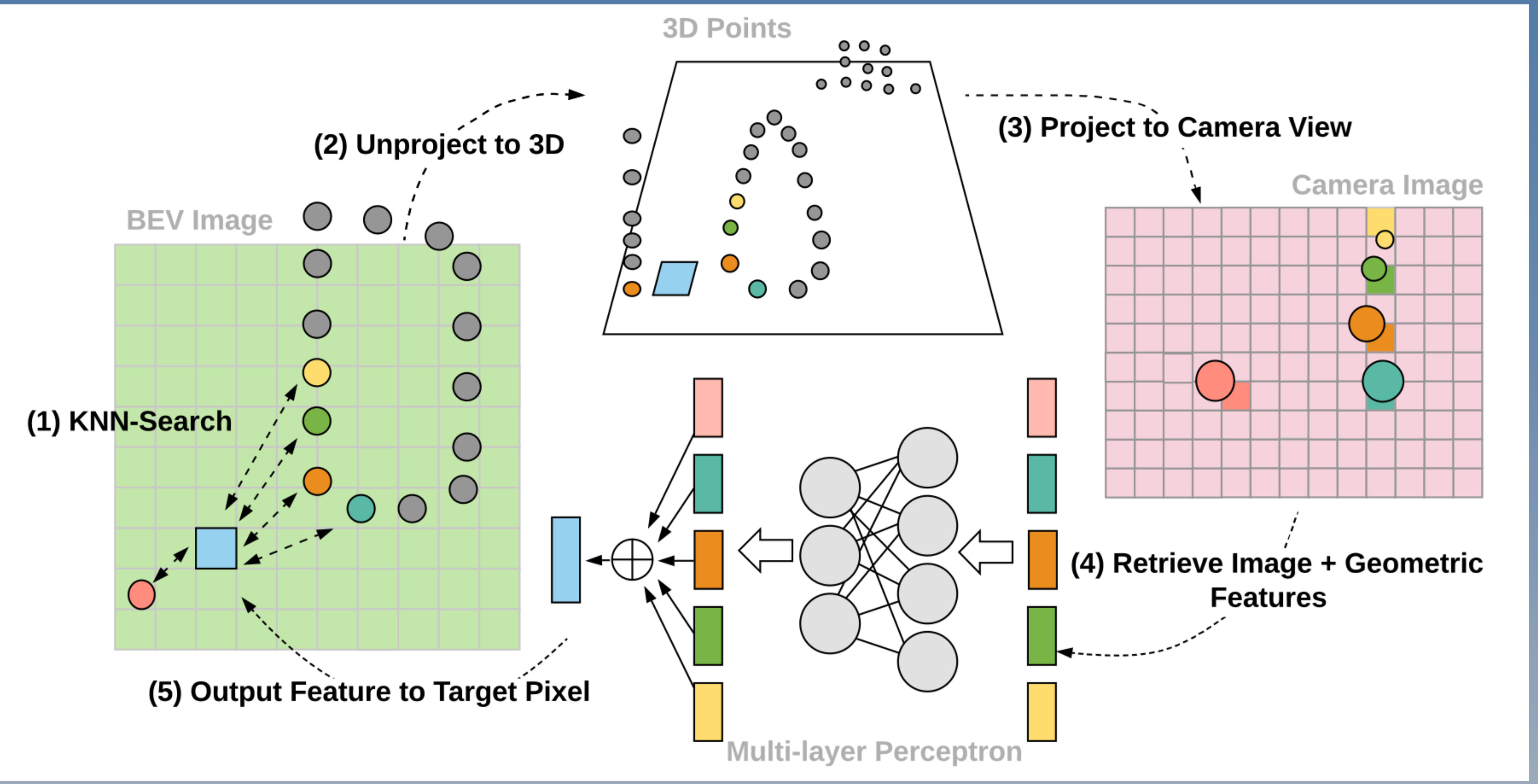
Drawback: constrained by the sparsity of point cloud.



Point-to-point fusion



Architecture of Deep Continuous Fusion [3].



Point-to-point feature retrieve process [3].

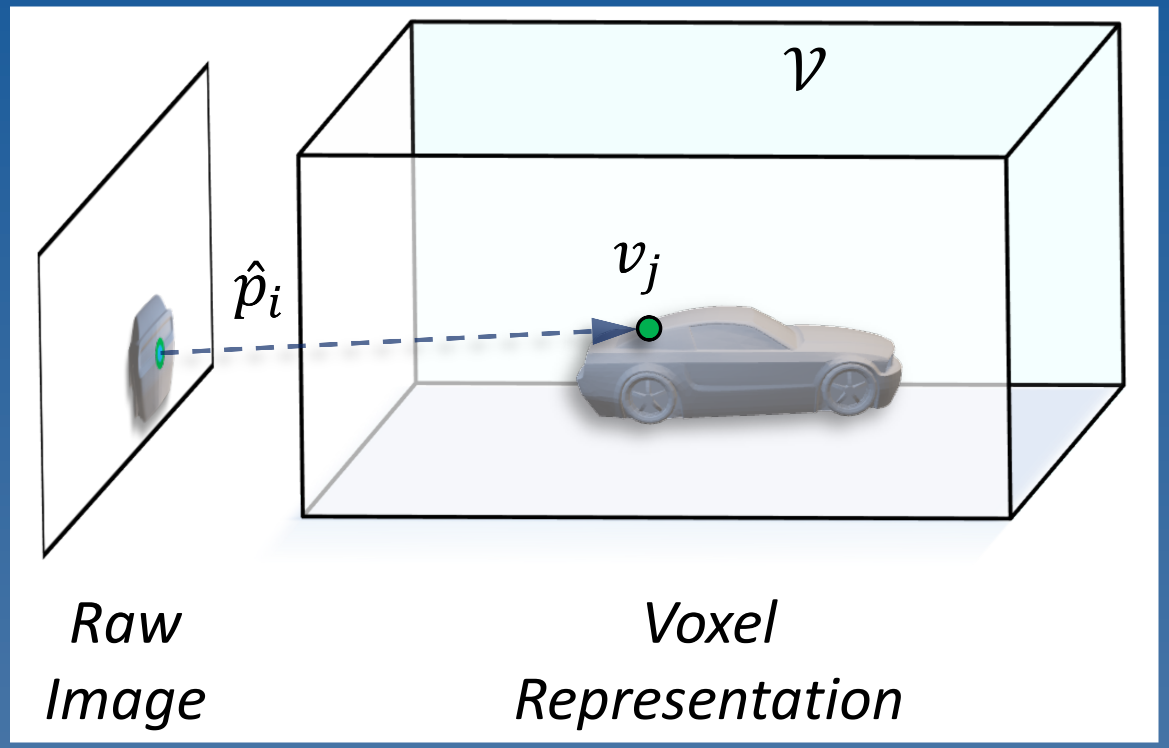
[3] Deep continuous fusion for multi-sensor 3d object detection. In *ECCV*, 2018.

Introduction

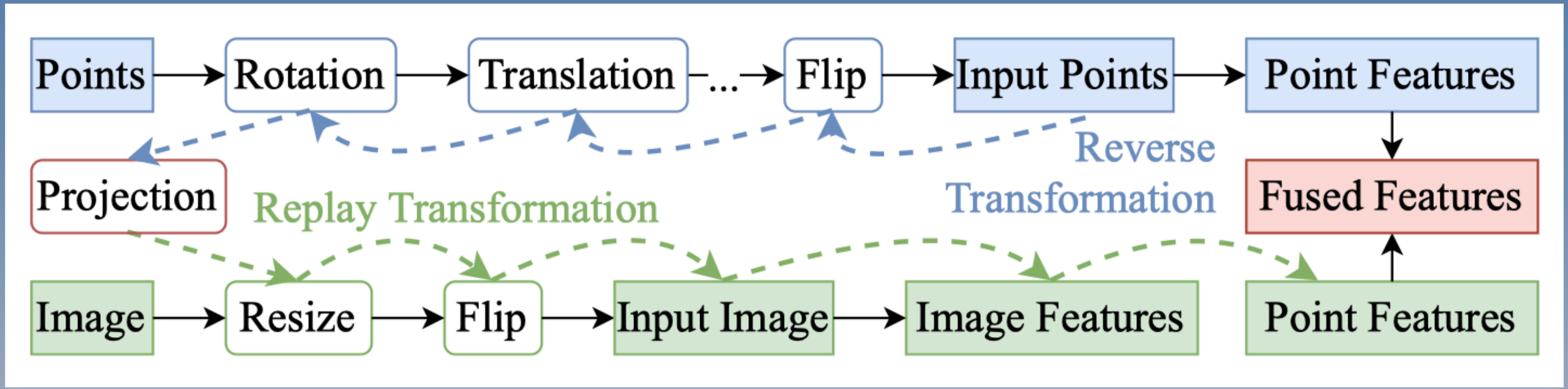
Previous methods damage consistent augmentation

- *LiDAR: scene-level flipping, rescaling, and rotation.*
- *Image: no image-level data augmentation.*
- *Cross-modality: inverse LiDAR point to find correspondence.*

Drawback: out-of-sync augmentation damage consistency.



Point-to-point fusion



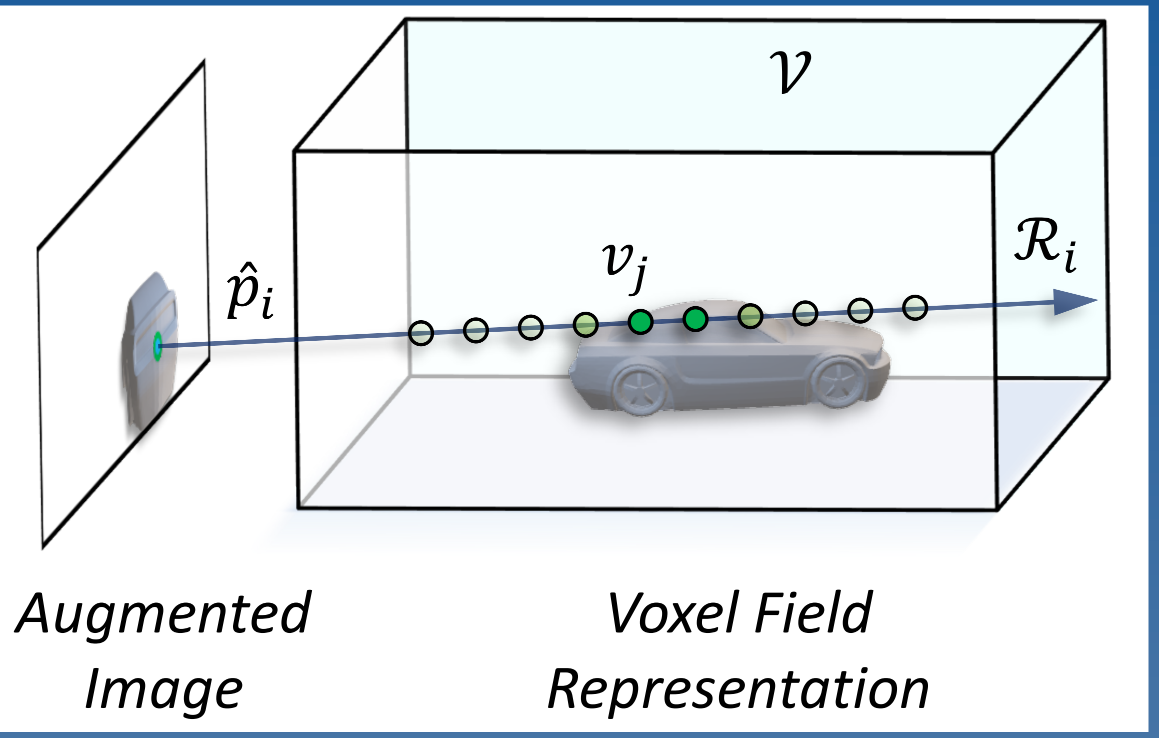
Classic multi-modality transformation flow [4].

[4] Exploring Data Augmentation for Multi-Modality 3D Object Detection. arXiv:2012.12741, 2020.

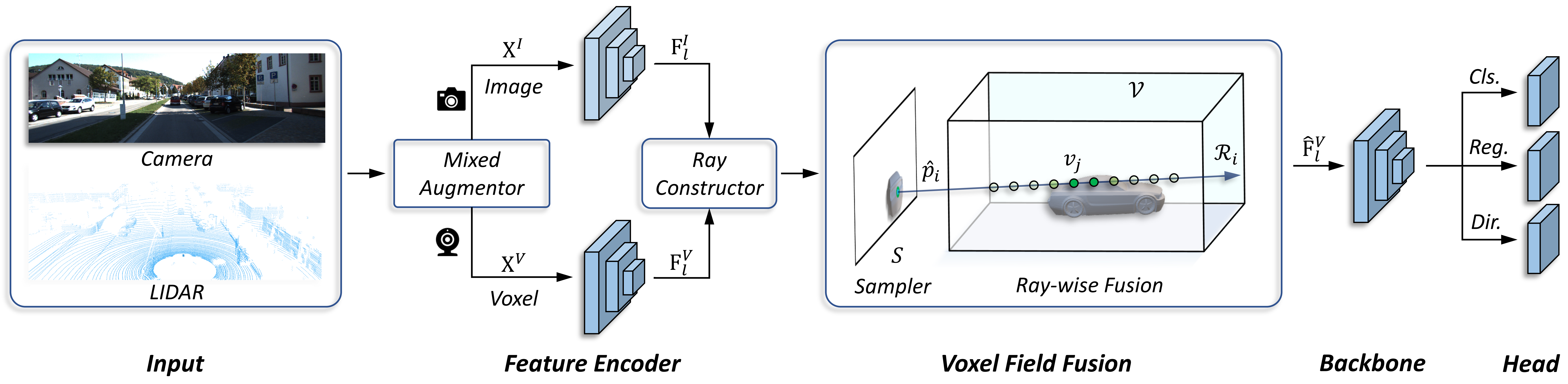
Voxel Field Fusion

Voxel Field Fusion maintain the consistency

- *Feature representation*: project augmented image features to voxel space and represent in a point-to-ray manner.
- *Data augmentation*: synced image-level augmentation according to that in point cloud.



Point-to-ray fusion



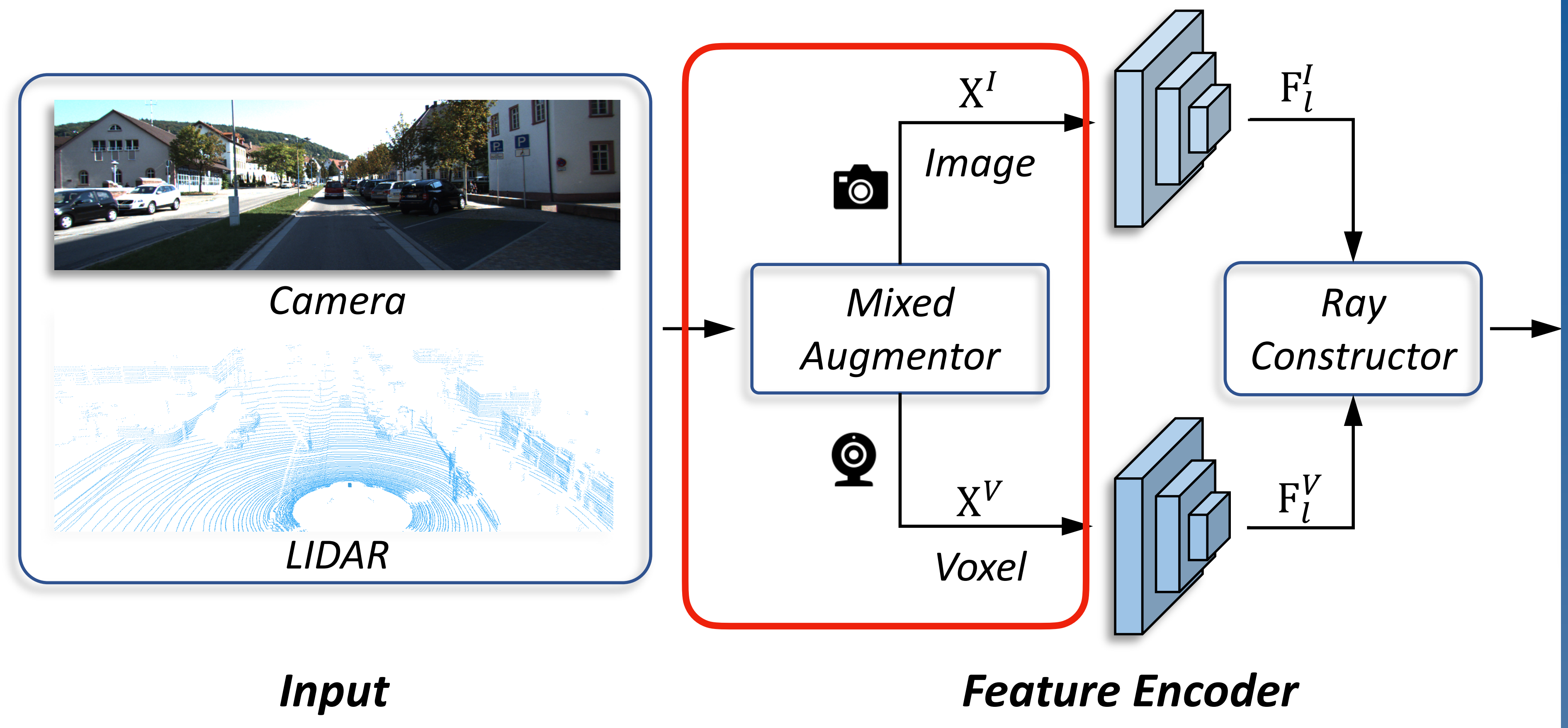
The framework for 3D detection with voxel field fusion.

[5] Voxel Field Fusion for 3D Object Detection. In *CVPR*, 2022.

Voxel Field Fusion

Mixed Augmentor

- *Sample-added*: supplement the RGB data of sampled 3D objects in a copy-paste manner, i.e., 3D GT-sampling.
- *Sample-static*: scene-level augmentation combined with image-level flipping and rescaling.



Voxel Field Fusion process.

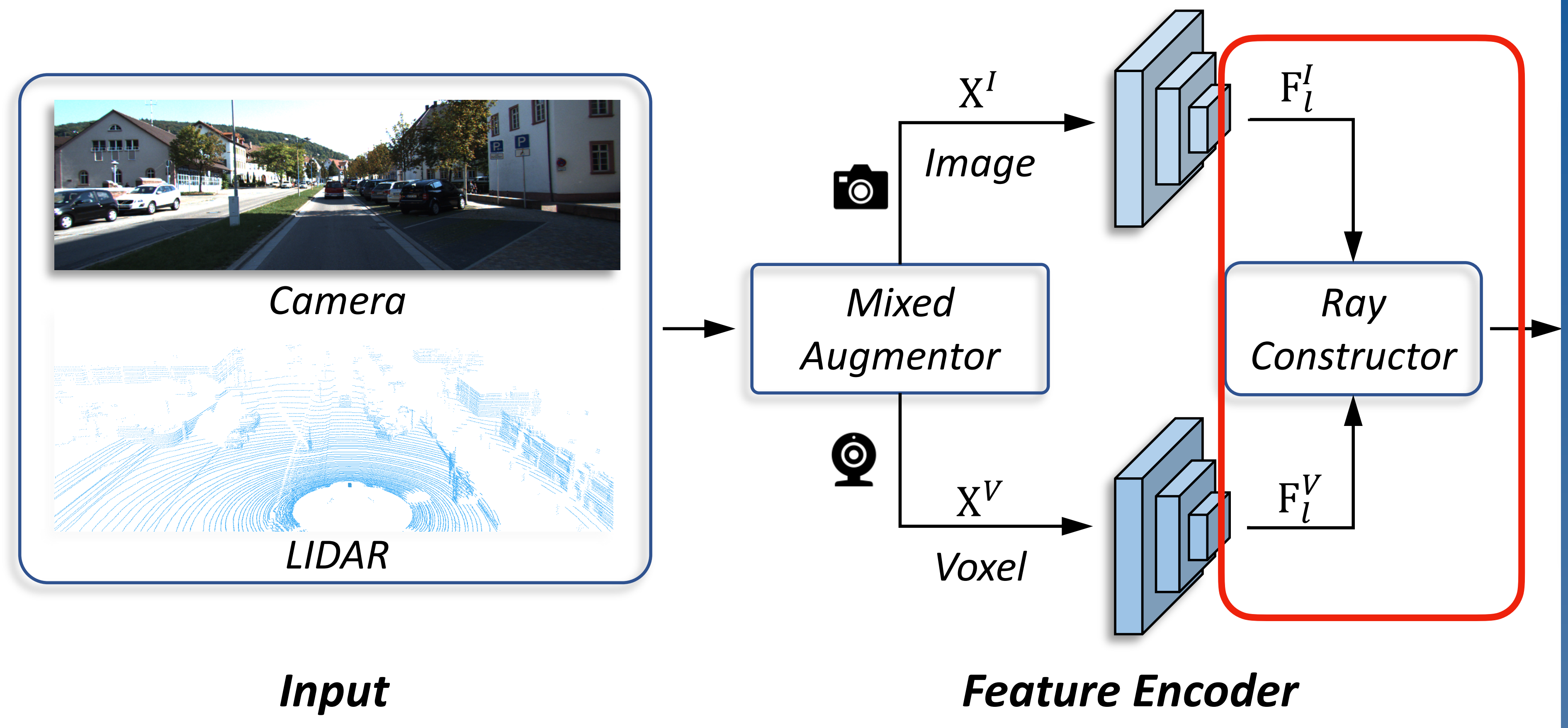
Corresponding operations in the mixed augmentor.

type	Point operation	Image operation
Sample-added	GT-sampling	Copy-paste
Sample-static	Flip Rescale Rotate	Image-flip Image-rescale Reproject

Voxel Field Fusion

Mixed Augmentor

- *Sample-added: supplement the RGB data of sampled 3D objects in a copy-paste manner, i.e., 3D GT-sampling.*
- *Sample-static: scene-level augmentation combined with image-level flipping and rescaling.*



Voxel Field Fusion process.

Ray Constructor

- *Establish the cross-modality correspondence from voxel bin v_j to image pixels p_i.*

$$p_i = v_j \mathbf{T}^T_{\text{Voxel}\to\text{Image}}, \forall v_j \in \mathcal{R}_i \,.$$

Corresponding operations in the mixed augmentor.

type	Point operation	Image operation
Sample-added	GT-sampling	Copy-paste
Sample-static	Flip Rescale Rotate	Image-flip Image-rescale Reproject

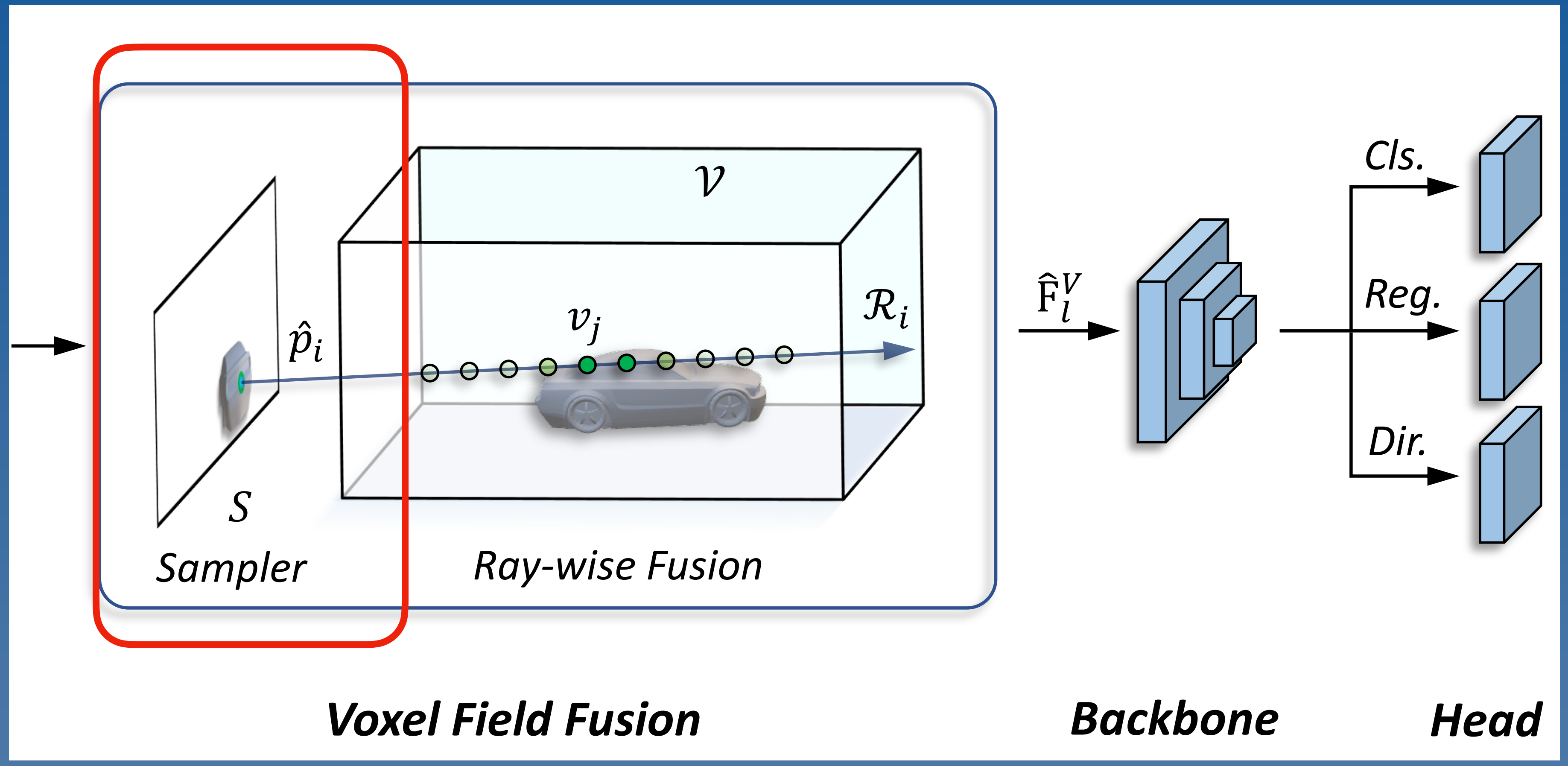
Voxel Field Fusion

Efficient learnable sampler

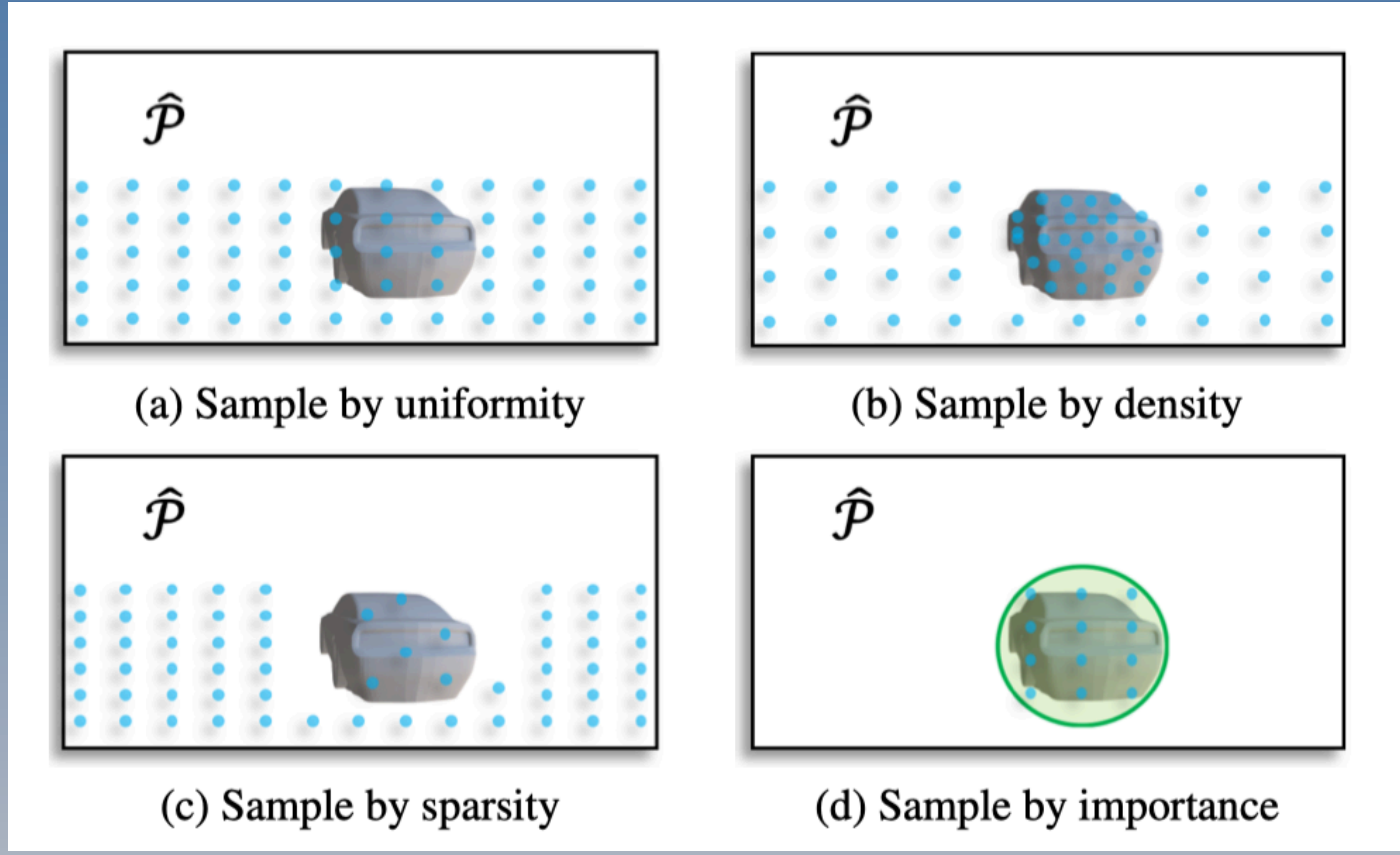
- Sample vital image features for ray construction according to the importance of foreground objects.

Design choices

- Sample by uniformity: uniformly sample image features for ray construction.
- Sample by density: sample image features for ray construction according to the density of projected LiDAR points.
- Sample by sparsity: sample image features for ray construction according to the density of projected LiDAR points.



Voxel Field Fusion process.



Toy examples of different sampling methods.

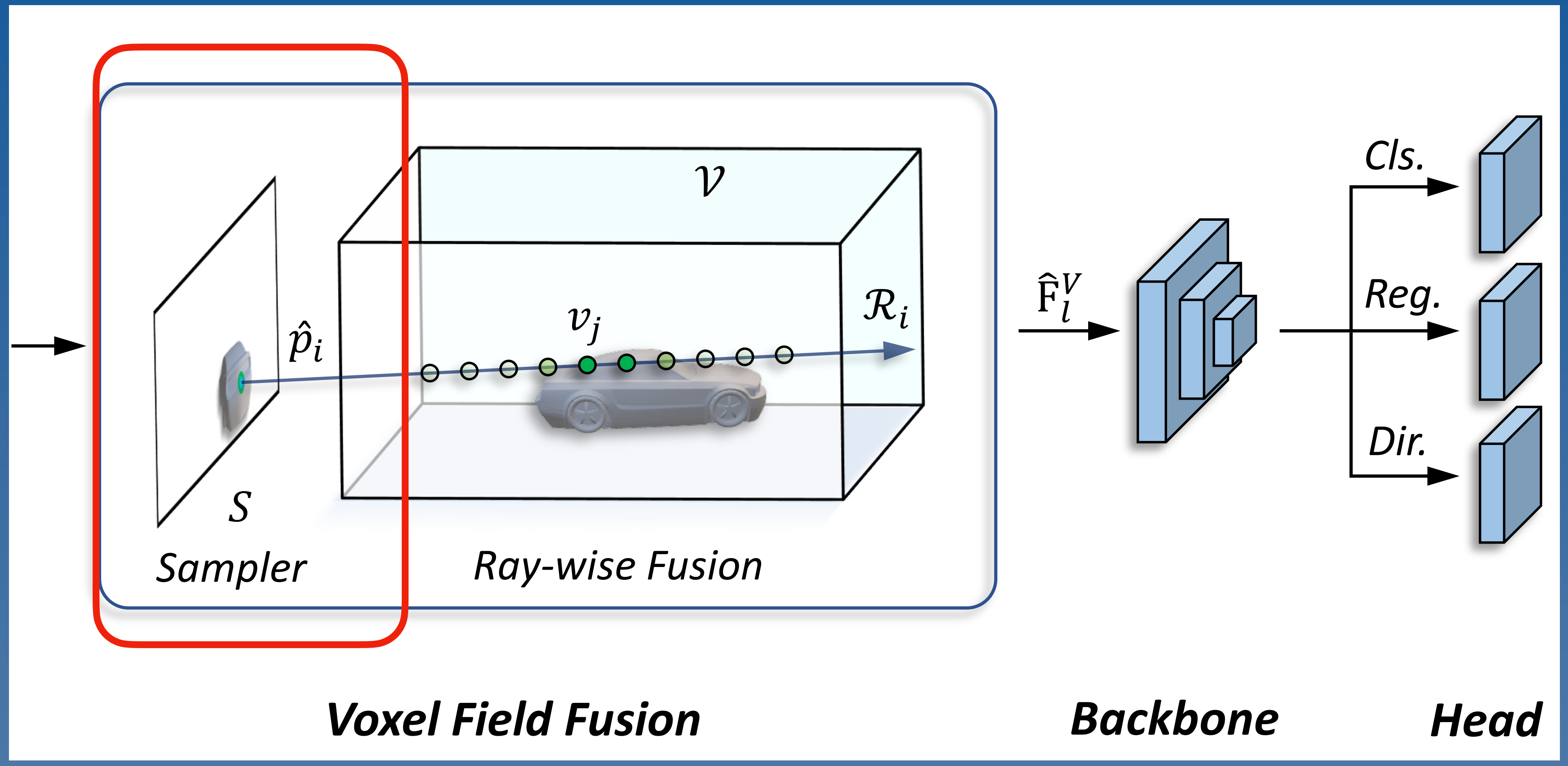
Voxel Field Fusion

Efficient learnable sampler

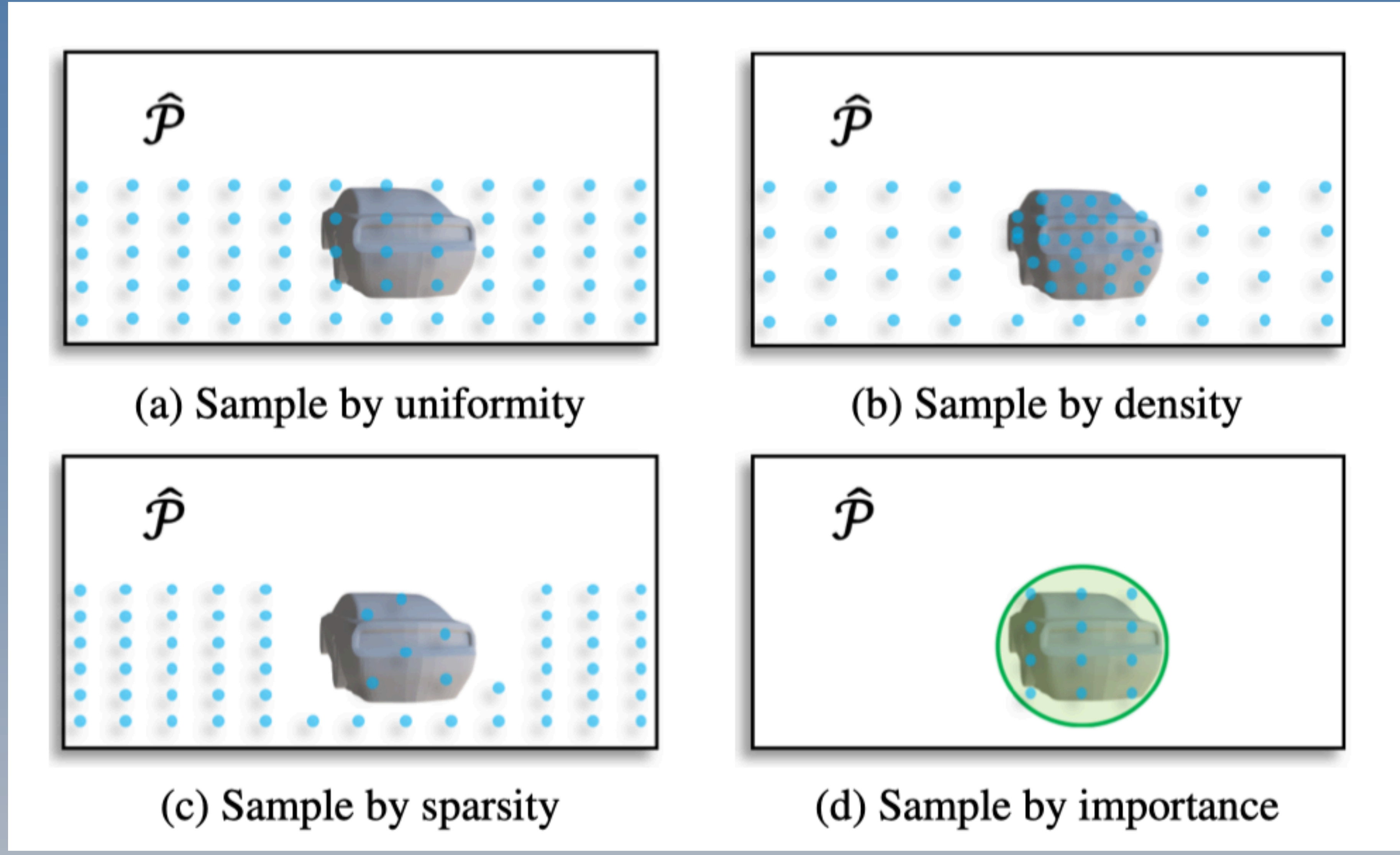
- *Sample vital image features for ray construction according to the importance of foreground objects.*

Design choices

- *Sample by uniformity: uniformly sample image features for ray construction.*
- *Sample by density: sample image features for ray construction according to the density of projected LiDAR points.*
- *Sample by sparsity: sample image features for ray construction according to the density of projected LiDAR points.*
- *Sample by importance: sample image features for ray construction according to the predicted importance.*



Voxel Field Fusion process.



Toy examples of different sampling methods.

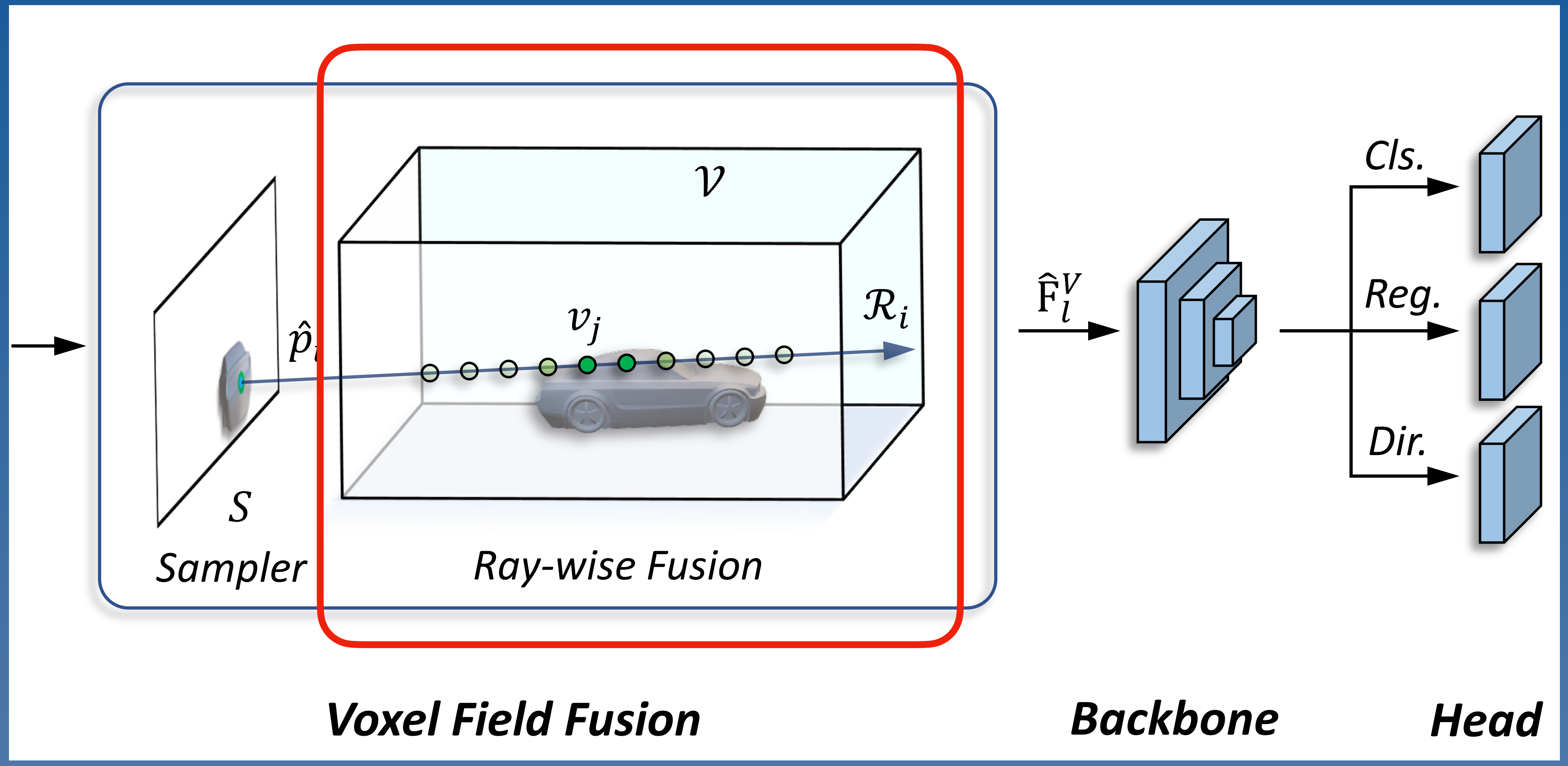
Voxel Field Fusion

Ray-voxel interaction

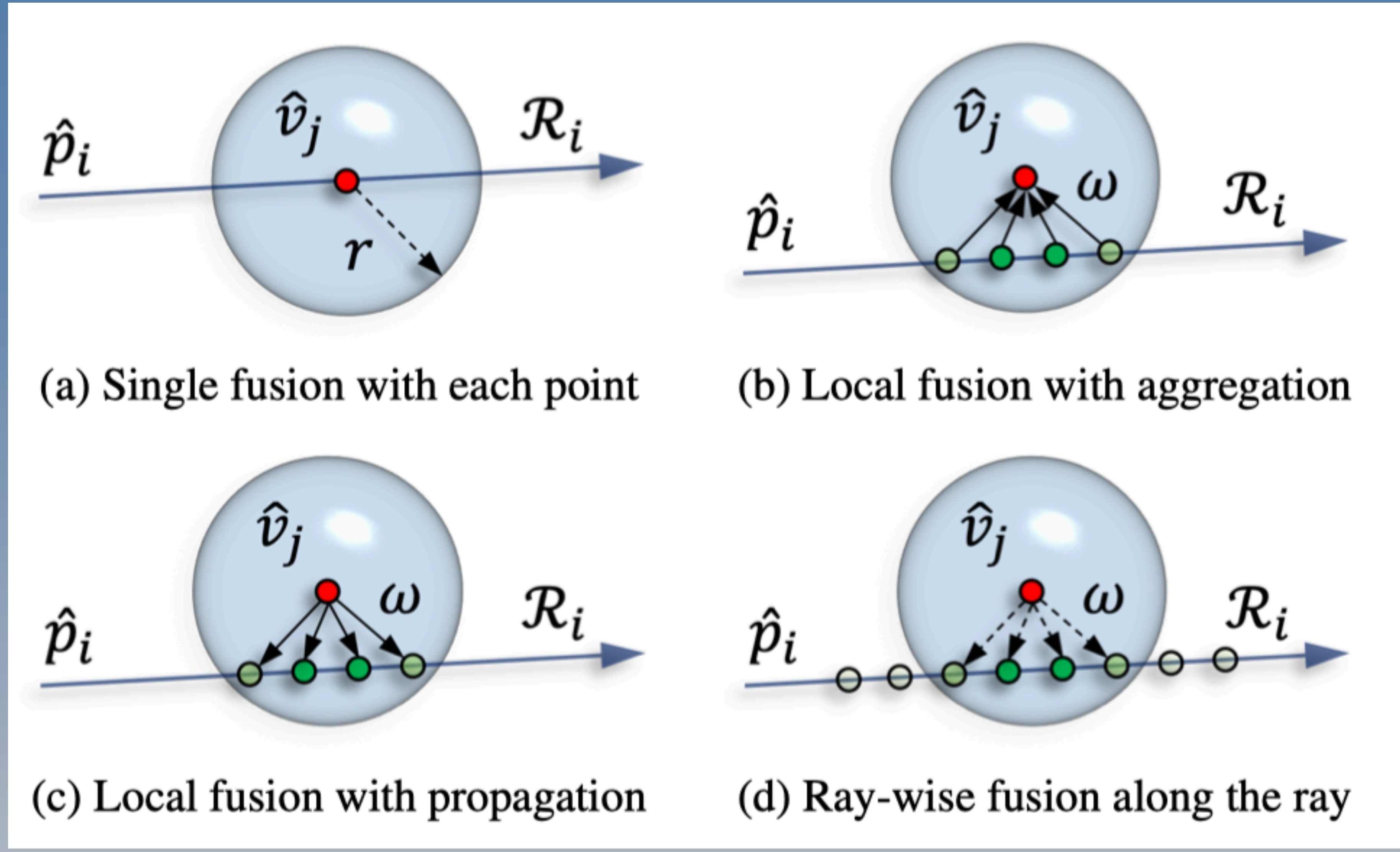
- Ray-wise fusion extends the operation and fuses, as well as newly generates, the high-responded features along the ray.

Design choices

- Single fusion: only fuses the single point as traditional method.
- Local aggregation: aggregates all the neighboring features to the anchor voxel within a radius along the ray.
- Local propagation: propagates the feature of anchor voxel to all the neighboring points within a radius along the ray.



Voxel Field Fusion process.

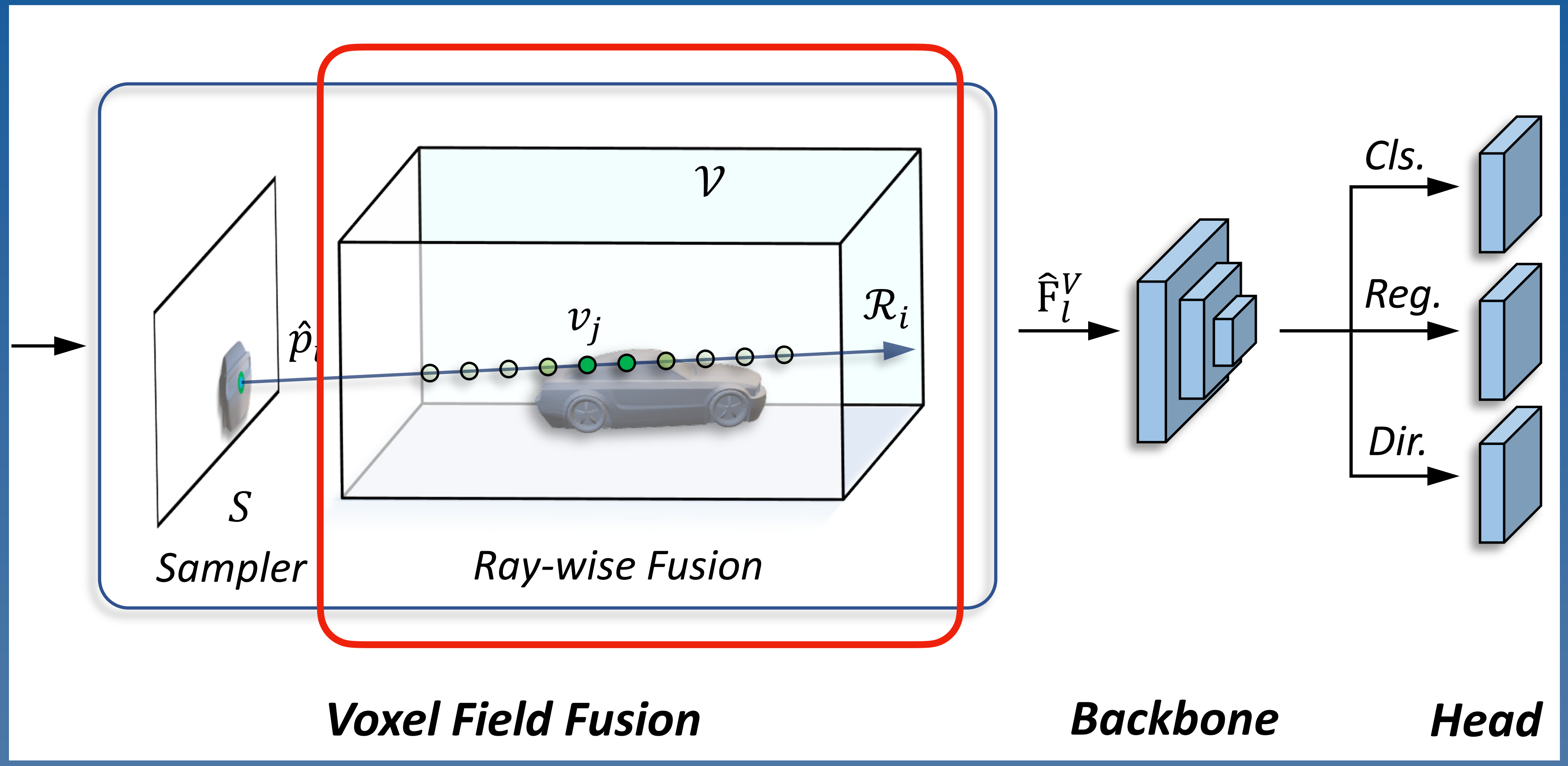


Toy examples of different fusion methods.

Voxel Field Fusion

Ray-voxel interaction

- *Ray-wise fusion extends the operation and fuses, as well as newly generates, the high-responded features along the ray.*

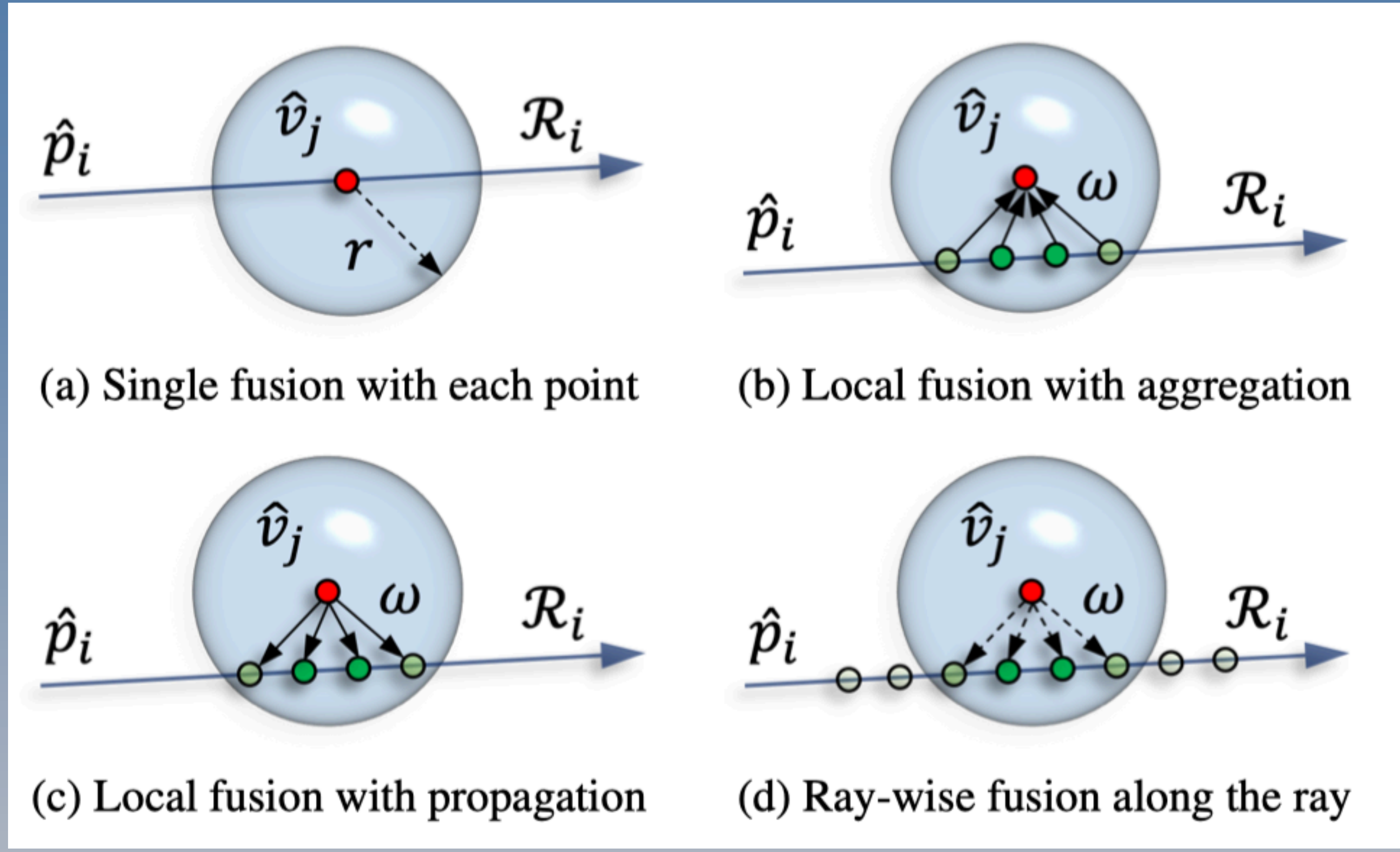


Voxel Field Fusion process.

Design choices

- *Single fusion: only fuses the single point as traditional method.*
- *Local aggregation: aggregates all the neighboring features to the anchor voxel within a radius along the ray.*
- *Local propagation: propagates the feature of anchor voxel to all the neighboring points within a radius along the ray.*
- *Ray-wise fusion: (1) Training: distributes the existence probability of each point within a radius along the ray; (2) Inference: fuses all the high-responded points.*

$$\widehat{\mathcal{F}}(x_j, y_j, z_j) = \mathcal{F}(x_j, y_j, z_j) + \omega_j f([\mathbf{F}^I_{l,i}, \mathbf{F}'_{l,v_j}])\ .$$



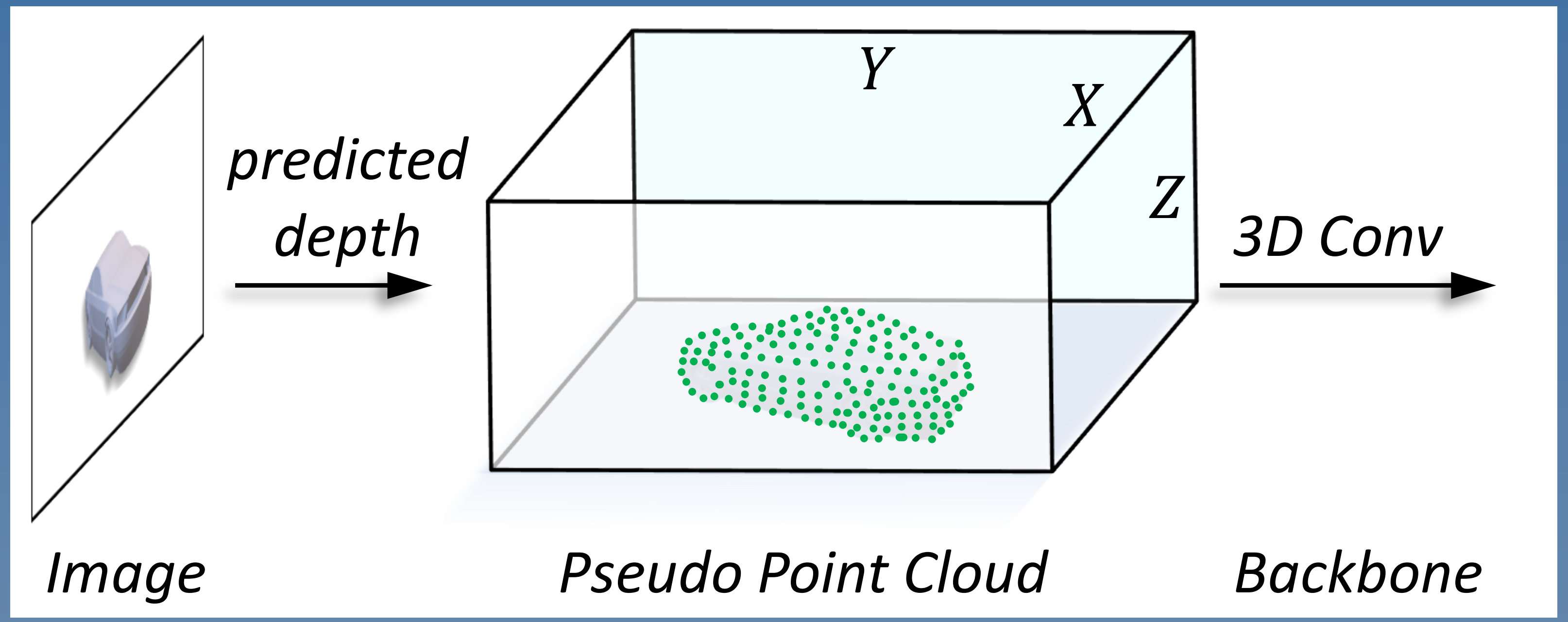
Toy examples of different fusion methods.

Unified Representation

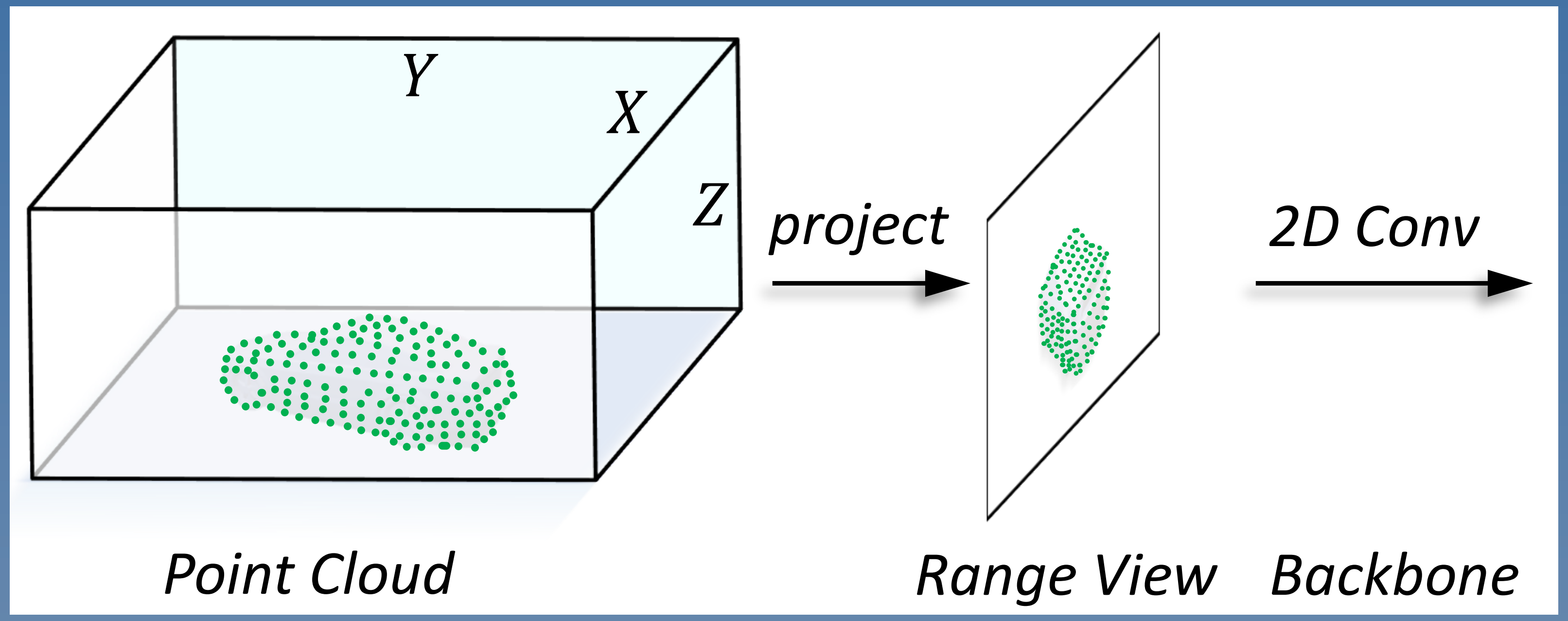
Weakness in previous work

- Using point cloud for feature reference reduces robustness of camera-only models.
- Previous approaches introduces semantic ambiguity.

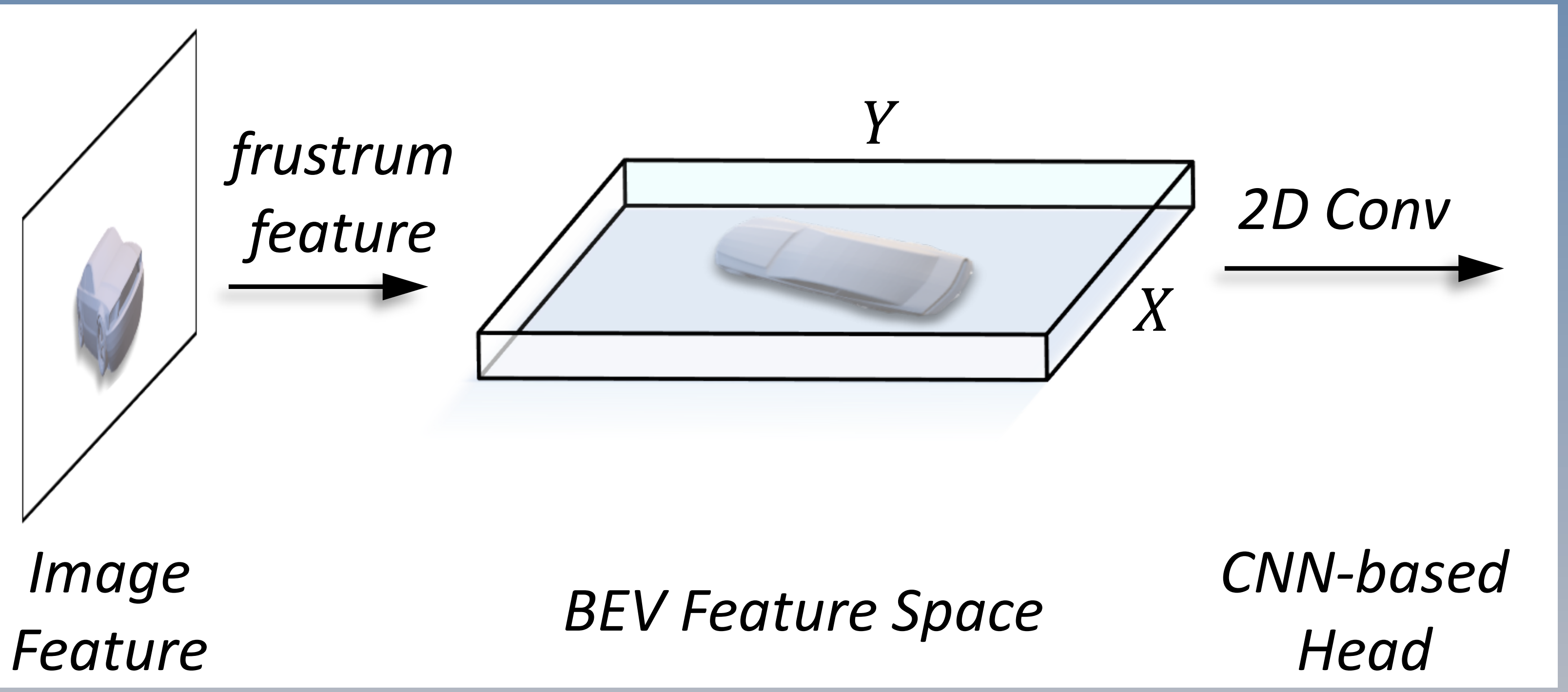
A more unified representation is desired to bridge modality gap and facilitate interactions.



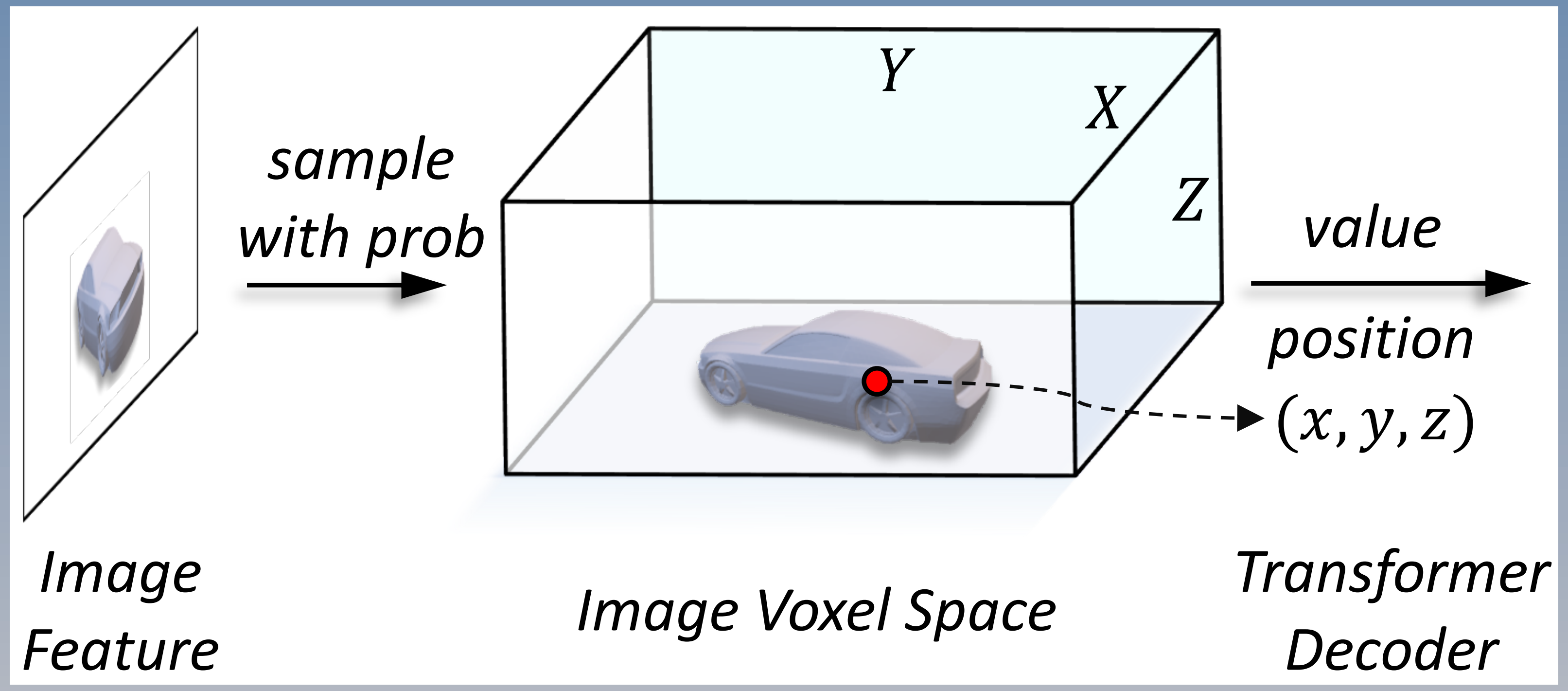
Pseudo point cloud transformed from image



Range view projected from point cloud



BEV feature space from image feature



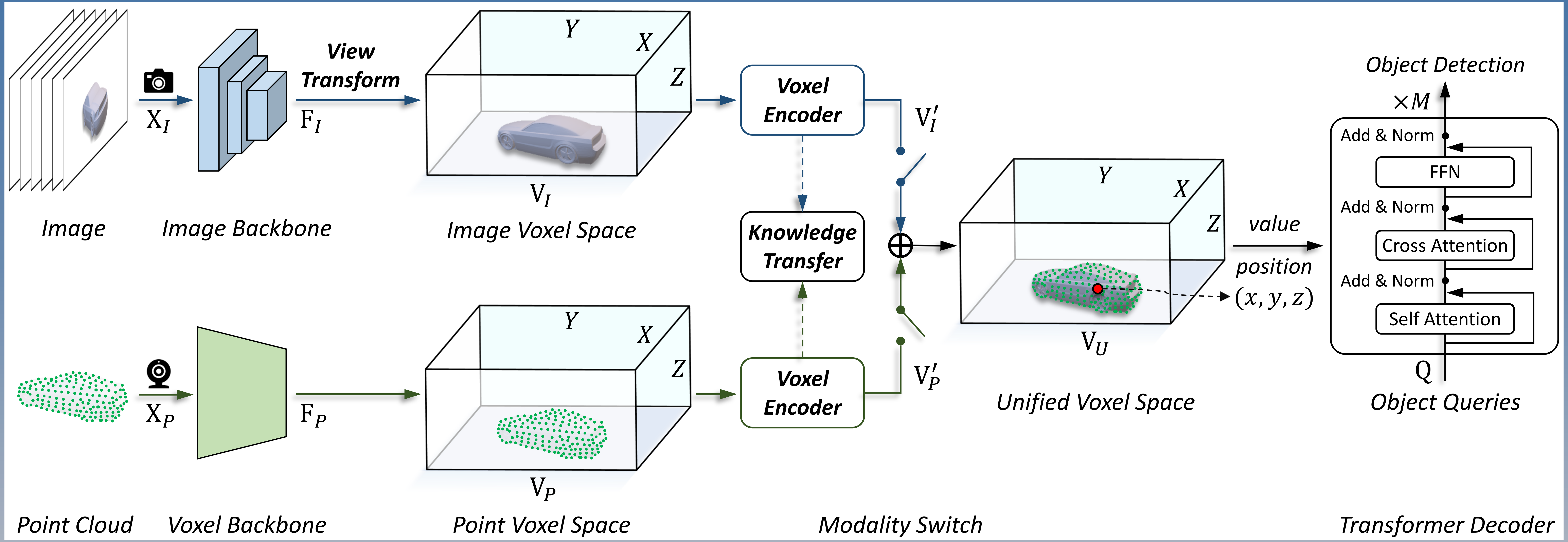
Voxel feature space from image feature

Unified Representation

UVTR for unified representation

- *Modality-specific Space*: construct unified representation.
- *Cross-modality interaction*: feature learning across spaces.
- *Transformer decoder*: object-level interaction and prediction.

Similar *feature representation* and *data augmentation* for different modalities.



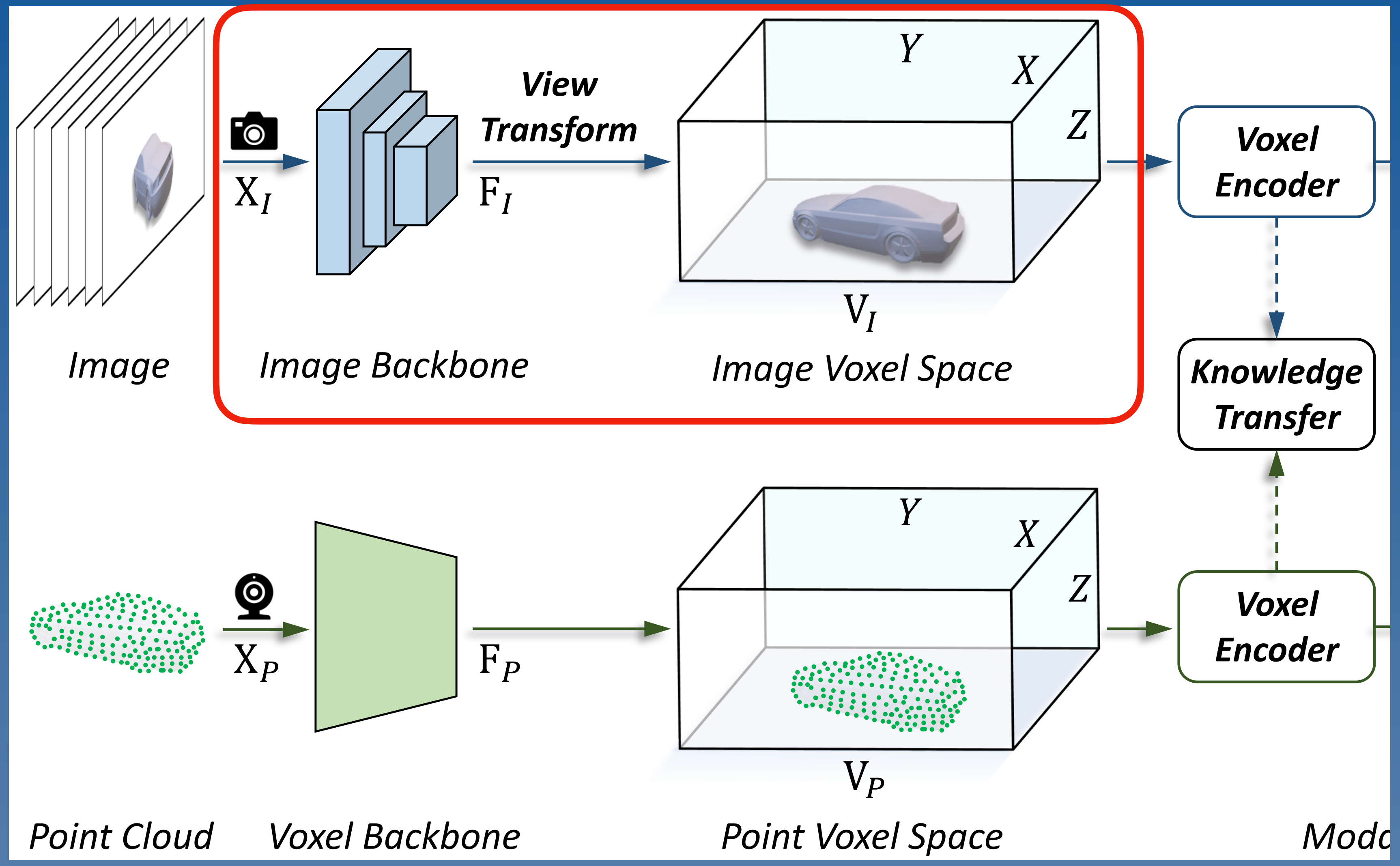
The framework of UVTR with multi-modality input.

[6] Unifying Voxel-based Representation with Transformer for 3D Object Detection. arXiv:2206.00630, 2022.

Unified Representation

Modality-specific Space

- *Given images captured from cameras and point cloud from LiDAR, different branches are utilized to generate and enhance voxel space for each modality*



Design choices

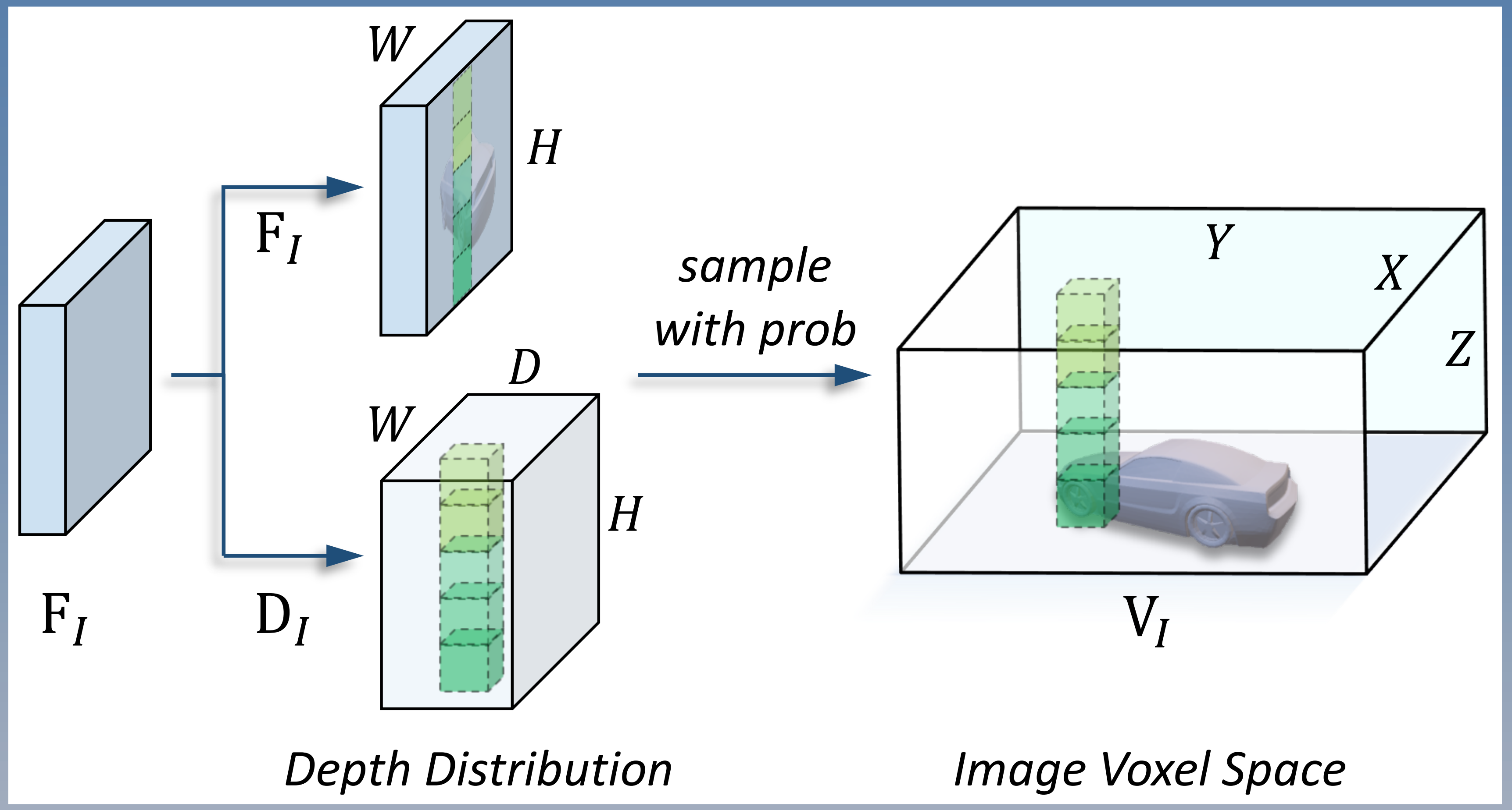
- *Image Voxel Space: construct voxel space from multi-view images using shared backbone and predicted depth.*

Generate the depth distribution

$$\mathbf{D}_I(u, v) = \mathrm{Softmax}(\mathrm{Conv}(\mathbf{F}_I)(u, v))$$

Transfer image feature to voxel space

$$\mathbf{V}_I(x, y, z) = \mathbf{D}_I(u, v, d) \times \mathbf{F}_I(u, v)$$

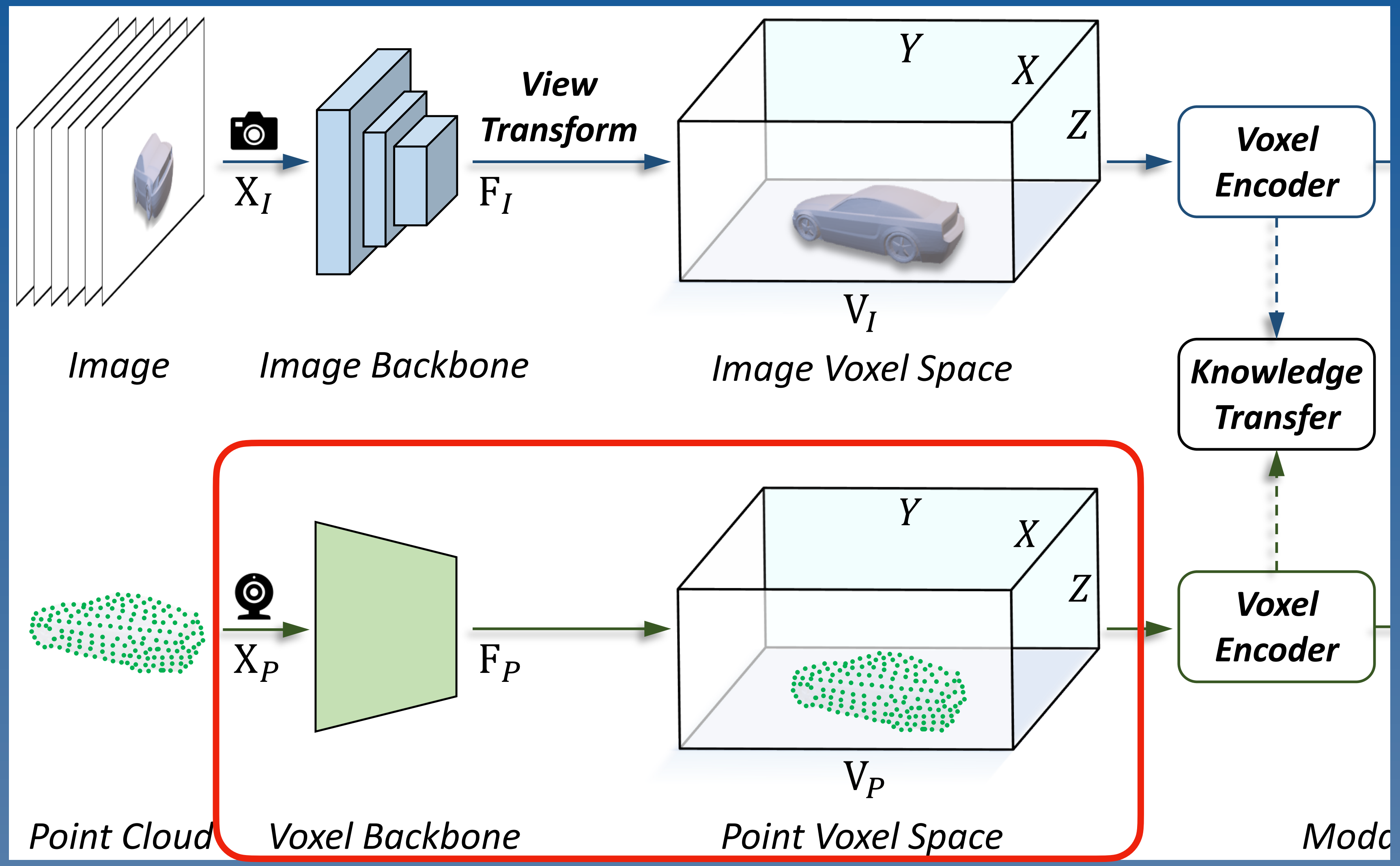


Details in the view transform.

Unified Representation

Modality-specific Space

- *Given images captured from cameras and point cloud from LiDAR, different branches are utilized to generate and enhance voxel space for each modality.*



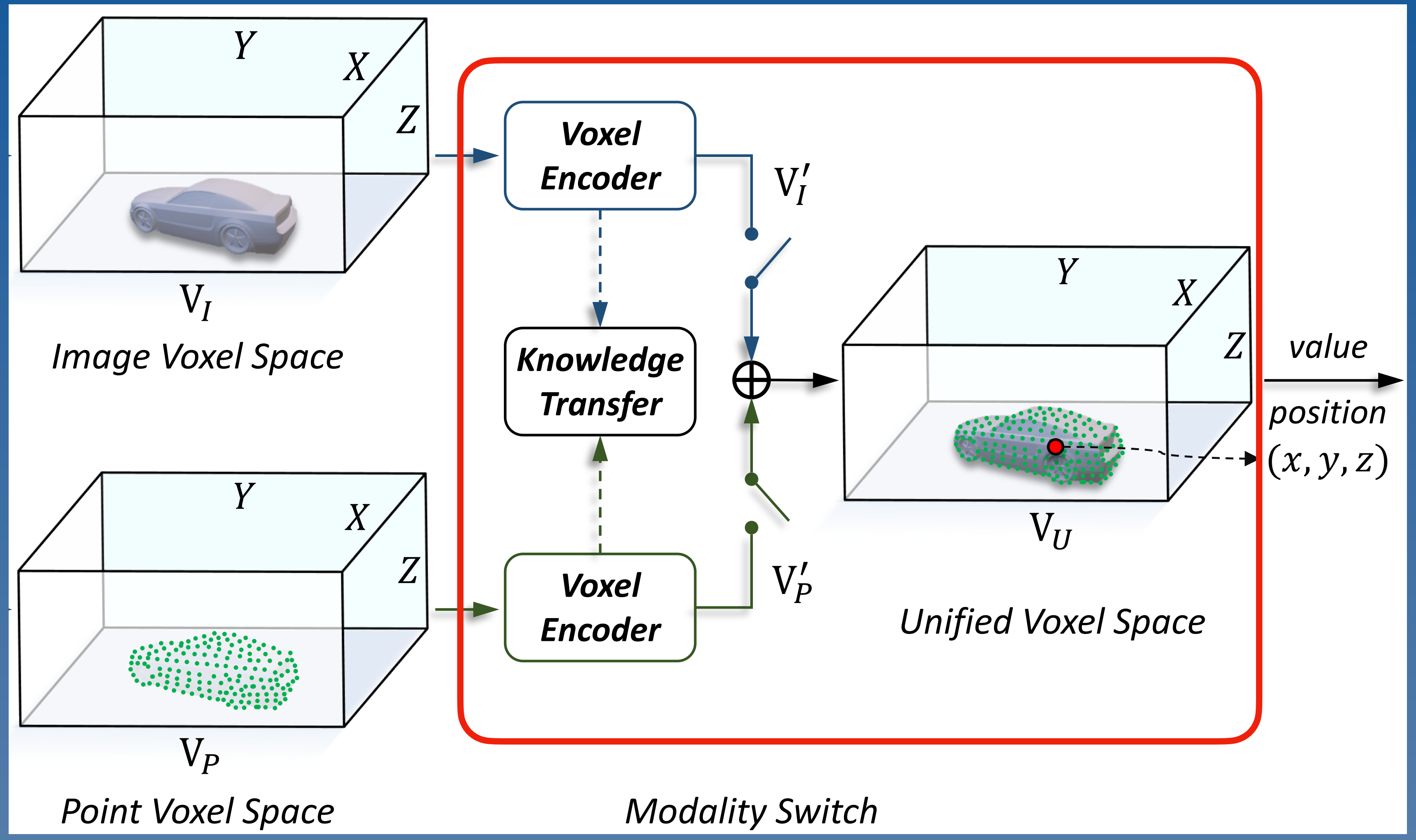
Design choices

- *Image Voxel Space: construct voxel space from multi-view images using shared backbone and predicted depth.*
- *Point Voxel Space: construct voxel space from point cloud using sparse convolution.*
- *Voxel Encoder: feature interaction among adjacent voxels.*

Unified Representation

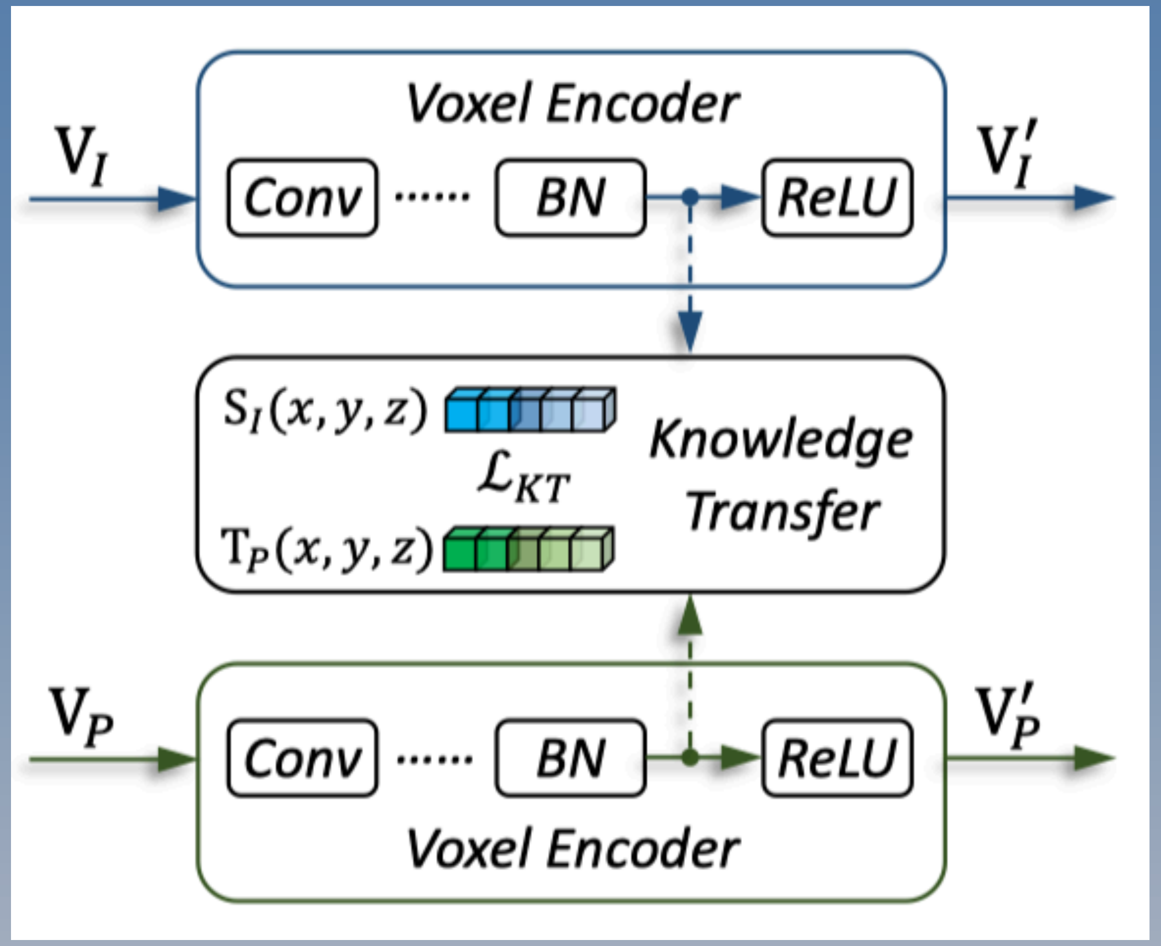
Cross-modality Interaction

- *The cross-modality interaction is proposed from two folds, i.e., transferring geometry-aware knowledge to images and fusing context-aware features with point clouds.*



Design choices

- *Knowledge Transfer: optimize features of the student with guidance from the teacher during training.*



Details in the knowledge transfer.

Feature distance for knowledge transfer

$$d_{KT} = PL_2(\mathbf{T}_P(x, y, z), \mathbf{S}_I(x, y, z))$$

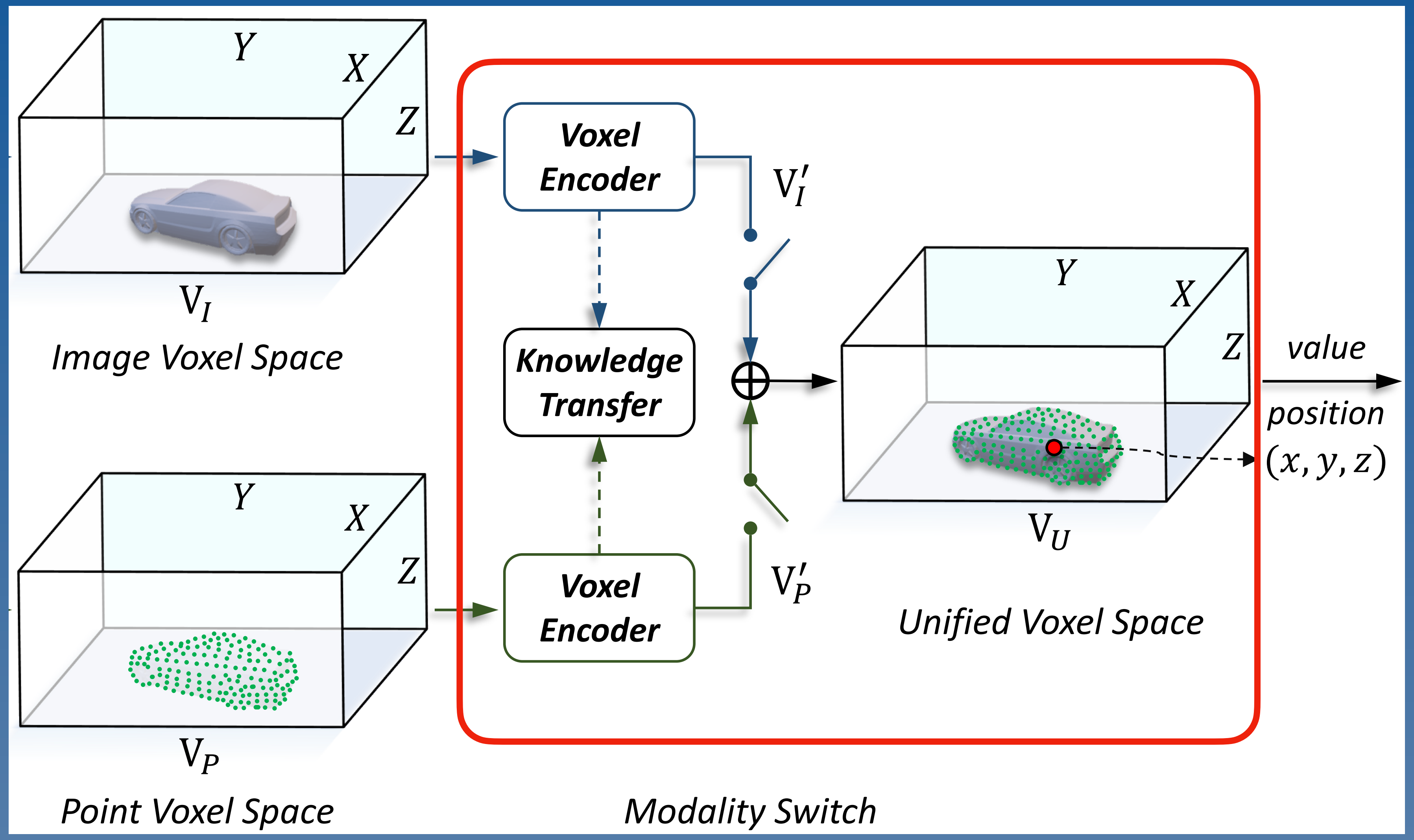
Optimization objective for knowledge transfer

$$\mathcal{L}_{KT} = \frac{1}{N} \sum_i (d_{KT})$$

Unified Representation

Cross-modality Interaction

- *The cross-modality interaction is proposed from two folds, i.e., transferring geometry-aware knowledge to images and fusing context-aware features with point clouds.*



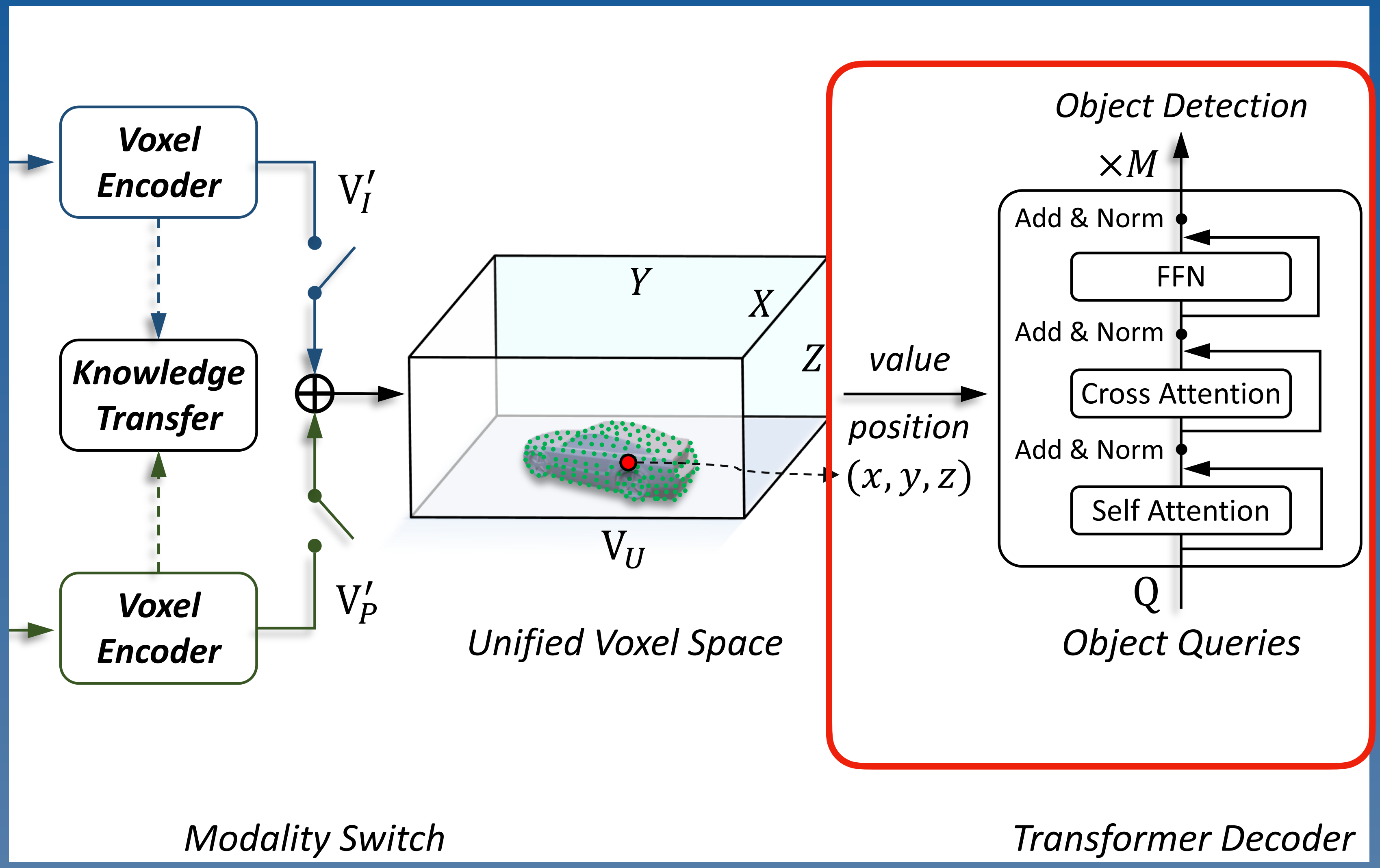
Design choices

- *Knowledge Transfer: optimize features of the student with guidance from the teacher during training.*
- *Modality Fusion: aim to better utilize all modalities in both training and inference stages.*

Unified Representation

Transformer Decoder

- Transformer decoder is utilized for further object-level interaction in the unified voxel space.



Design choices

- Transformer Design: apply reference positions to efficiently sample representative features.
- Deformable Attention: use cross-attention module like that in Deformable DETR.

$$\mathrm{CrossAttn}(q, \mathbf{V}_U(p)) = \mathrm{DeformAttn}(q, p, \mathbf{V}_U)$$

Results & Analysis

Results of UVTR

It surpasses previous multi-modality methods and improves consistently.

Comparisons on different methods with a single model on the nuScenes val set.

Method	Backbone	**NDS(%)**	**mAP(%)**	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
LiDAR-based								
CenterPoint† [24]	V0.1	64.9	56.6	0.291	0.252	0.324	0.284	0.189
UVTR-L	V0.1	66.4	59.3	0.345	0.259	0.313	0.218	0.185
UVTR-L	V0.075	**67.7**	**60.9**	0.334	0.257	0.300	0.204	0.182
Camera-based								
DETR3D [8]	R101	42.5	34.6	0.773	0.268	0.383	0.842	0.216
UVTR-C	R50	41.9	33.3	0.793	0.276	0.454	0.760	0.196
UVTR-C	R101	44.1	36.2	0.758	0.272	0.410	0.758	0.203
UVTR-CS	R50	47.2	36.2	0.756	0.276	0.399	0.467	0.189
UVTR-CS	R101	48.3	37.9	0.731	0.267	0.350	0.510	0.200
UVTR-L2C	R101	45.0	37.2	0.735	0.269	0.397	0.761	0.193
UVTR-L2CS	R101	**48.8**	**39.2**	0.720	0.268	0.354	0.534	0.206
LiDAR+Camera								
FUTR3D [9]	V0.075-R101	68.3	64.5	-	-	-	-	-
UVTR-M	V0.075-R101	**70.2**	**65.4**	0.332	0.258	0.268	0.212	0.177

Results & Analysis

Results of UVTR

It surpasses previous multi-modality methods and improves consistently.

Comparisons on different methods with a single model on the nuScenes test set.

Method	Backbone	NDS(%)	mAP(%)	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
			LiDAR-based					
3DSSD [45]	Point-based	56.4	42.6	-	-	-	-	-
CenterPoint [24]	V0.075	65.5	58.0	-	-	-	-	-
HotSpotNet [46]	V0.1	66.0	59.3	0.274	0.239	0.384	0.333	0.133
AFDetV2 [47]	V0.075	68.5	62.4	0.257	0.234	0.341	0.299	0.137
UVTR-L	V0.075	**69.7**	**63.9**	0.302	0.246	0.350	0.207	0.123
			Camera-based					
FCOS3D [27]	R101	42.8	35.8	0.690	0.249	0.452	1.434	0.124
DD3D [48]	V2-99	47.7	41.8	0.572	0.249	0.368	1.014	0.124
DETR3D [8]	V2-99	47.9	41.2	0.641	0.255	0.394	0.845	0.133
BEVDet [6]	V2-99	48.8	42.4	0.524	0.242	0.373	0.950	0.148
PETR [10]	V2-99	50.4	44.1	0.593	0.249	0.383	0.808	0.132
UVTR-L2C	V2-99	52.2	45.2	0.612	0.256	0.385	0.664	0.125
UVTR-L2CS3	V2-99	**55.1**	**47.2**	0.577	0.253	0.391	0.508	0.123
			LiDAR+Camera					
FusionPainting [49]	V0.075-R50	70.4	66.3	-	-	-	-	-
MVP [32]	V0.075-DLA34	70.5	66.4	-	-	-	-	-
PointAugmenting [50]	V0.075-DLA34	71.0	66.8	-	-	-	-	-
UVTR-M	V0.075-R101	**71.1**	**67.1**	0.306	0.245	0.351	0.225	0.124

Results & Analysis

Framework analysis

UVTR achieves robust results with dropped views and sensor noises.

Different heights Z in voxel space.

modality	*height*	NDS(%)	mAP(%)
Camera	1	31.4	24.9
	5	34.5	27.0
	11	**35.6**	**28.7**
LiDAR	1	62.8	54.4
	5	63.8	55.5
	11	**63.8**	**56.3**

Different operations in voxel encoder.

modality	*type*	NDS(%)	mAP(%)
Camera	None	12.0	2.5
	Conv2D	31.9	24.8
	Conv3D	**34.5**	**27.0**
LiDAR	None	63.1	54.3
	Conv2D	63.2	54.6
	Conv3D	**63.8**	**55.5**

Different knowledge transfer settings.

student	*teacher*	NDS(%)	mAP(%)
Camera	–	34.5	27.0
	CS	36.3	28.1
	LiDAR	36.4	28.2
	Multi-mod	**37.1**	**28.8**
LiDAR	–	63.8	55.5
	Multi-mod	**64.4**	**56.1**

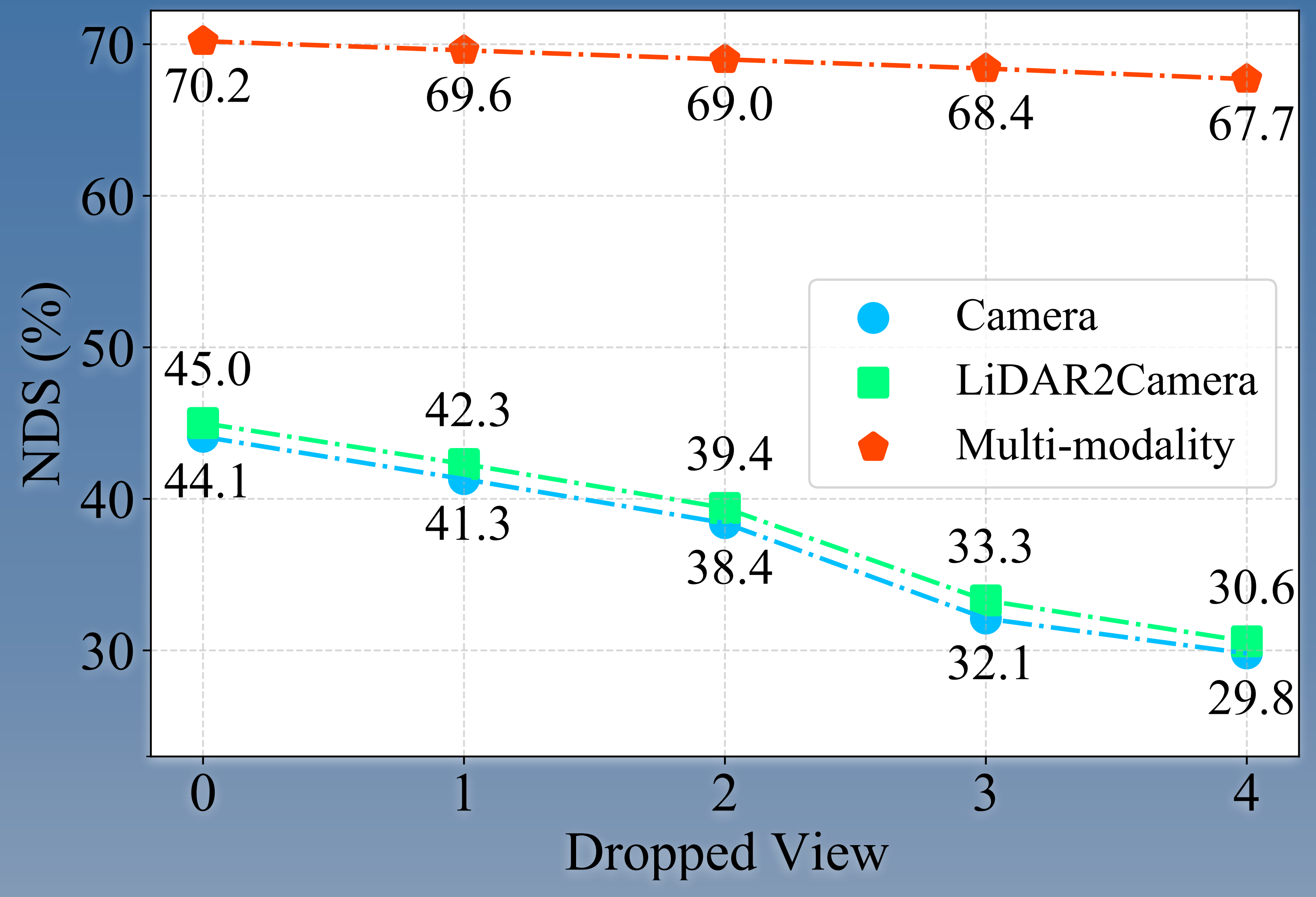
Different cross-modality fusion settings.

camera	*lidar*	NDS(%)	mAP(%)
R50	–	34.5	27.0
–	V0.1	63.8	55.5
R50	V0.1	65.1	59.0
	V0.075	**65.6**	**60.1**
R101	V0.1	65.4	59.4
	V0.075	**66.3**	**61.0**

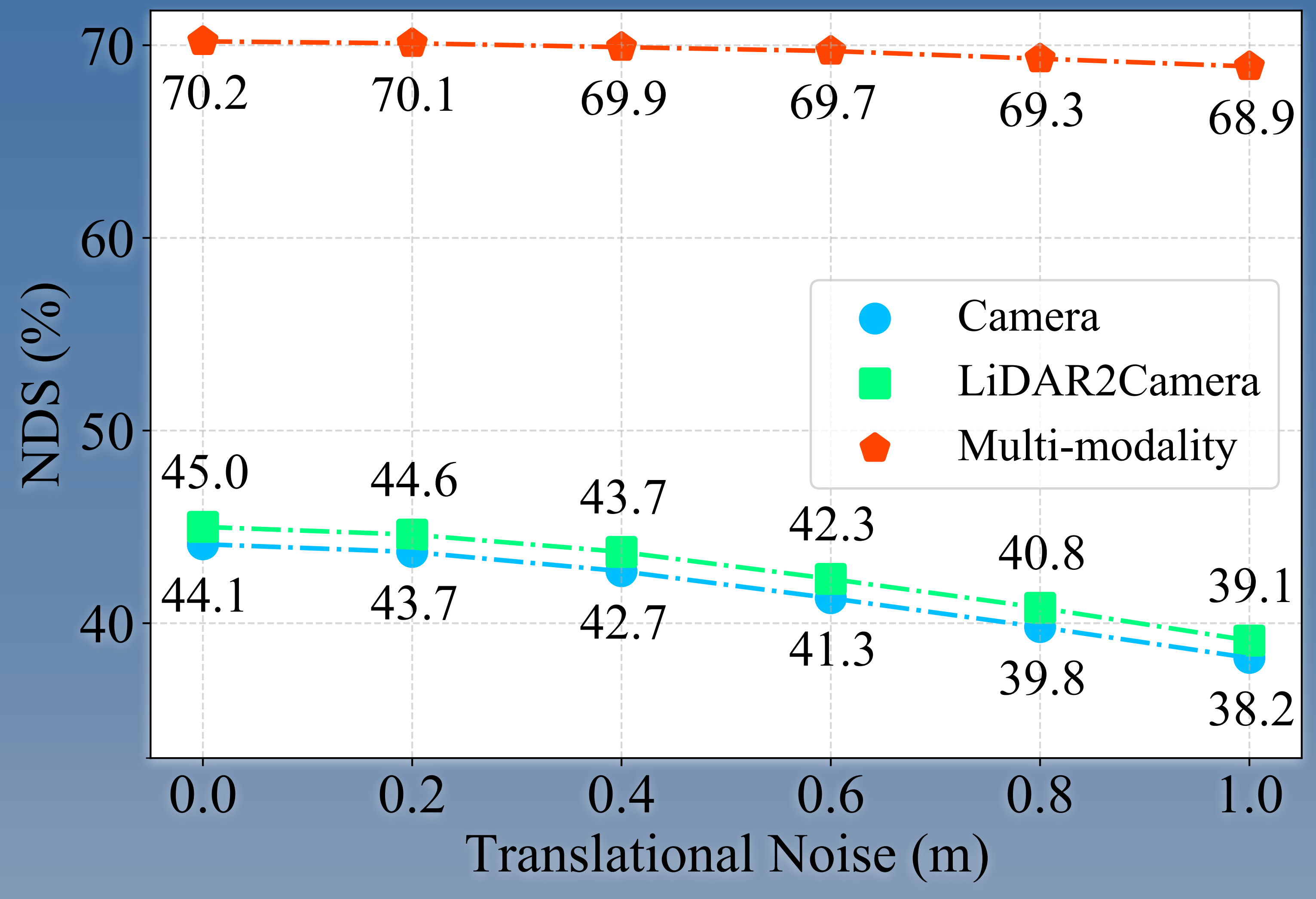
Results & Analysis

Framework analysis

UVTR achieves robust results with dropped views and sensor noises.



Robustness of dropped camera view.



Robustness of sensor calibration noise.

Future Work

Current multi-modality frameworks still exists several problems that can be solved in the future work:

1. *Reduce computation cost: current camera-based approaches process all of them in the shared image backbone, which brings computational cost, especially for multi-frame setting.*
2. *Unified framework extension: current multi-modality frameworks mainly focuses on object detection, which can be extended to support following tasks, like segmentation, tracking, and planning.*
3. *Open-world and long tail: current work mainly focus on predefined vehicles, ignoring numerous long-tail instances in real world scenes, like unseen objects in training set.*

Q&A

For more questions, please contact

www.yanwei-li.com
ywli@cse.cuhk.edu.hk

Slides