Qualcomm

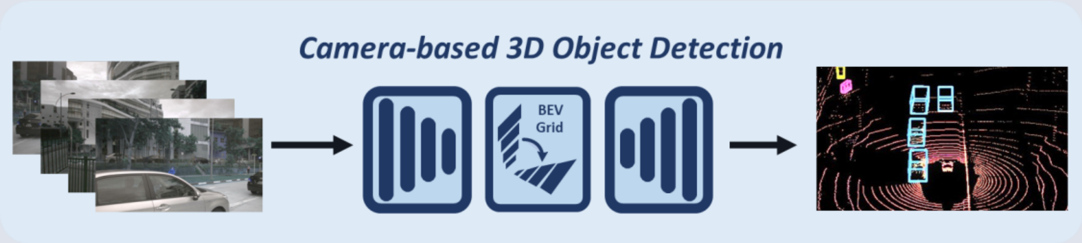
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

X^{3}KD: Knowledge Distillation Across Modalities, Tasks and Stages for Multi-Camera 3D Object Detection

Marvin Klingner* Shubhankar Borse* Varun Ravi Kumar*
Behnaz Rezaei Venkatraman Narayanan Senthil Yogamani Fatih Porikli

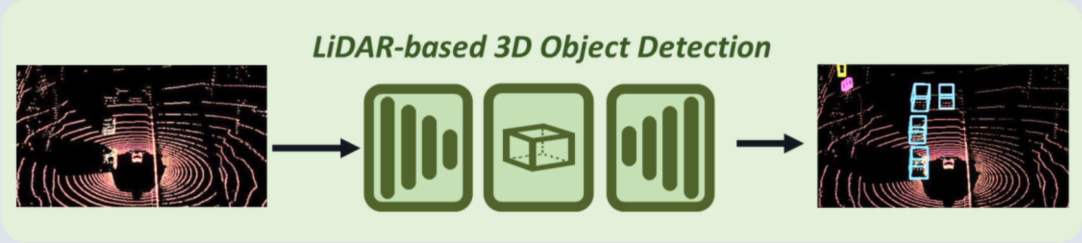
WED-PM-094

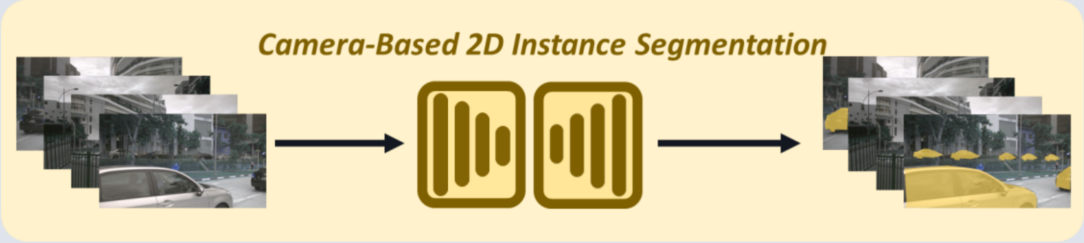
Motivation



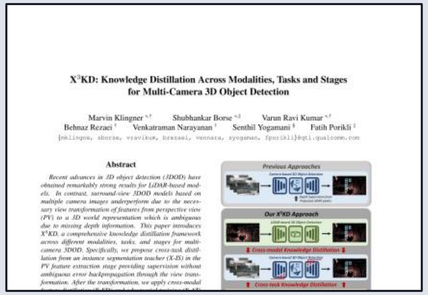
LiDAR-based 3D object detection outperform multi-camera 3D object detection methods

Pretraining of the feature extraction on 2D instance segmentation improves 3D object detection





Main Contribution

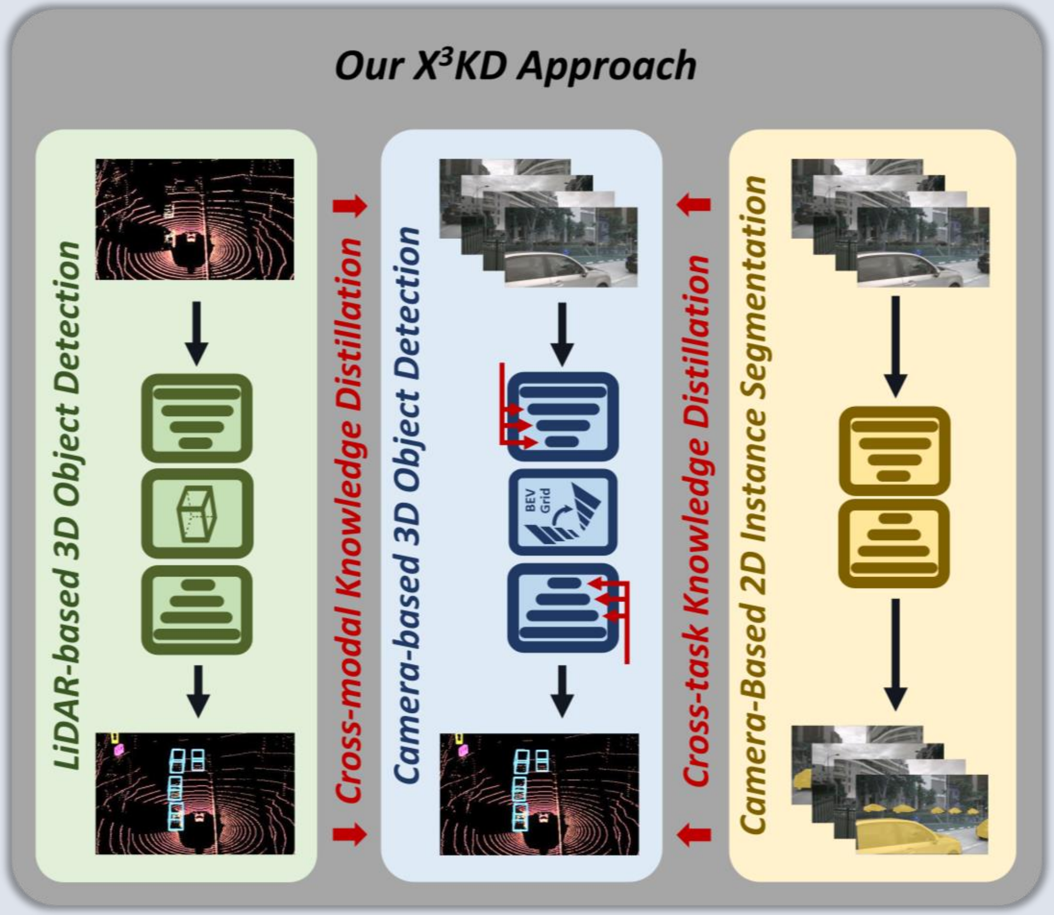


X^3KD: Knowledge Distillation Across Modalities, Tasks and Stages for Multi-Camera 3D Object Detection

Marvin Klingner, Shubhankar Borse, Varun Ravi Kumar, Behnaz Rezaei, Venkatraman Narayanan, Senthil Yogamani, Fatih Porikli

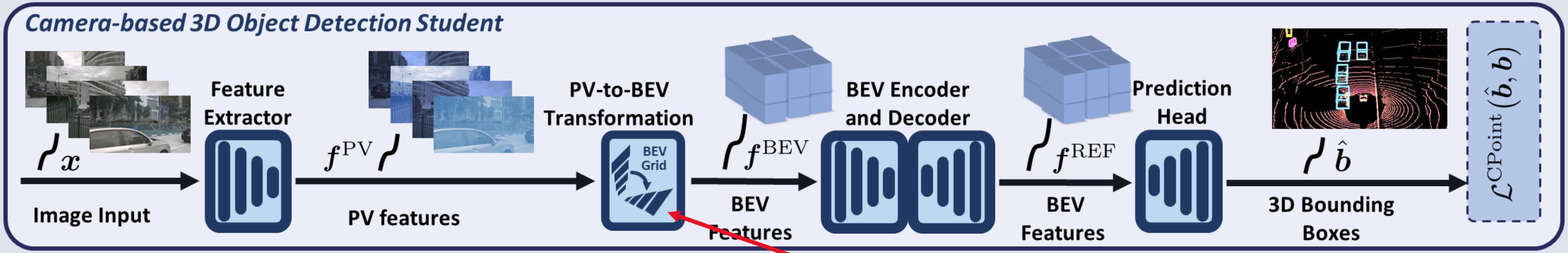
Abstract

Recent advances in 3D object detection (3DOD) have obtained remarkably strong results for LiDAR-based models. In contrast, surround-view 3DOD models based on multiple camera images underperform due to the necessary view transformation of features from perspective view (PV) to a 3D world representation which is ambiguous due to missing depth information. This paper introduces X^3KD, a comprehensive knowledge distillation framework across different modalities, tasks, and stages for multi-camera 3DOD. Specifically, we propose cross-task distillation from an instance segmentation teacher (X-IS) in the PV feature extraction stage providing supervision without ambiguous error backpropagation through the view transformation. After the transformation, we apply cross-modal



We improve multi-camera 3D object detection using knowledge distillation from LiDAR-based 3D object detection (3DOD) and instance segmentation teacher models

Baseline Method



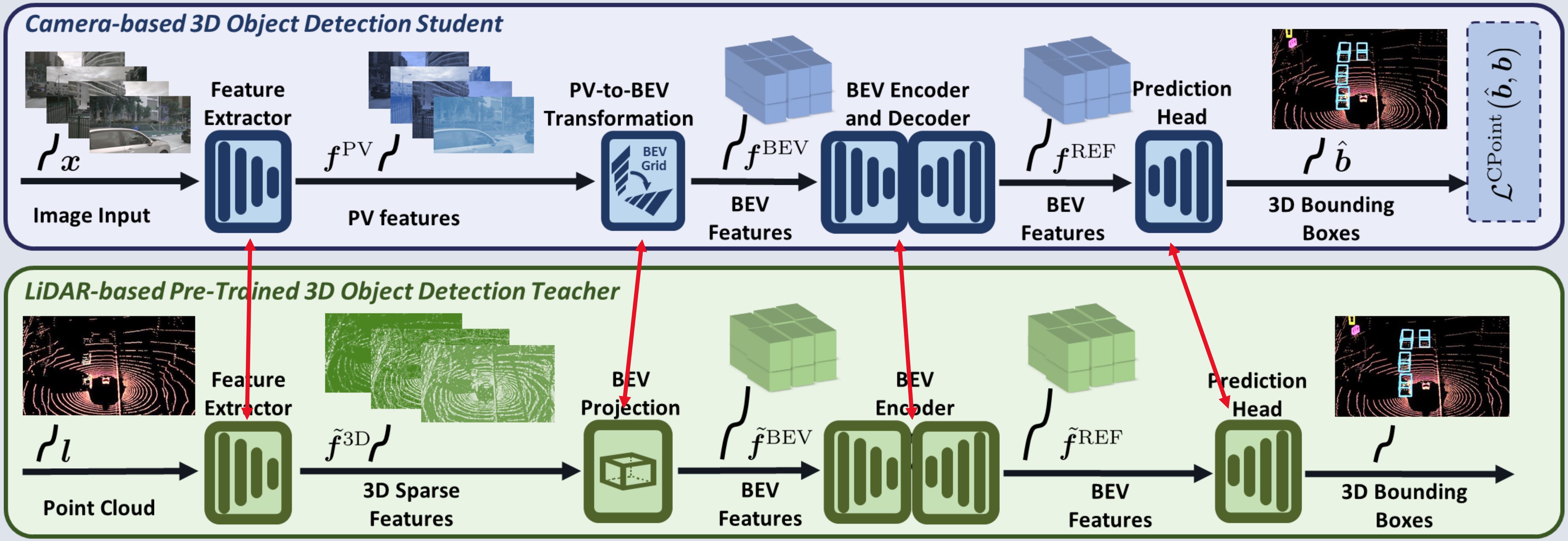
Depth supervision through LiDAR points

- Recent 3DOD approaches learn a bird's eye view (BEV) representation using depth to transform image features to bird's eye view
- Our baseline BEVDepth [1] uses projected LiDAR points to supervise the depth prediction used to transform the features
- In our work we aim at improving the way the LiDAR is used for supervision of the multi-camera 3D object detection

Model	*LSS++*	*DS*	GFLOPS	*mAP*↑	*NDS*↑
BEVDepth†	✗	✗	298	32.4	44.9
	✗	✓	298	33.1	44.9
	✓	✗	316	34.9	47.0
	✓	✓	316	35.9	47.2
X^{3}KD (Ours)	✓	✓	316	**39.0**	**50.5**

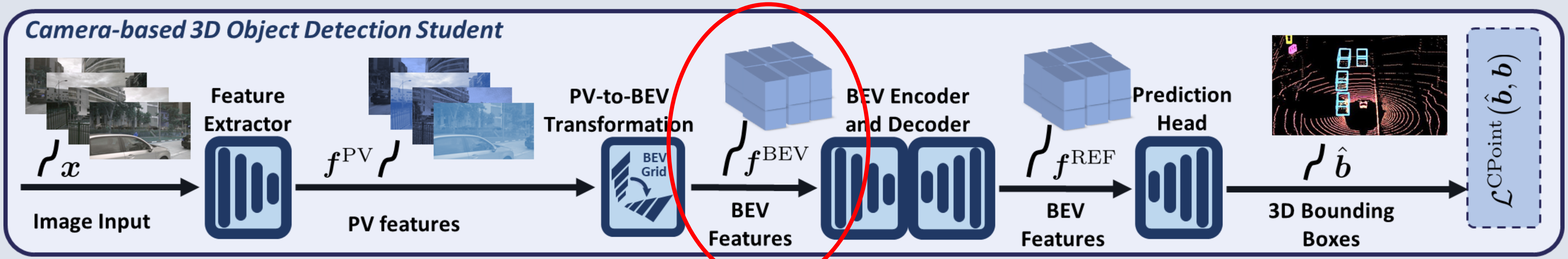
[1] Li et al., "BEVDepth: Acquisition of Reliable Depth for Multi-view 3D Object Detection," in Proc. of AAAI, 2023.

Multi-camera vs. LiDAR-based Models

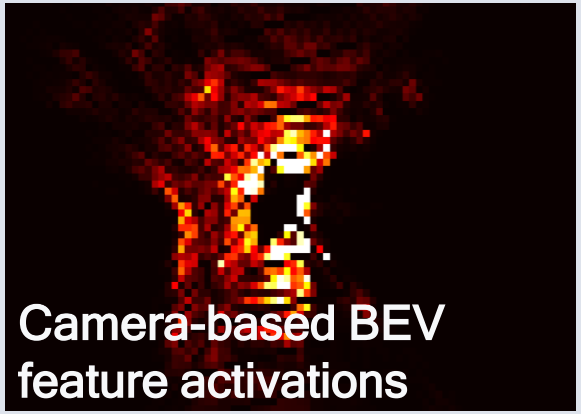


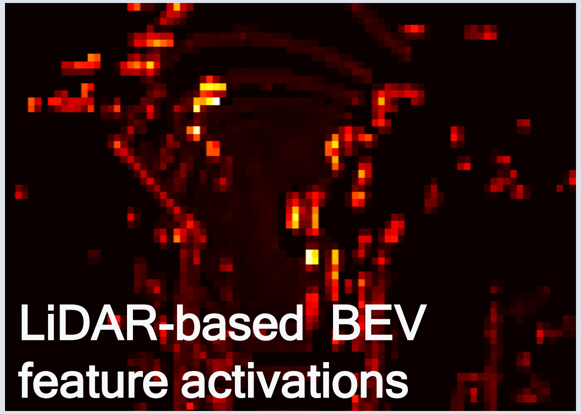
- LiDAR-based 3D object detection clearly outperform multi-camera 3D object detection
- The architectural components of both models are very similar: feature extractor, view transformation, BEV network, prediction head

Transformation to Bird's Eye View

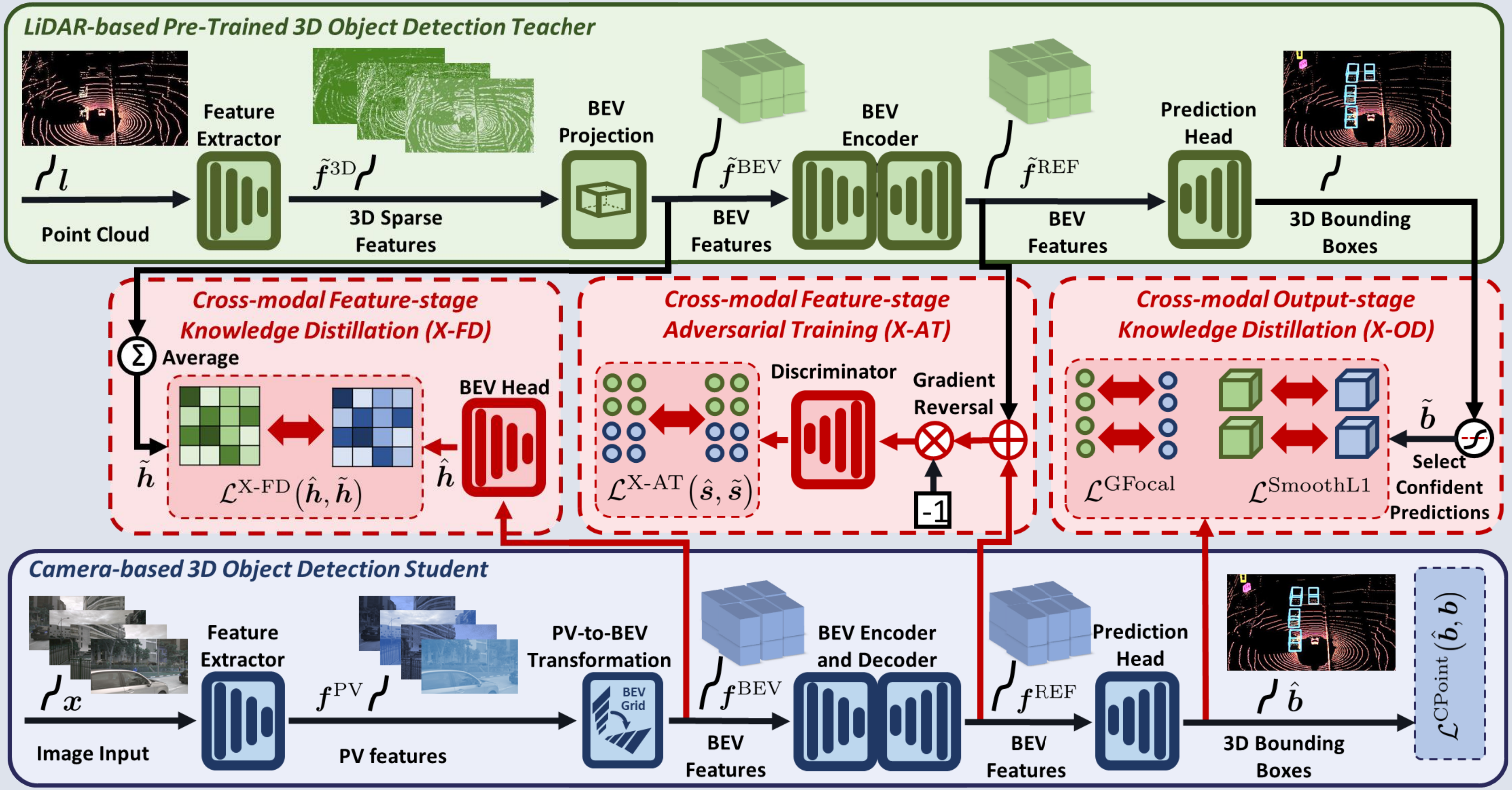


- The multi-camera 3DOD model has a strong focus on the center grid cells and is less precise due to the view transformation based on depth
- The LiDAR-based 3DOD model is able to learn a precise and spatially balanced feature representation in bird's eye view
- We use the learned feature representation of the LiDAR model to distill knowledge into the multi-camera model

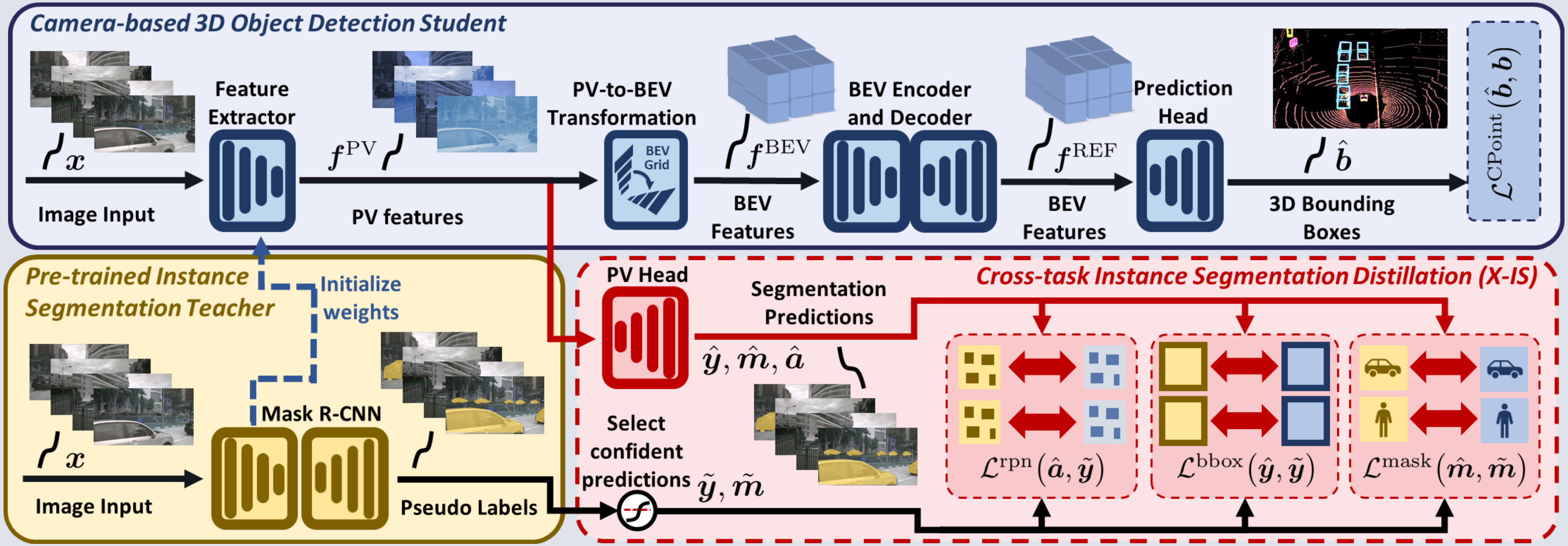




X³KD Method: Cross-modal Distillation



X³KD Method: Cross-task Distillation



- Pretraining of the image feature extraction on instance segmentation improves the multi-camera 3D object detection model's performance
- To retain the knowledge of the image feature extraction, we propose cross-task distillation from a pretrained instance segmentation model

State-of-the-art Comparison

Set	*Model*	*Backbone*	*Resolution*	*mATE*↓	*mASE*↓	*mAOE*↓	*mAVE*↓	*mAAE*↓	*mAP*↑	*NDS*↑
Validation	BEVDet [18]	ResNet-50	256 × 704	0.725	0.279	0.589	0.860	0.245	29.8	37.9
	BEVDet4D [49]			0.703	0.278	0.495	0.354	0.206	32.2	45.7
	BEVDepth [28]			0.629	**0.267**	0.479	0.428	0.198	35.1	47.5
	BEVDepth†			0.636	0.272	0.493	0.499	0.198	35.9	47.2
	STS* [47]			0.601	0.275	0.450	0.446	0.212	37.7	48.9
	BEVStereo* [27]			**0.598**	0.270	**0.438**	0.367	**0.190**	37.2	50.0
	X^3KD_{all}	ResNet-50	256 × 704	0.615	0.269	0.471	**0.345**	0.203	**39.0**	**50.5**
Validation	PETR [32]	ResNet-101	512 × 1408	0.710	0.270	0.490	0.885	0.224	35.7	42.1
	BEVDepth†			0.579	0.265	0.387	0.364	0.194	40.9	53.1
	BEVDepth [28]			0.565	0.266	0.358	0.331	**0.190**	41.2	53.5
	STS* [47]			**0.525**	0.262	0.380	0.369	0.204	43.1	54.2
	X^3KD_{all}	ResNet-101	512 × 1408	0.552	**0.257**	**0.338**	**0.328**	0.199	**44.8**	**55.3**

- We compare our X³KD method against previous state-of-the-art methods on the nuScenes dataset in comparable settings, i.e., same backbone and input image resolution
- We outperform all previous state-of-the-art methods
- We significantly improve over our reimplementation of the baseline BEVDepth [1]

[1] Li et al., "BEVDepth: Acquisition of Reliable Depth for Multi-view 3D Object Detection," in Proc. of AAAI, 2023.

State-of-the-art Comparison

Set	Model	Backbone	Resolution	*mATE*↓	*mASE*↓	*mAOE*↓	*mAVE*↓	*mAAE*↓	*mAP*↑	*NDS*↑
Validation	DETR3D [46]	ResNet-101	900 × 1600	0.716	0.268	0.379	0.842	0.200	34.9	43.4
	BEVFormer [30]			0.673	0.274	0.372	0.394	0.198	41.6	51.7
	PolarFormer [21]			0.648	0.270	0.348	0.409	0.201	43.2	52.8
	BEVDepth†	ResNet-101	640 × 1600	0.571	0.260	0.379	0.374	**0.196**	42.8	53.6
	X^3KD_{all}	ResNet-101	640 × 1600	**0.539**	**0.255**	**0.320**	**0.324**	**0.196**	**46.1**	**56.7**
Test	BEVFormer [30]	ResNet-101	640 × 1600	0.631	0.257	0.405	0.435	0.143	44.5	53.5
	BEVDepth†			0.533	0.254	0.443	0.404	**0.129**	43.1	53.9
	PolarFormer [21]			0.610	0.258	**0.391**	0.458	**0.129**	**45.6**	54.3
	X^3KD_{all}	ResNet-101	640 × 1600	**0.506**	**0.253**	0.414	**0.366**	0.131	**45.6**	**56.1**

- We compare our X^3KD method against previous state-of-the-art methods on the nuScenes dataset in comparable settings, i.e., same backbone and input image resolution
- We outperform all previous state-of-the-art methods
- We significantly improve over our reimplementation of the baseline BEVDepth [1]
- We show the efficacy of our method at various input image resolutions
- We also submit our method's results to the nuScenes benchmark server to evaluate on the non-public test set, where we also outperform previous results

Ablation Studies

Model	X-OD	X-FD	X-AT	X-IS	*mATE*↓	*mASE*↓	*mAOE*↓	*mAVE*↓	*mAAE*↓	*mAP*↑	*NDS*↑
BEVDepth†	✗	✗	✗	✗	0.636	0.272	0.493	0.499	0.198	35.9	47.2
X-OD	✓	✗	✗	✗	0.642	0.278	**0.456**	**0.338**	**0.188**	35.7	48.7
X-FD	✗	✓	✗	✗	0.644	0.276	0.479	0.361	0.200	36.1	48.5
X-AT	✗	✗	✓	✗	0.648	0.277	0.492	0.354	0.192	35.5	48.1
X^{3}KD$_{\text{modal}}$	✓	✓	✓	✗	0.632	0.271	**0.456**	0.342	0.203	36.8	49.4
X-IS	✗	✗	✗	✓	0.635	0.273	0.462	0.350	0.204	38.7	50.1
X^{3}KD$_{\text{all}}$	✓	✓	✓	✓	**0.615**	**0.269**	0.471	0.345	0.203	**39.0**	**50.5**
LiDAR Teacher	NA	NA	NA	NA	0.301	0.257	0.298	0.256	0.195	59.0	66.4

- We show the effectiveness of our single contributions by an ablation study
- Each method component individually improves the 3DOD model’s performance

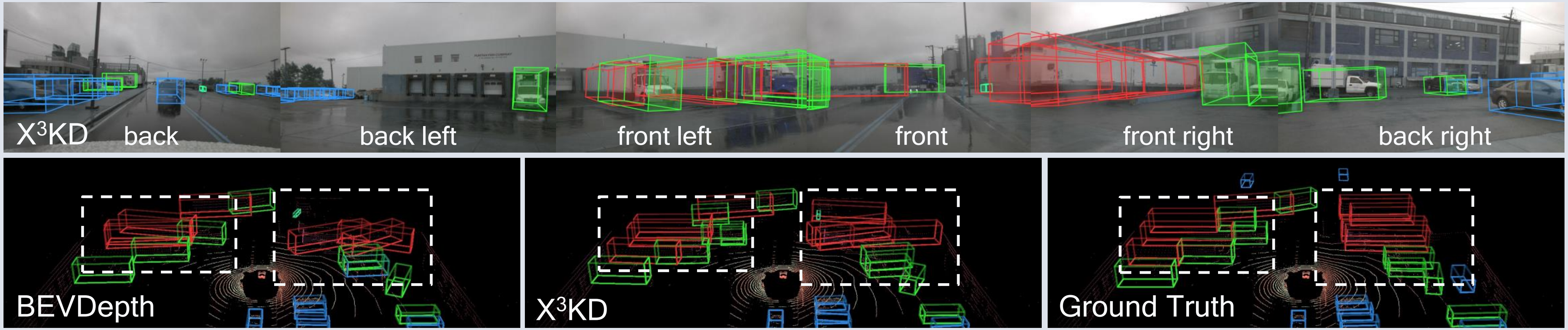
Ablation Studies

Model	*Student Backbone*	*Teacher Backbone*	*Pre.*	*Dist.*	*mAP↑*	*NDS↑*
BEVDepth†	ResNet-50	NA	✗	✗	35.9	47.2
	ResNet-50	ResNet-50	✗	✓	36.4	48.8
	ResNet-50	NA	✓	✗	37.7	49.5
X-IS	ResNet-50	ResNet-50	✓	✓	**38.7**	**50.1**
X-IS	ResNet-50	ConvNeXt-T	✓	✓	38.5	49.9

Model	*Dist.*	*Weight*	*w/o GT*	*mAOE↓*	*mAVE↓*	*mAP↑*	*NDS↑*
BEVDepth†	✗	✗	✗	0.493	0.499	**35.9**	47.2
	✓	✗	✗	0.477	0.342	35.6	48.5
X-OD	✓	✓	✗	**0.456**	**0.338**	35.7	**48.7**
	✓	✗	✓	1.090	0.972	36.1	35.3
X-OD$_{\text{w/o GT}}$	✓	✓	✓	**0.724**	**0.570**	**36.5**	**43.7**

- We also provide ablations on different variants of our method components
- For cross-task distillation (X-IS), we show the effect of instance segmentation pretraining (Pre.) and distillation during 3DOD training (Dist.)
- We also show that student and teacher backbones do not have to be identical
- For cross-modal output-level distillation (X-OD), we show the effect of confidence-based weighting (Weight) and not using 3DOD labels but only pseudo labels from the LiDAR-based teacher (w/o GT)

Qualitative Results



- We also provide a qualitative analysis of our X^{3}KD method
- We observe improved classification and detection performance of X^{3}KD compared to the baseline BEVDepth [1]

[1] Li et al., “BEVDepth: Acquisition of Reliable Depth for Multi-view 3D Object Detection,” in Proc. of AAAI, 2023.

Summary

- We introduce a cross-modal knowledge distillation from a LiDAR-based 3DOD model to a multi-camera 3DOD model
- We propose cross-task knowledge distillation from a pretrained instance segmentation model to the feature extraction of a multi-camera 3DOD model
- We provide a detailed ablation study on the effectiveness of our single contributions
- We outperform previous state-of-the-art approaches at no additional complexity during inference

Qualcomm

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Thank you