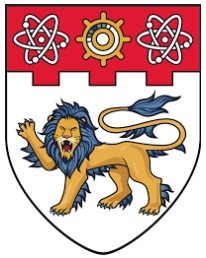


ViKIENet: Towards Efficient 3D Object Detection with Virtual Key Instance Enhanced Network

Zhuochen Yu[1*] Bijie Qiu[1*†] Andy W. H. Khong[1,2]

Nanyang Technological University, Singapore

[1]School of Electrical and Electronic Engineering, [2]Lee Kong Chian School of Medicine

June 14, 2025

Background and Motivation

Background:

3D Object Detection is critical for autonomous driving, robotics, and intelligent systems.

LiDAR-based detection provides accurate depth but suffers from:

1. Sparse point clouds, especially for distant or small objects.
2. Limited semantic information.

RGB images offer rich semantic content and dense structure, but lack depth information.

Motivation:

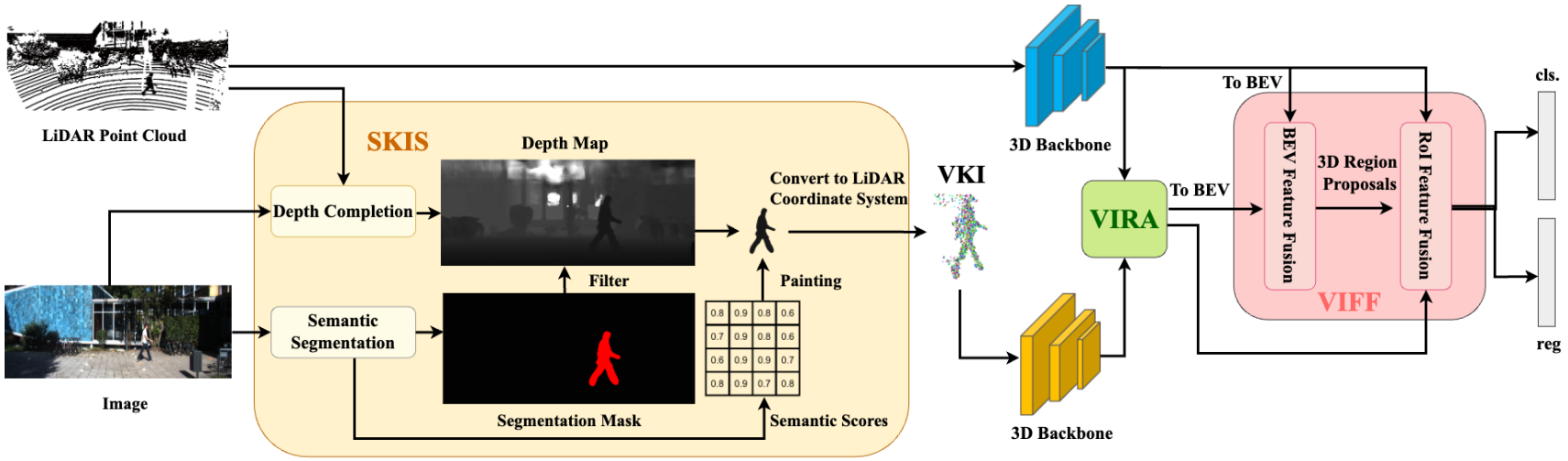
Existing virtual-point-based fusion methods face three key challenges:

1. **High computational cost**
2. **Noisy depth completion**
3. **Insufficient semantic utilization**

Our Method-ViKIENet

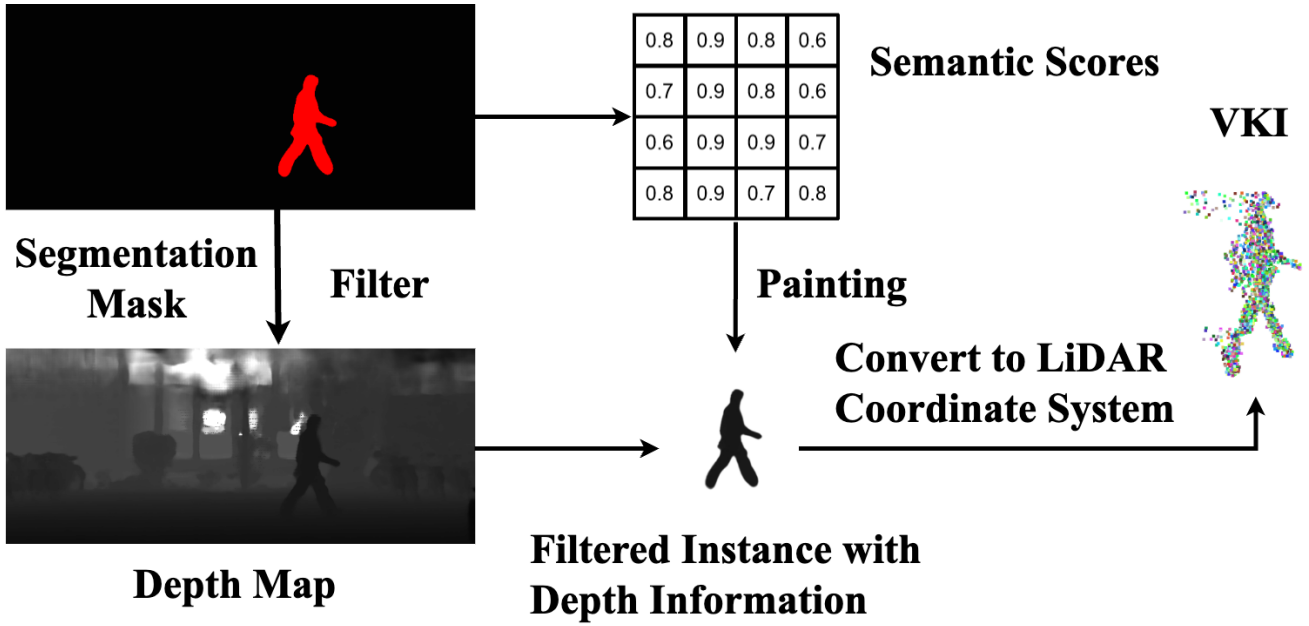
We propose a multi-modal fusion framework for 3D object detection:

- **SKIS**: Semantic Key Instance Selection.
- **VIFF:** Virtual-Instance-Focused Fusion
- **VIRA:** Virtual-to-Real Attention
- We further introduce **ViKIENet-R** with **VIFF-R**, incorporating rotationally equivariant features.



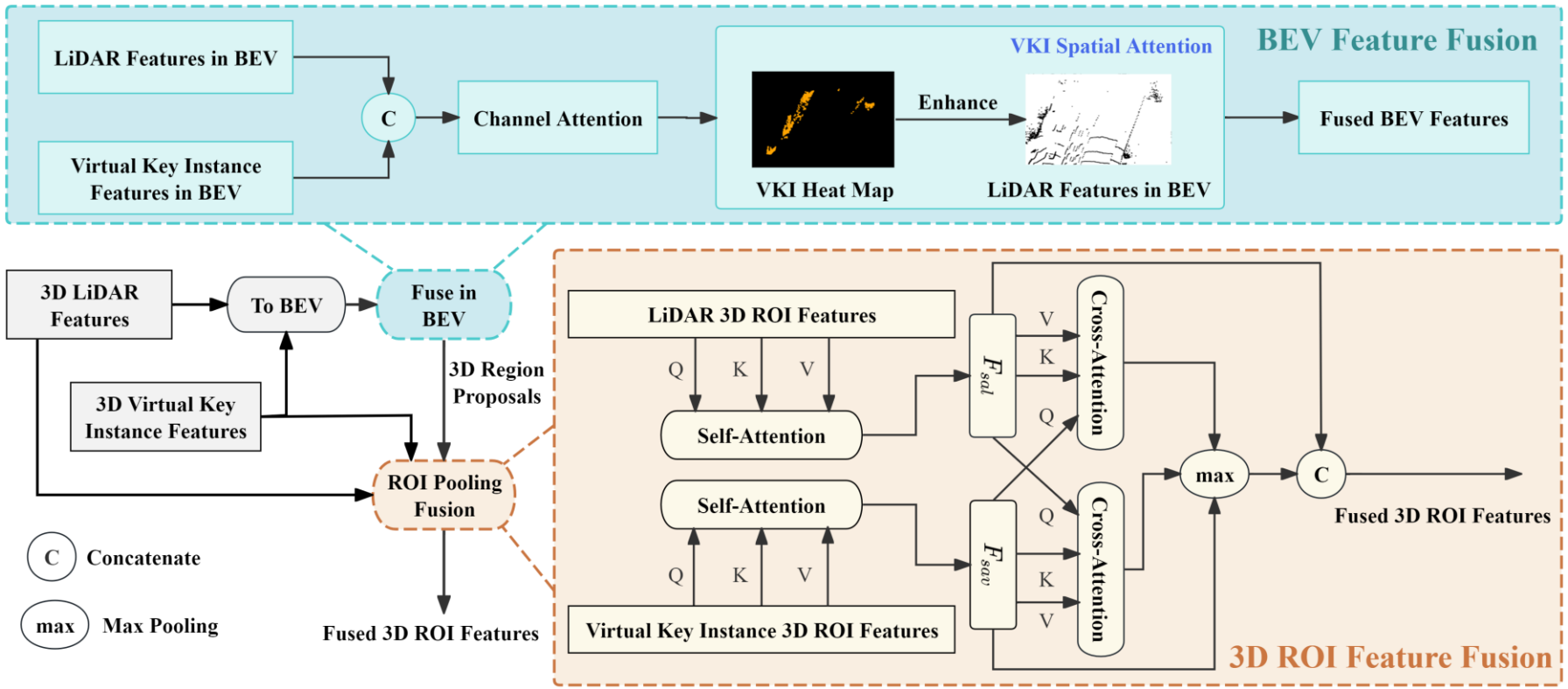
The proposed ViKIENet architecture with SKIS, VIRA, and VIFF.

SKIS: Semantic Key Instance Selection



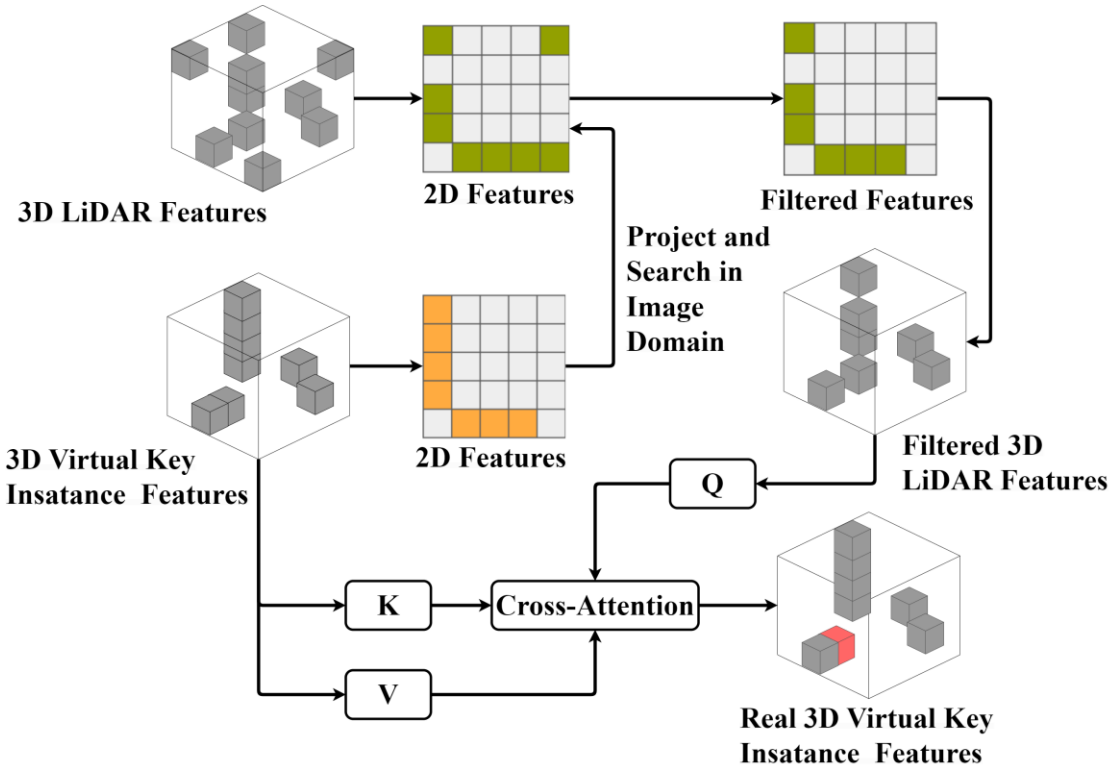
The SKIS module, where semantic segmentation is performed on the original input image to obtain a segmentation mask from which the segmentation scores are derived for each pixel. The segmentation mask is used to extract VKIs while the segmentation scores are employed as semantic information to enrich the VKIs.

VIFF: Virtual-Instance-Focused Fusion



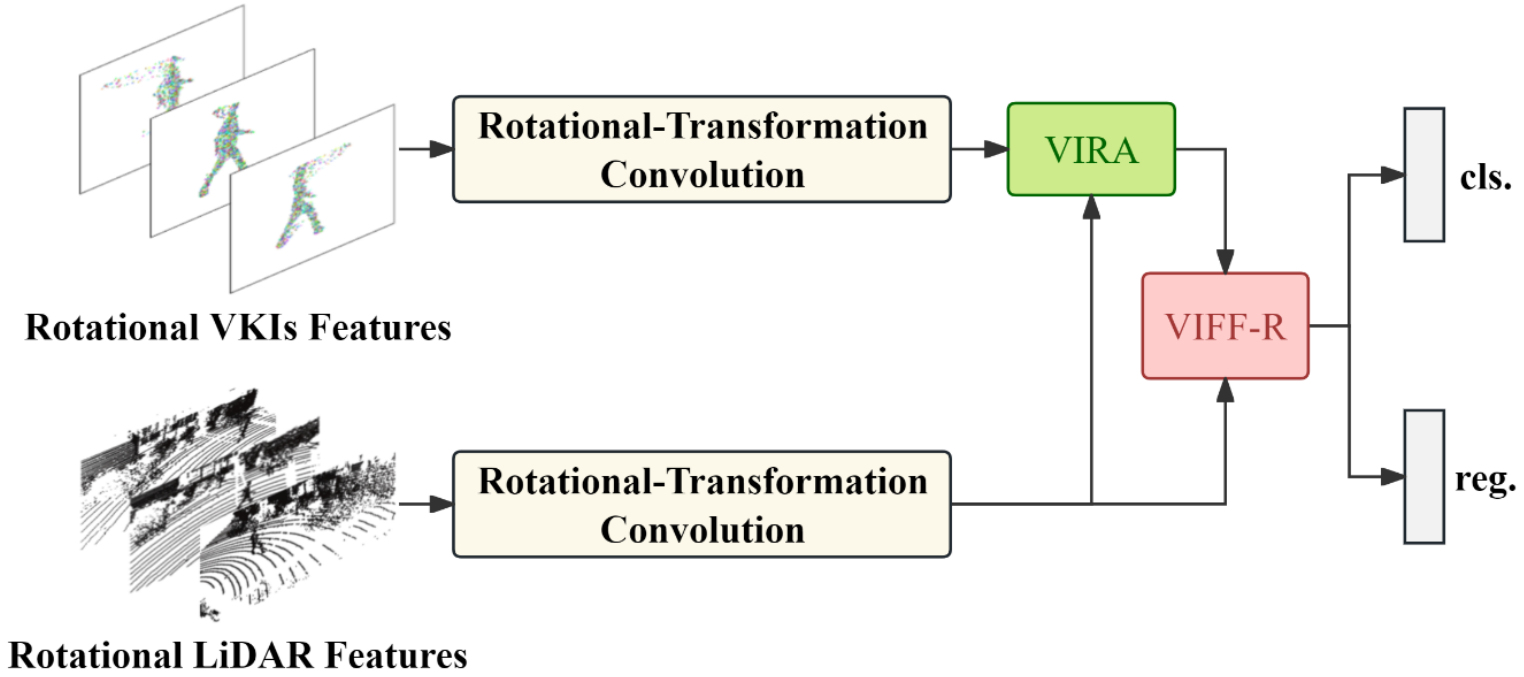
The VIFF module with BEV feature fusion and RoI feature fusion are shown in blue and yellow, respectively. The 3D region proposals from BEV will be subjected to RoI pooling within the 3D features to refine the proposals.

VIRA: Virtual-to-Real Attention



VIRA leverages the precise depth of LiDAR points to form more accurate feature representations of the VKIs

ViKIENet-R



The structure of ViKIENet-R.

Experiment

Method	BEV (R40)			3D AP (R40)		
	Easy	Mod.	Hard	Easy	Mod.	Hard
PV-RCNN [24]	95.76	91.11	88.93	92.57	84.83	82.69
Voxel-RCNN [3]	95.68	91.25	88.95	92.38	85.29	82.86
LoGoNet [13]	-	-	-	92.04	85.04	84.31
Focals Conv [2]	95.45	91.51	91.21	92.86	85.85	85.29
SFD [33]	-	-	-	95.47	**88.56**	**85.74**
VirConv-L [32]	-	-	-	93.18	88.23	85.48
ViKIENet (Ours)	**96.31**	**93.73**	**91.64**	**95.58**	88.49	**86.07**
VirConv-T [32]	**96.58**	93.35	91.25	94.58	**89.87**	**87.78**
ViKIENet-R (Ours)	95.45	**93.58**	**91.33**	**94.63**	89.52	87.76

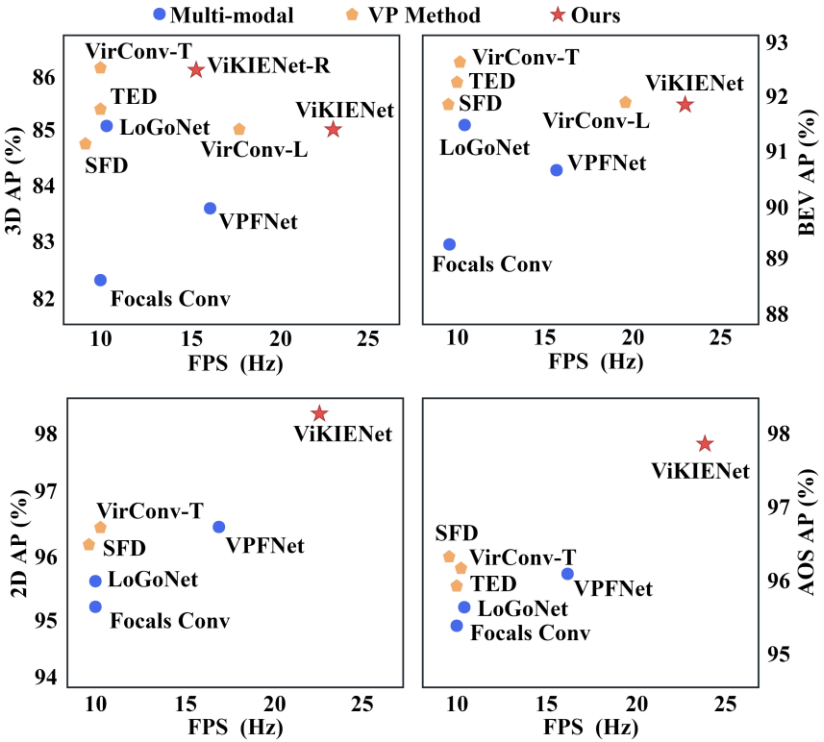
Table 1. Comparison with state-of-the-art methods on the KITTI val set for car 3D detection.

Method	Modality	Car 3D AP (R40)			FPS (Hz)
		Easy	Mod.	Hard	
PV-RCNN [24]	LiDAR	90.25	81.43	76.82	12.5*
Voxel-RCNN [3]	LiDAR	90.90	81.62	77.06	25.2
BtcDet [35]	LiDAR	90.64	82.86	78.09	11.1
Focals Conv [2]	LiDAR+RGB	90.55	82.28	77.59	10.0*
VPFNet [46]	LiDAR+RGB	91.02	83.21	78.20	16.13
LoGoNet [13]	LiDAR+RGB	91.80	85.06	80.74	10.7
SFD [33]	LiDAR+RGB	91.73	84.76	77.92	10.2
VirConv-L [32]	LiDAR+RGB	91.41	85.05	80.22	17.9
ViKIENet (Ours)	LiDAR+RGB	91.79	84.96	80.20	**22.7**
TSSTDet [8]	LiDAR	91.84	85.47	80.65	10.5
TED [31]	LiDAR+RGB	91.61	85.28	80.68	10.6
VirConv-T [32]	LiDAR+RGB	92.54	86.25	81.24	10.7
ViKIENet-R (Ours)	LiDAR+RGB	91.20	86.04	81.18	**15.0**

Table 2. 3D detection results for the car class on the KITTI test set.

VIFF		VIRA	AP (R40)		
BEV Fusion	RoI Fusion		Easy	Moderate	Hard
No	No	No	93.14	85.41	83.20
Yes	No	No	93.47	85.63	84.71
No	Yes	No	92.89	87.75	85.62
Yes	Yes	No	95.33	88.09	85.67
Yes	Yes	Yes	95.58	88.49	86.07

Table 3. Ablation study results of VIFF and VIRA on the KITTI validation set.



The proposed ViKIENet and ViKIENet-R accelerate the virtual-point-based methods while enhancing performance.