ETH zürich | Computer Vision and Geometry Lab

ICCV OCT 19-23, 2025 HONOLULU HAWAII

3D-MOOD: Lifting 2D to 3D for Monocular Open-Set Object Detection

Yung-Hsu Yang[1], Luigi Piccinelli[1], Mattia Segu[1], Siyuan Li[1], Rui Huang[1,2],

Yuqian Fu[3], Marc Pollefeys[1,4], Hermann Blum[1,5], Zuria Bauer[1]

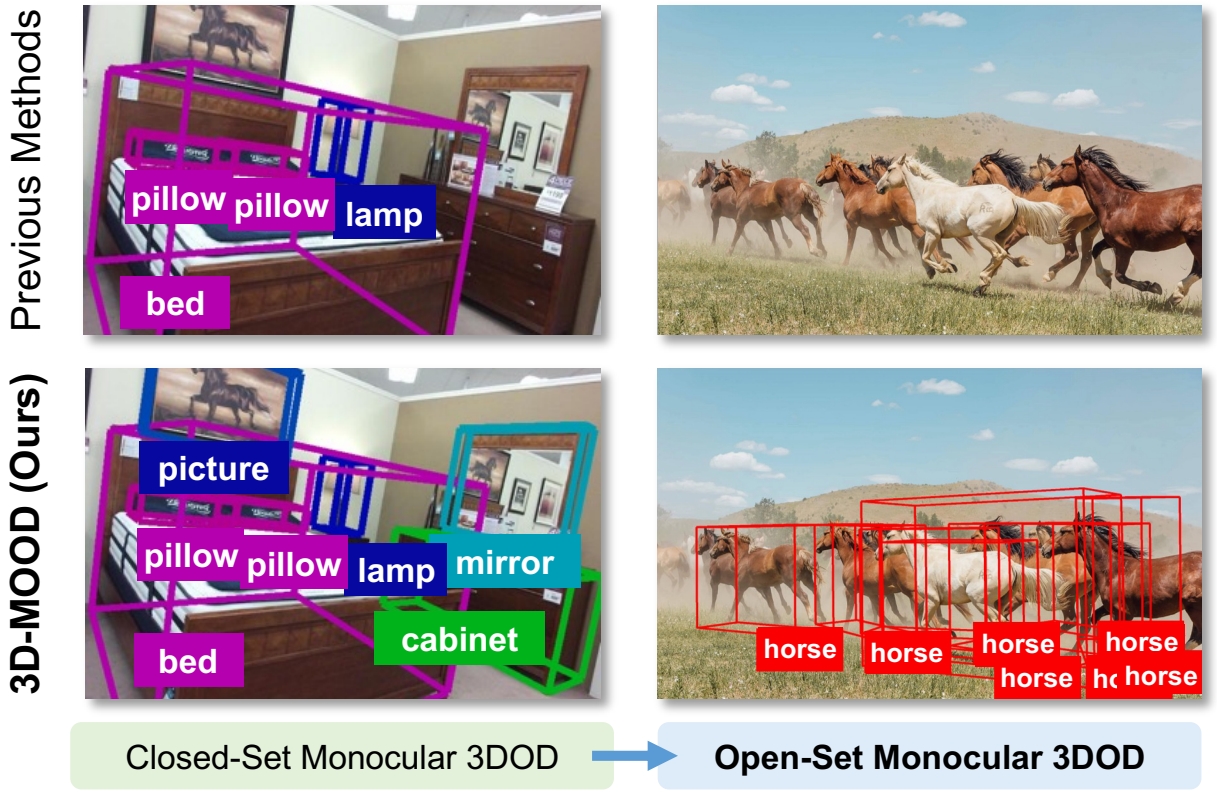
[1]ETH Zürich [2]Tsinghua University [3]INSAIT [4]Microsoft [5]University of Bonn

Project page: royyang0714.github.io/3D-MOOD

Code and models: github.com/cvg/3D-MOOD

Overview

- Classification
  - Open-Vocabulary
  - Language Model
- Regression
  - Open Scenes
  - Metric Depth Estimation



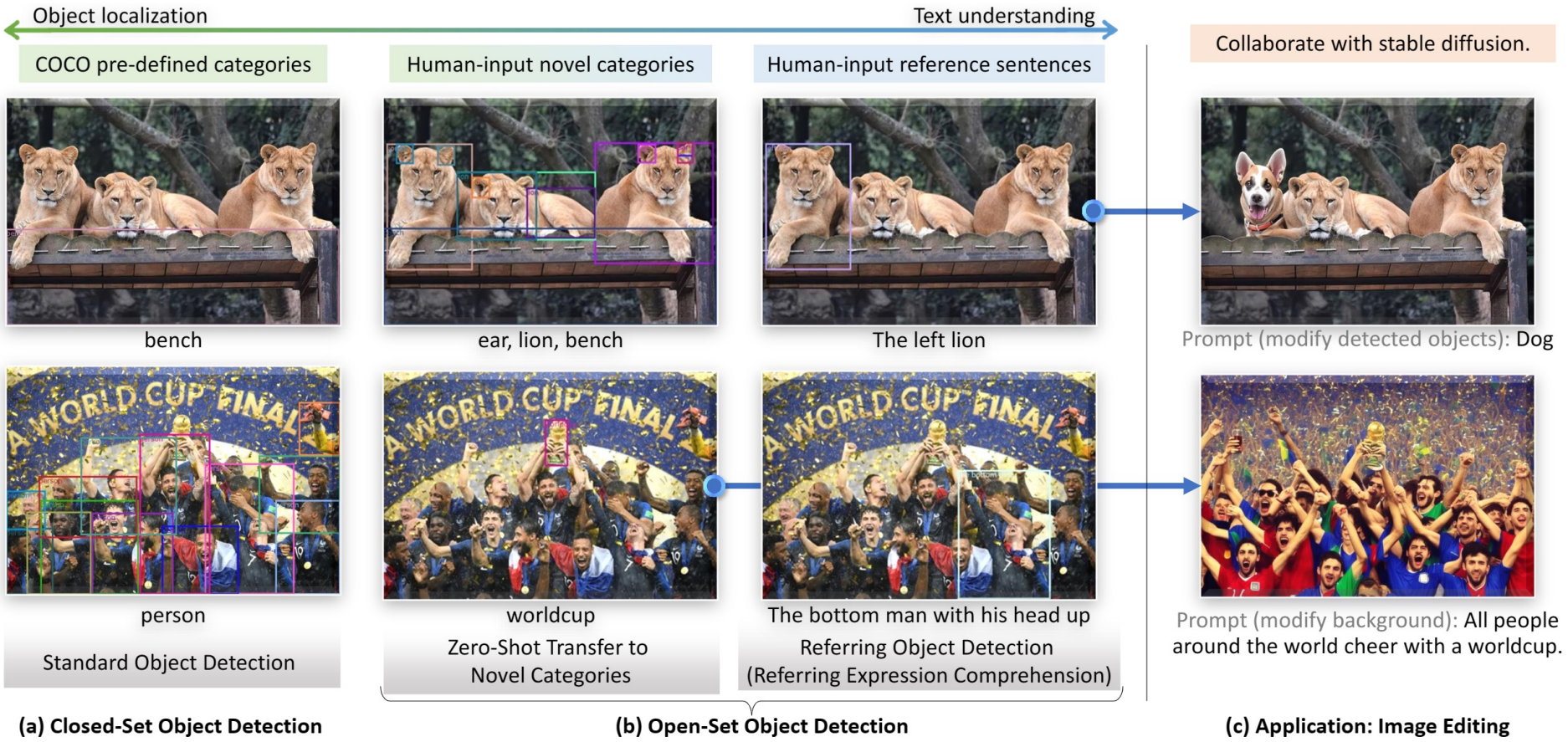
3D Monocular Object Detection

- Fixed Categories
- Cross-dataset Generalization

Trained on	AP_{3D}^{HYP}	AP_{3D}^{SUN}	AP_{3D}^{AR}	Trained on	AP_{3D}^{KIT}	AP_{3D}^{NU}
Hypersim	15.2	9.5	7.5	KITTI	37.1	12.7
SUN	5.8	34.7	13.1	nuScenes	20.2	38.6
ARKit	5.9	14.2	38.6			

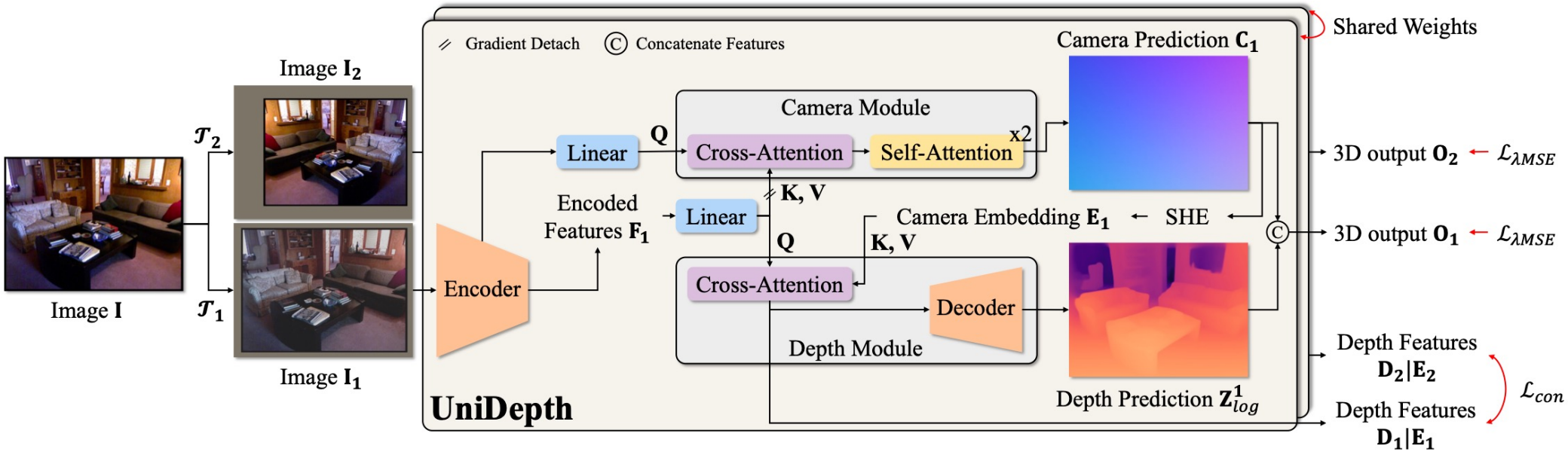
Brazil, Garrick, et al. "Omni3d: A large benchmark and model for 3d object detection in the wild." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

2D Open-set Object Detection



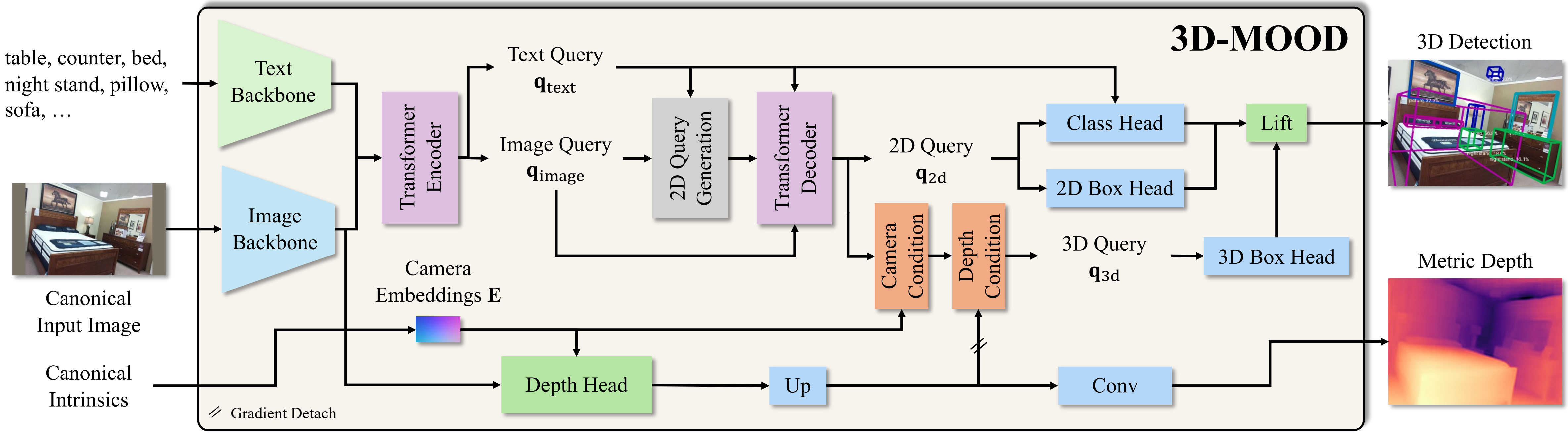
Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." European Conference on Computer Vision 2024.

Metric Monocular Depth Estimation



Piccinelli, Luigi, et al. "UniDepth: Universal monocular metric depth estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

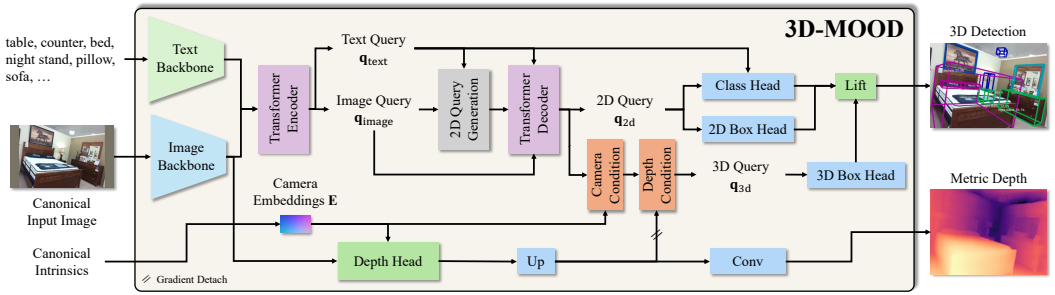
3D-MOOD



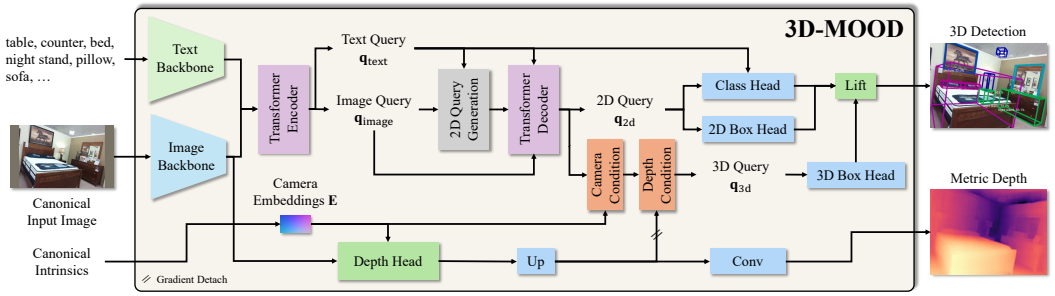
3D-MOOD
table, counter, bed, night stand, pillow, sofa, …
Text Backbone
Image Backbone
Canonical Input Image
Canonical Intrinsics
Transformer Encoder
Text Query $\mathbf{q}_{\text{text}}$
Image Query $\mathbf{q}_{\text{image}}$
2D Query Generation
Transformer Decoder
2D Query $\mathbf{q}_{2d}$
Class Head
2D Box Head
Lift
Camera Condition
Depth Condition
3D Query $\mathbf{q}_{3d}$
3D Box Head
Camera Embeddings $\mathbf{E}$
Depth Head
Up
Conv
Gradient Detach
3D Detection
Metric Depth

3D-MOOD

- End-to-end 2D to 3D Lifting
  - Grounding DINO

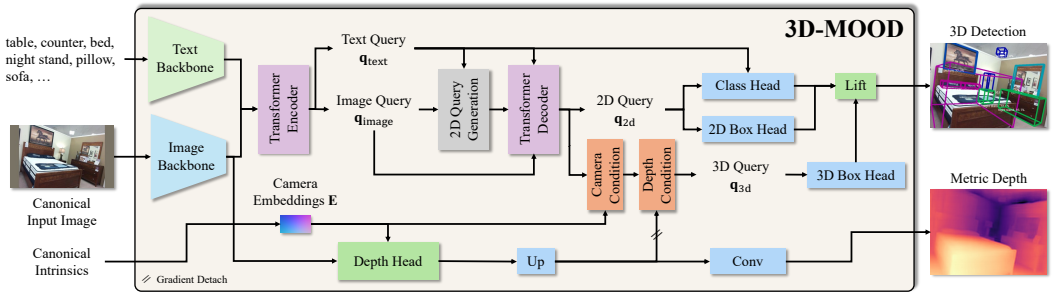


3D-MOOD



- End-to-end 2D to 3D Lifting
  - Grounding DINO
- Geometry-aware 3D Query Generation
  - Camera
  - Depth

3D-MOOD



- End-to-end 2D to 3D Lifting
  - Grounding DINO
- Geometry-aware 3D Query Generation
  - Camera
  - Depth
- Canonical Image Space

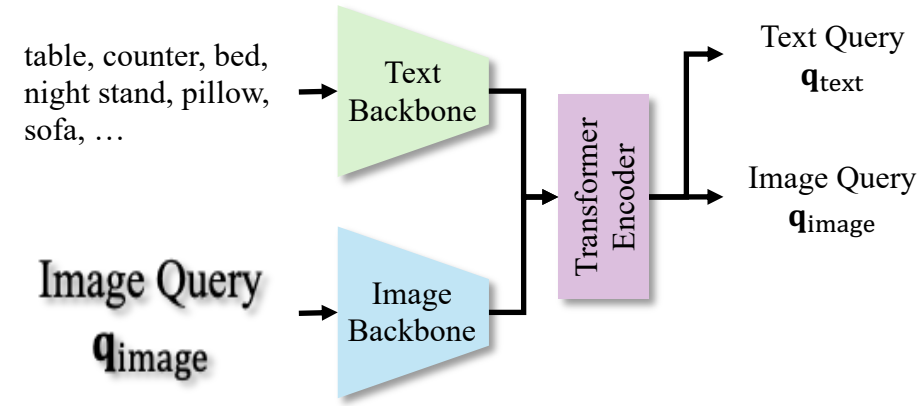
Architecture

table, counter, bed,
night stand, pillow,
sofa, …

Canonical Input Image

Canonical Intrinsics

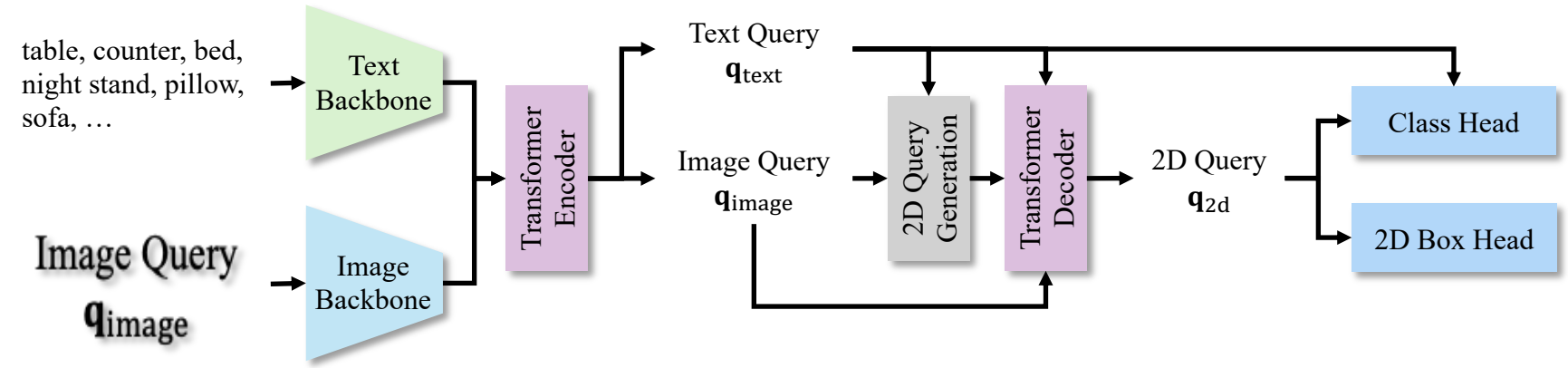
Architecture



Canonical Input Image

Canonical Intrinsics

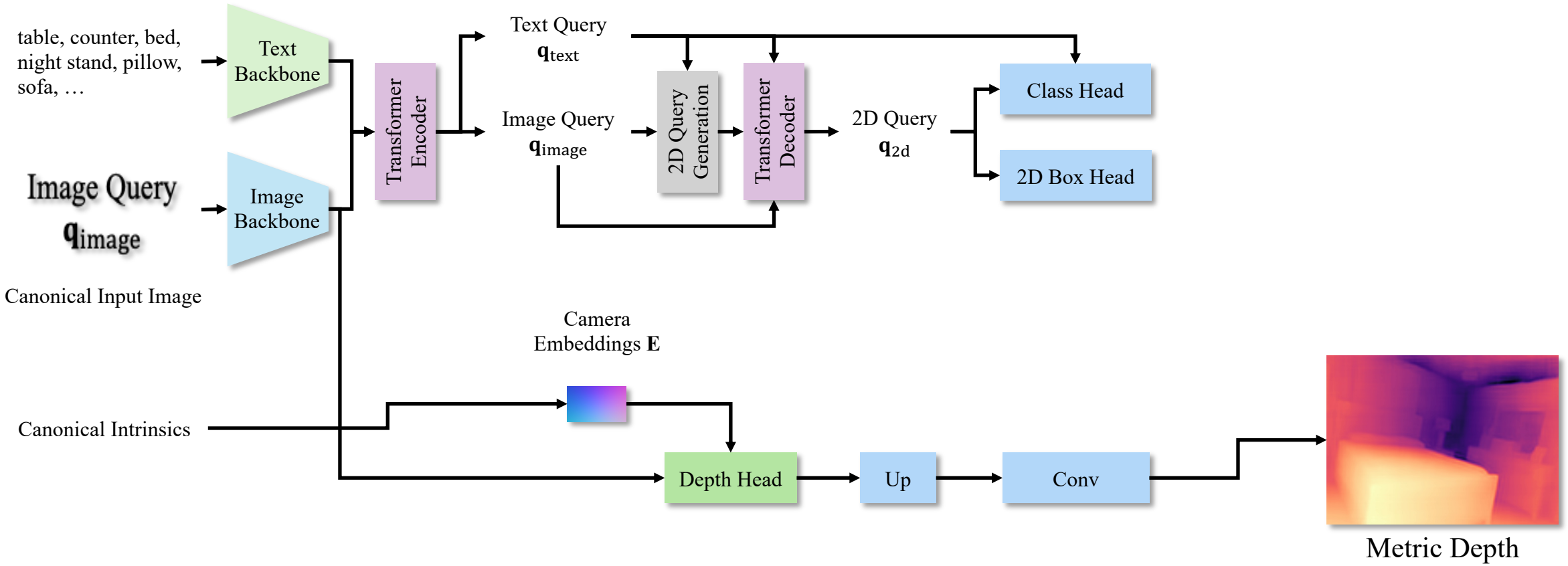
Architecture



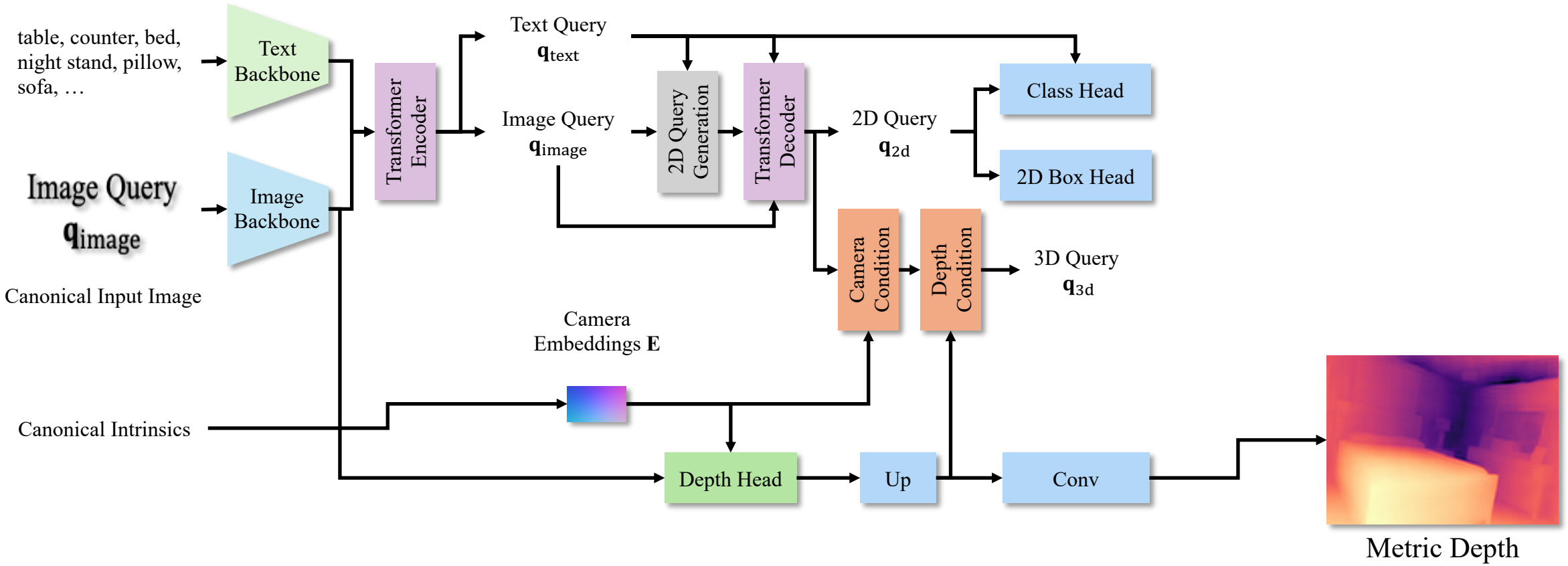
Canonical Input Image

Canonical Intrinsics

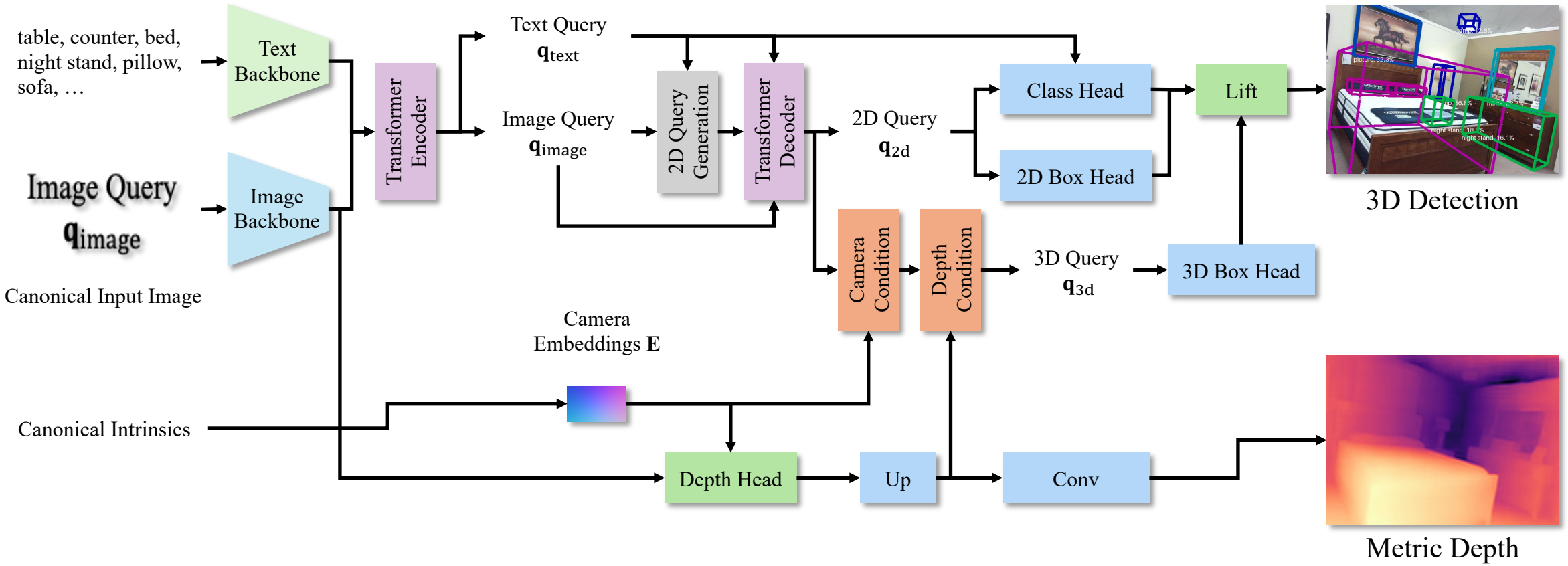
Architecture



Architecture

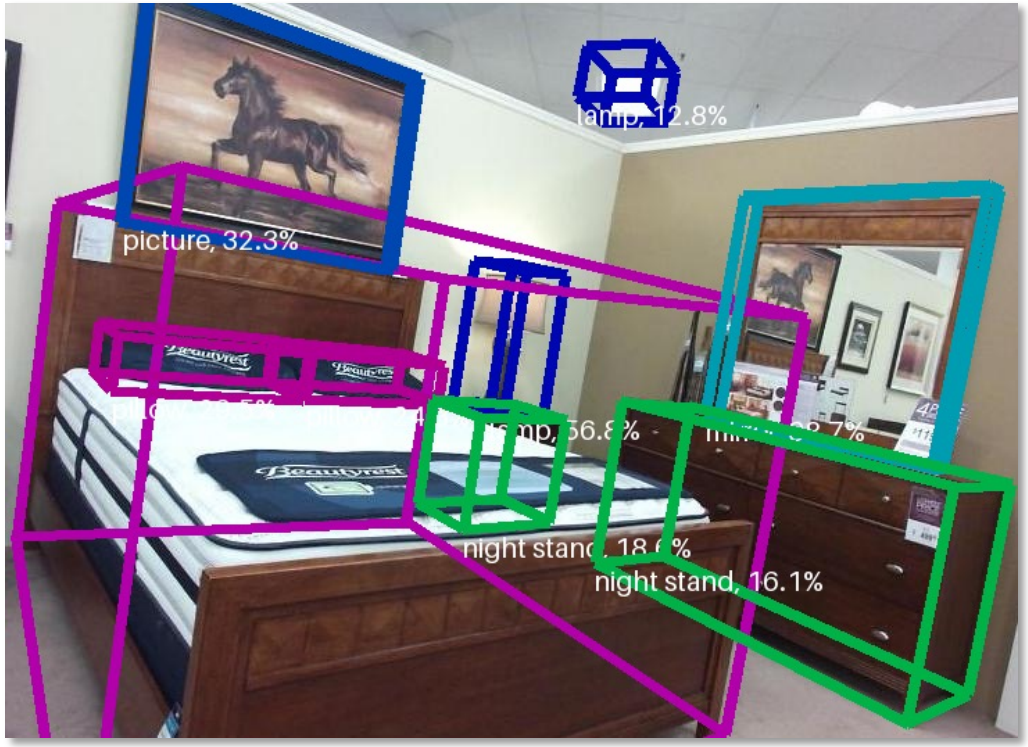


Architecture

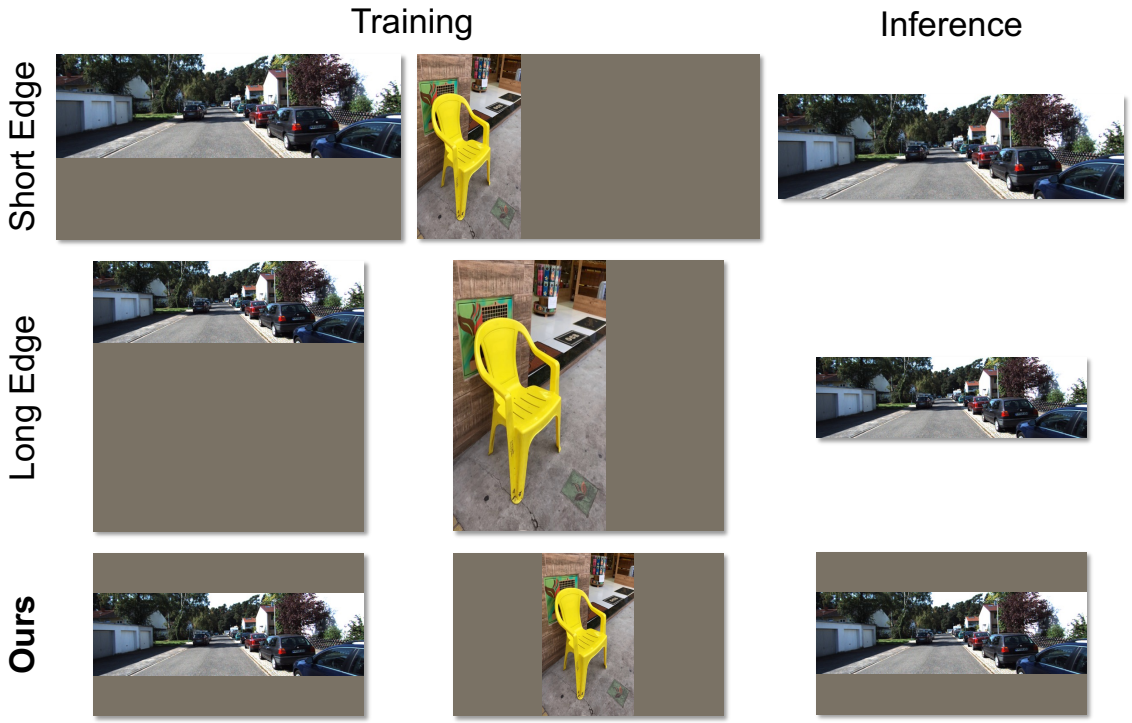


3D Bounding Box Head

- 2D-to-3D Prediction
  - Projected 3D Center
  - Center Metric Depth
  - Width, Length, Height
  - Rotation Matrix



Canonical Image Space

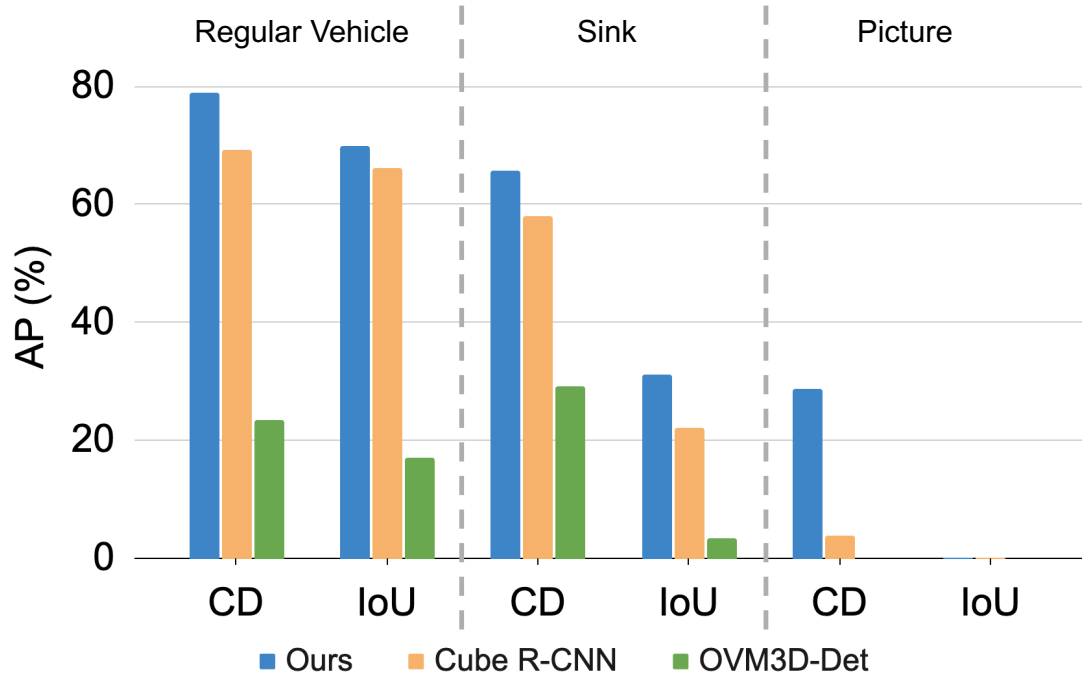


Resize	Padding	Image Resolutions	GPU RAM (G)
Short Edge	Right-Bottom	Short Edge to 512	21
Long Edge	Right-Bottom	Long Edge to 1333	23
Ours	Center	800×1333	17

Open Detection Score (ODS)

- Matching
  - Normalized Distance (Radius)
- True Positive Errors
  - Translation, Shape, Orientation

$$\text{ODS} = \frac{1}{6}[3 \times \text{AP}_{\text{3D}}^{\text{dist}} + \sum(1 - \text{mTPE})]$$



Open-set Benchmark

Method	Argoverse 2							ScanNet						
	AP_{3D}^{dist} ↑	mATE ↓	mASE ↓	mAOE ↓	ODS ↑	ODS (B) ↑	ODS (N) ↑	AP_{3D}^{dist} ↑	mATE ↓	mASE ↓	mAOE ↓	ODS ↑	ODS (B) ↑	ODS (N) ↑
Cube R-CNN (In)	-	-	-	-	-	-	-	19.5	0.725	0.771	0.858	20.5	24.6	0.0
Cube R-CNN (Out)	10.5	0.896	0.869	0.991	9.3	19.5	0.0	-	-	-	-	-	-	-
Cube R-CNN [4]	8.6	0.903	0.867	0.953	8.9	18.6	0.0	20.0	0.733	0.774	0.921	19.5	23.4	0.0
OVM3D-Det† [18]	7.7	0.914	0.893	0.899	8.8	16.5	1.7	15.6	0.798	0.871	0.818	16.3	17.8	8.8
Ours (Swin-T)	**14.8**	0.782	0.697	0.612	22.5	31.7	14.2	27.3	0.630	0.726	**0.650**	30.2	33.6	13.4
Ours (Swin-B)	14.7	**0.755**	**0.680**	**0.580**	**23.8**	**33.6**	**14.8**	**28.8**	**0.612**	**0.706**	0.655	**31.5**	**34.7**	**15.7**

Cross-domain Generalization

Method	Trained on Hypersim			Trained on SUN RGB-D			Trained on ARKitScenes		
	AP_{3D}^{hyp} ↑	AP_{3D}^{sun} ↑	AP_{3D}^{ark} ↑	AP_{3D}^{hyp} ↑	AP_{3D}^{sun} ↑	AP_{3D}^{ark} ↑	AP_{3D}^{hyp} ↑	AP_{3D}^{sun} ↑	AP_{3D}^{ark} ↑
Cube R-CNN [4]	15.2	9.5	7.5	9.5	34.7	14.2	7.5	13.1	38.6
Uni-MODE [23]	14.7	5.6	3.6	3.0	28.5	8.8	4.2	13.0	35.0
Ours	**25.6**	**15.9**	**14.5**	**13.8**	**42.1**	**21.4**	**12.9**	**23.8**	**43.9**

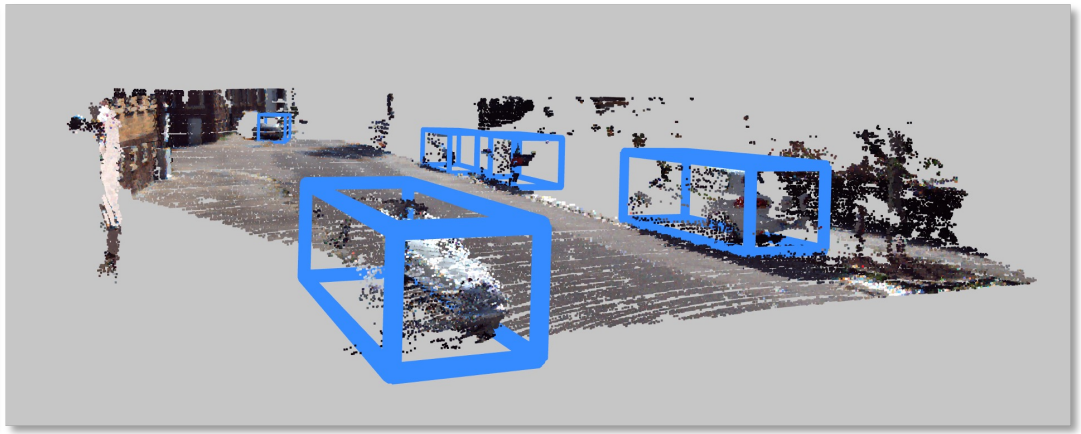
Closed-set Results

Method	AP_{3D}^{kit} ↑	AP_{3D}^{nus} ↑	AP_{3D}^{sun} ↑	AP_{3D}^{hyp} ↑	AP_{3D}^{ark} ↑	AP_{3D}^{obj} ↑	AP_{3D}^{omni} ↑
ImVoxelNet [44]	-	-	-	-	-	-	9.4
SMOKE [29]	-	-	-	-	-	-	9.6
FCOS3D [51]	-	-	-	-	-	-	9.8
PGD [52]	-	-	-	-	-	-	11.2
Cube R-CNN [4]	32.6	30.1	15.3	7.5	41.7	50.8	23.3
Uni-MODE* [23]	29.2	**36.0**	23.0	8.1	48.0	66.1	28.2
Ours (Swin-T)	**32.8**	31.5	21.9	**10.5**	51.0	64.3	28.4
Ours (Swin-B)	31.4	35.8	**23.8**	9.1	**53.9**	**67.9**	**30.0**

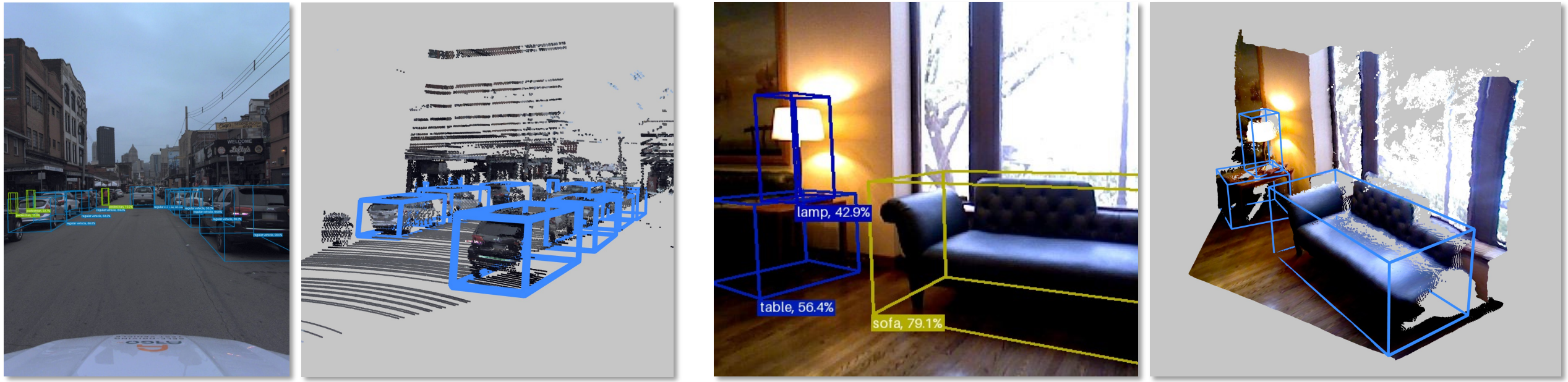
Ablation Studies

	CI	Depth	GA	AP_{3D}^{kit} ↑	AP_{3D}^{nus} ↑	AP_{3D}^{hyp} ↑	AP_{3D}^{sun} ↑	AP_{3D}^{ark} ↑	AP_{3D}^{obj} ↑	AP_{3D}^{omni} ↑	ODS^{av2} ↑	ODS^{scan} ↑	ODS^{open} ↑
1	-	-	-	**32.5**	29.7	8.1	17.3	46.5	54.9	24.1	18.2	29.0	23.6
2	✓	-	-	31.1	30.5	9.1	19.1	47.7	58.1	25.5	19.5	29.5	24.5
3	✓	✓	-	29.8	30.7	**10.3**	19.9	48.6	58.8	26.2	20.0	29.4	24.7
4	✓	✓	✓	32.1	**31.9**	9.9	**20.8**	**49.1**	**60.2**	**26.8**	**22.0**	**30.0**	**26.0**

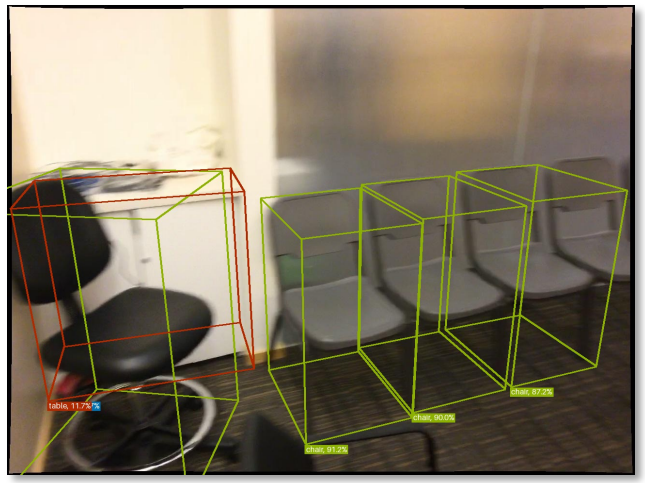
Qualitative Results

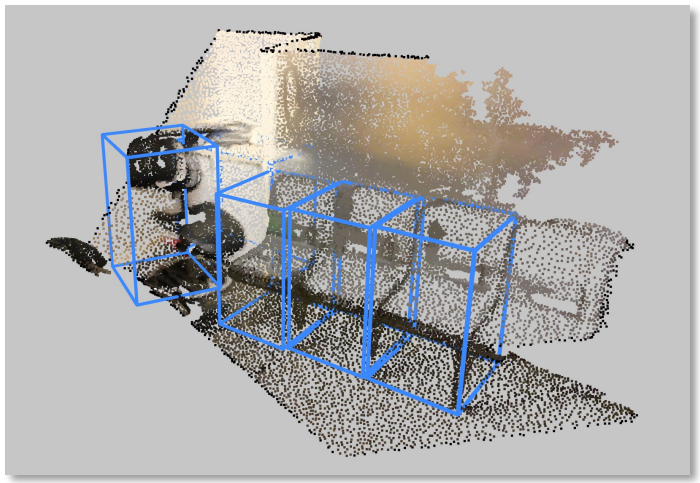


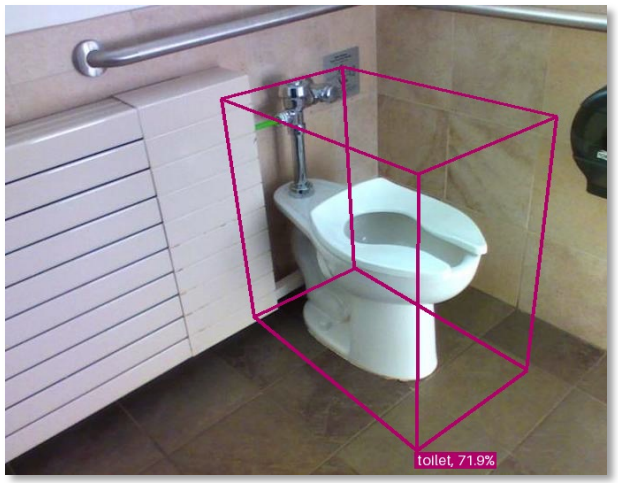
Qualitative Results

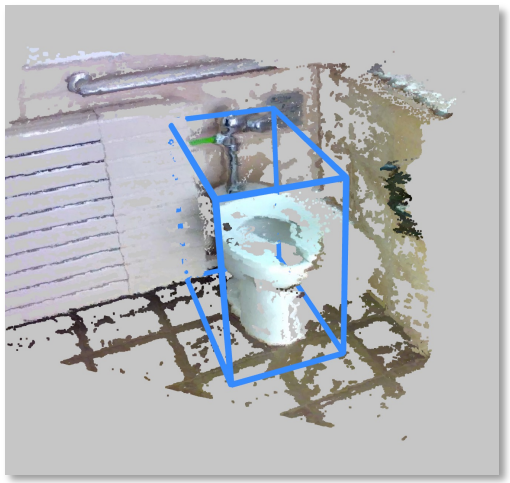


Qualitative Results









ETH zürich | Computer Vision and Geometry Lab

ICCV OCT 19-23, 2025 HONOLULU HAWAII

3D-MOOD: Lifting 2D to 3D for Monocular Open-Set Object Detection

Yung-Hsu Yang · Luigi Piccinelli · Mattia Segu · Siyuan Li · Rui Huang

Yuqian Fu · Marc Pollefeys · Hermann Blum · Zuria Bauer

Project Page

Code & Models

Live Demo