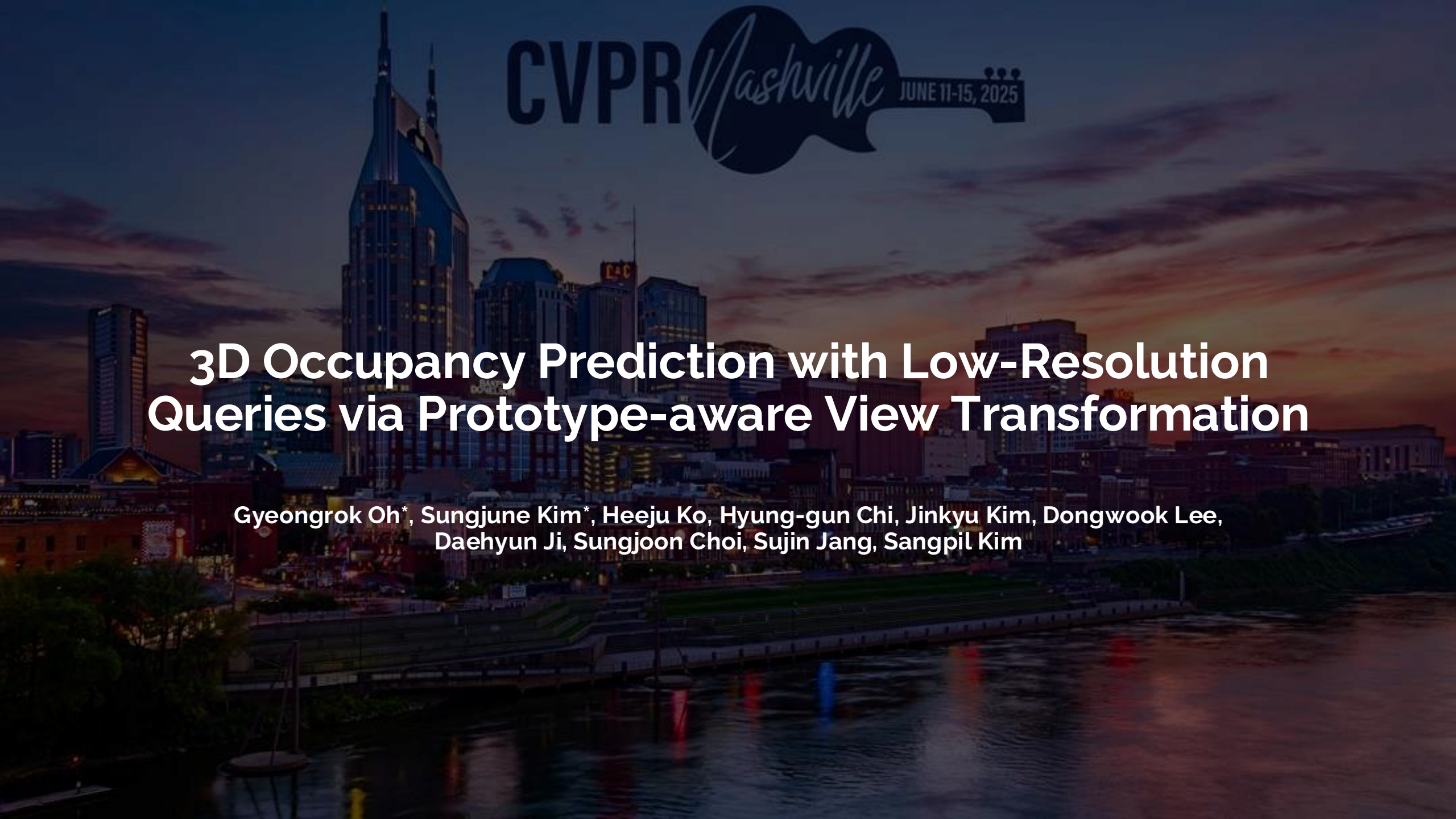
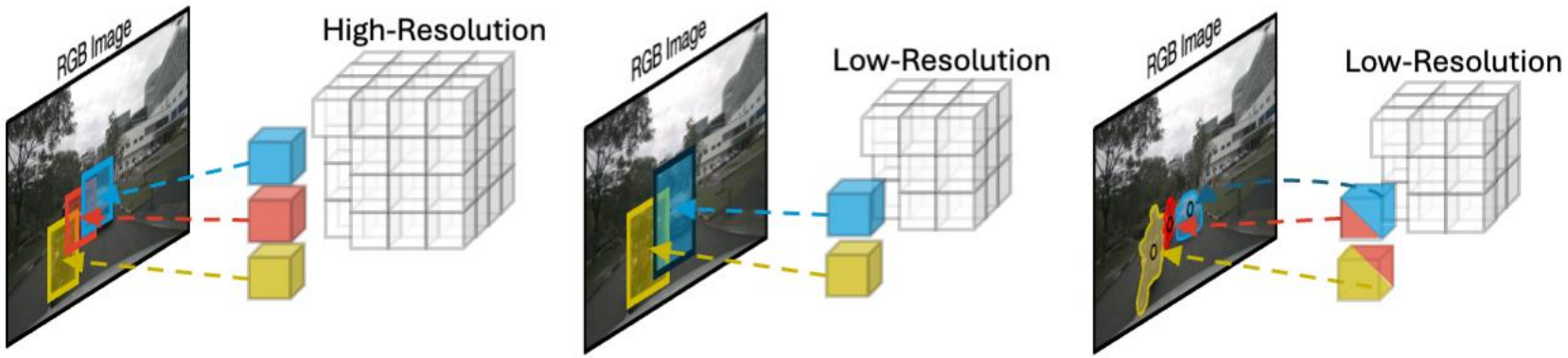


3D Occupancy Prediction with Low-Resolution Queries via Prototype-aware View Transformation

Gyeongrok Oh*, Sungjune Kim*, Heeju Ko, Hyung-gun Chi, Jinkyu Kim, Dongwook Lee, Daehyun Ji, Sungjoon Choi, Sujin Jang, Sangpil Kim

Motivation

- Efficient 3D Voxelization is critical for camera-based 3D Occupancy Prediction (3DOP).
- While reducing voxel query size speeds up inference, it often sacrifices performance by missing fine-grained spatial details.
- To overcome this challenges, we **integrate comprehensive visual contexts** into low-resolution voxel queries, enabling accurate and efficient 3DOP.



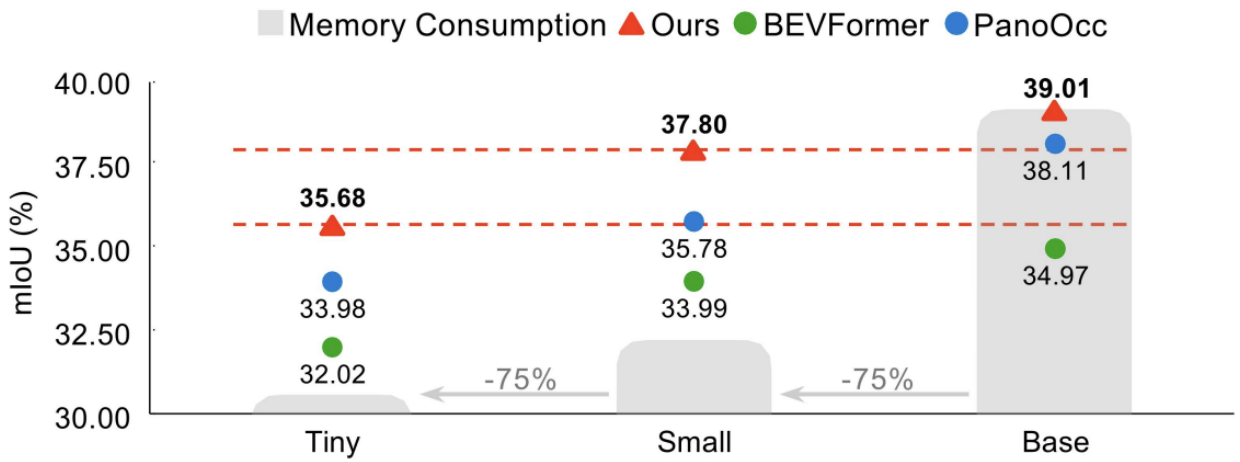
(b) Standard VT w/ High-Resolution

(c) Standard VT w/ Low-Resolution

(d) Prototype-based VT **(Ours)** w/ Low-Resolution

Contributions

- Introducing **ProtoOcc** as an exemplar in 3DOP by using computationally efficient low-resolution voxel queries
- A **prototype-aware view transformation** and **multi-perspective decoding** strategy for enhancing the representations of low-resolution voxel queries.
- Demonstrating clear improvements over previous state-of-the-art methods in two major benchmarks, along with detailed analyses.



Overview

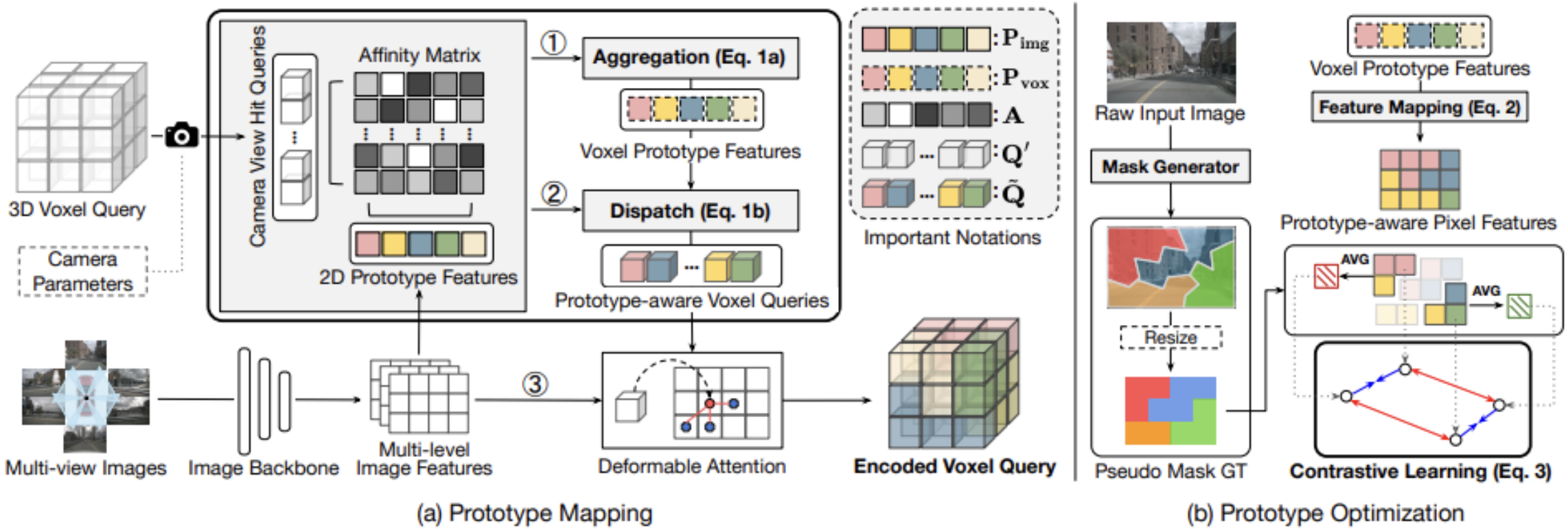
We introduce ProtoOcc, a novel occupancy network leveraging image prototype representations in view transformation for low-resolution context enhancement.

1. **Prototype-aware View Transformation**
   - Prototype Mapping
   - Prototype Optimization

2. **Multi-perspective Occupancy Decoding**
   - Multi-perspective View Generation
   - Scene Consistency Regularization

Method

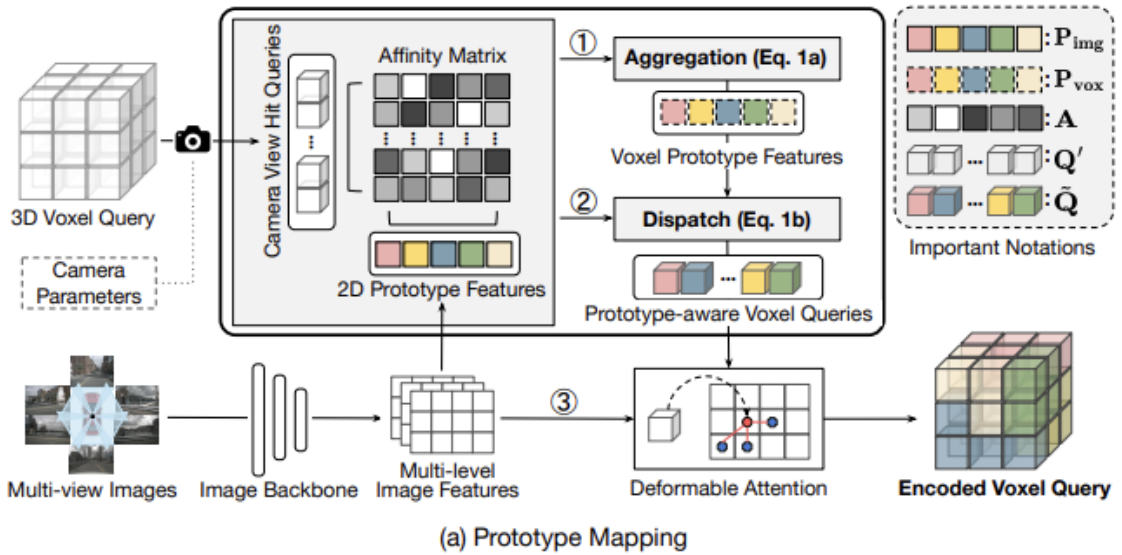
Prototype-aware View Transformation



Method

Prototype-aware View Transformation

- Iterative Feature Grouping
  - Iteratively update the image features by utilizing iterative grouping strategy to obtain 2D prototype features.
- Prototype Mapping
  - Lift 2D prototype features into the 3D voxel query space.



(a) Prototype Mapping

$$\texttt{Aggregate:}\ \mathbf{P}_{\text{vox}} = \frac{1}{R}\left(\hat{\mathbf{P}}_{\text{img}} + \sigma(\mathbf{A}) \cdot \mathbf{Q}'\right)$$

$$\texttt{Dispatch:}\ \tilde{\mathbf{Q}} = \mathbf{Q}' + \texttt{MLP}\left(\sigma(\mathbf{A})^{\mathsf{T}} \cdot \mathbf{P}_{\text{vox}}\right)$$

Method

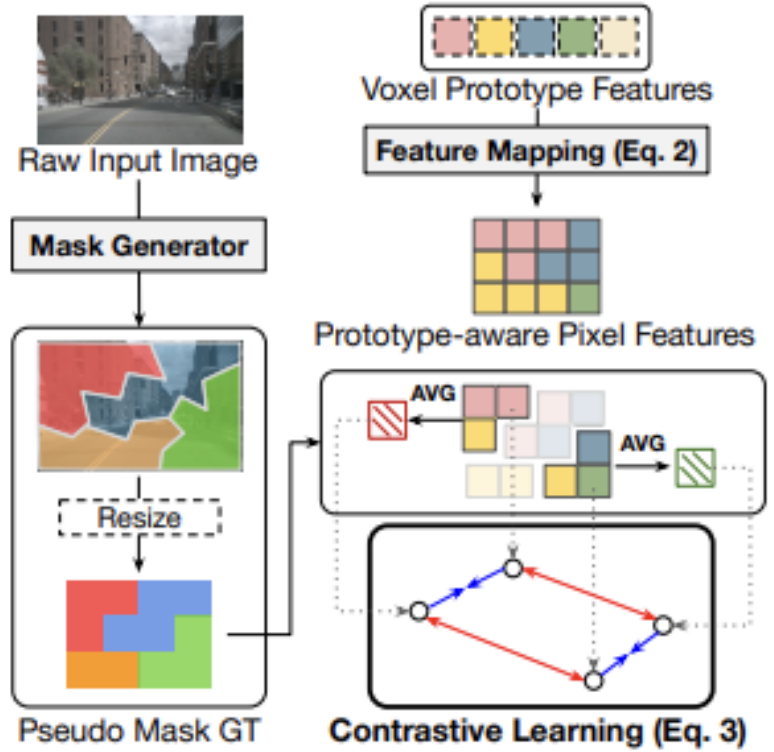
Prototype-aware View Transformation

- Prototype Optimization
  - We introduce pseudo-2D supervision to guide cluster learning absent in standard 3DOP.

$$\mathbf{X} = \{\mathbf{G} \odot \mathcal{H}(\mathbf{A})\} * \mathbf{P}_{\text{vox}}$$

  - The above process acts as a bridge for mapping 3D voxel queries onto 2D grid cells.
  - Pseudo masks from clustering or a foundation model guide contrastive loss, enhancing prototype distinctiveness.

$$\mathcal{L}_{\text{cls}}^{(x,y)} = -\log \frac{\sum_{s=1}^{S} m_s \exp\left(\langle \mathbf{M}_s, \mathbf{X}_{(x,y)} \rangle / \tau_{\text{cls}}\right)}{\sum_{s=1}^{S} \exp\left(\langle \mathbf{M}_s, \mathbf{X}_{(x,y)} \rangle / \tau_{\text{cls}}\right)}$$



(b) Prototype Optimization

Method

Multi-perspective Occupancy Decoding

- Multi-perspective view generation
  - Introduce feature-level and spatial-level voxel augmentation to generate diverse context views.
  - Shared kernel yields diverse scene view from local voxel variants.

- Scene Consistency Regularization
  - Enforce semantic consistency via GRAND by minimizing L2 distances between augmented predictions and their average.

$$\mathcal{L}_{\text{cons}} = \frac{\sum_{p=0}^{P}\sum_{k=1}^{Z}\sum_{j=1}^{W}\sum_{i=1}^{H}\left\|\tilde{\mathbf{V}}_{(i,j,k)} - \mathcal{G}(\mathbf{V}^{(p)}_{(i,j,k)})\right\|_2^2}{H \cdot W \cdot Z \cdot (P+1)}$$

Experiments

- We train and evaluate our method on two representative tasks in 3D scene understanding: 3D occupancy prediction (3DOP) and 3D semantic scene completion (3DSSC).

- Benchmark Datasets:
  - Occ3D-nuScense
  - SemanticKITTI

Experiments

Quantitative Results

- Below table indicates the quantitative results on the Occ3D-nuScenes and SemanticKITTI validation set.
- Not only does ProtoOcc stand out in its category, but also by comparing the results within the same color marks.
- It is apparent that ProtoOcc can overcome query deficiencies, performing on par even with higher-resolution counterparts.

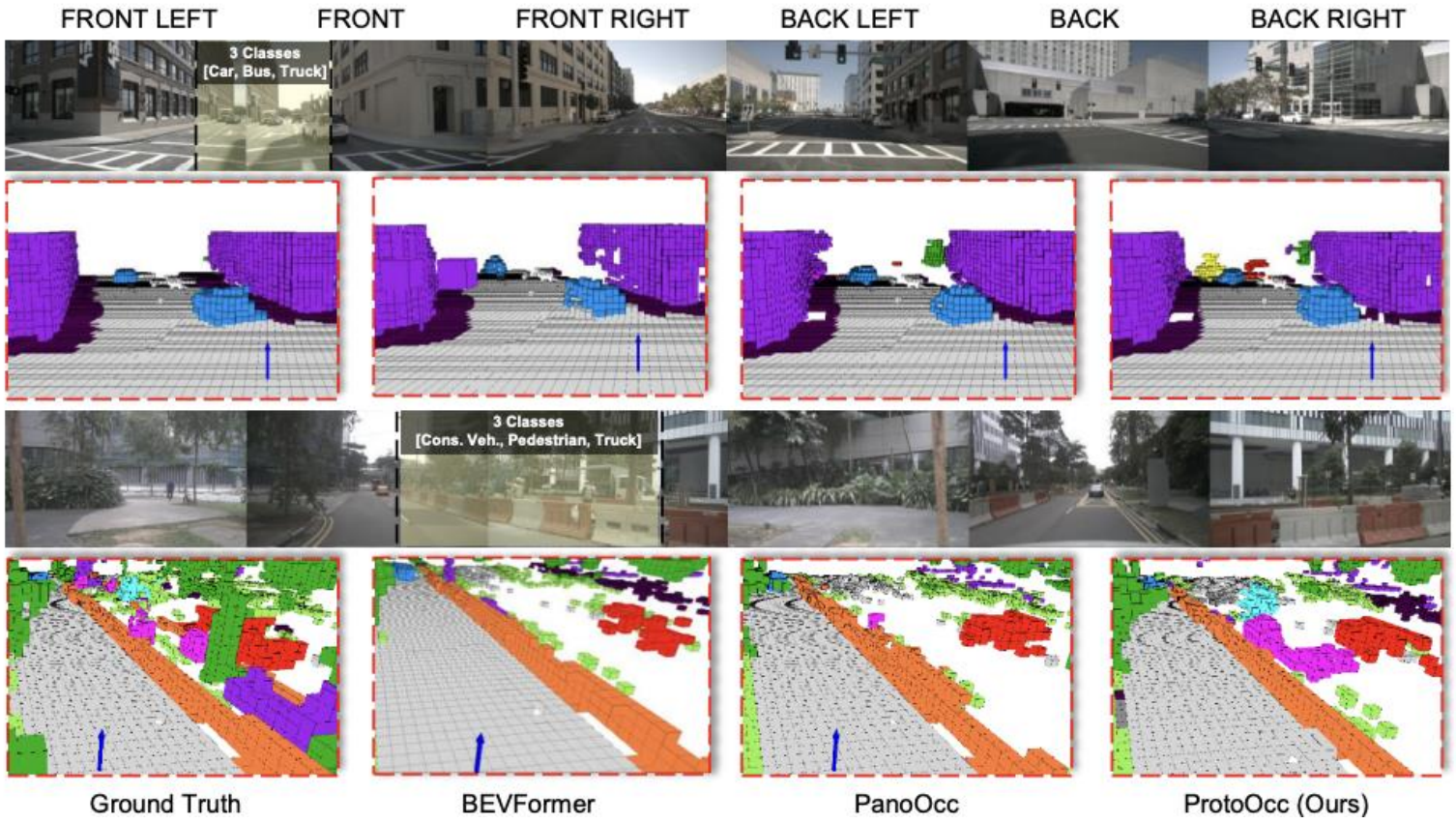
Query Size	Model	mIoU	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
Base	CTF-Occ [36]	28.50	8.09	39.33	20.56	38.29	42.24	16.93	24.52	22.72	21.05	22.98	31.11	53.33	33.84	37.98	33.23	20.79	18.00
	TPVFormer [12]	34.20	7.68	44.01	17.66	40.88	46.98	15.06	20.54	24.69	24.66	24.26	29.28	79.27	40.65	48.49	49.44	32.63	29.82
	SurroundOcc [47]	34.60	9.51	38.50	22.08	39.82	47.04	20.45	22.48	23.78	23.00	27.29	34.27	78.32	36.99	46.27	49.71	35.93	32.06
	OccFormer [51]	37.04	9.15	45.84	18.20	42.80	50.27	24.00	20.80	22.86	20.98	31.94	38.13	80.13	38.24	50.83	54.3	**46.41**	**40.15**
	BEVFormer* [22]	34.97	7.53	41.77	16.39	44.06	48.48	17.27	20.01	23.36	21.16	28.88	35.59	80.12	35.35	47.65	51.89	40.68	34.28
	PanoOcc* [45]	38.11	9.75	45.31	22.45	43.13	50.19	22.25	27.35	24.49	**25.17**	31.74	37.95	81.74	42.29	50.82	54.80	40.81	37.14
	ProtoOcc (Ours)	**39.01**	**9.75**	**46.08**	**24.34**	**46.09**	**52.45**	**24.21**	**28.11**	**24.72**	19.79	**32.90**	**40.50**	**82.29**	**43.02**	**52.47**	**55.94**	42.46	38.13
Small	BEVFormer* [22]	33.98	6.75	41.67	13.91	41.97	48.49	17.83	18.01	22.19	19.08	29.64	33.23	79.42	36.48	46.82	49.26	39.04	33.91
	PanoOcc* [45]	35.78	8.18	41.60	20.79	41.25	47.78	21.87	23.42	21.03	19.29	29.71	36.10	81.20	40.00	49.22	53.94	38.09	34.83
	ProtoOcc (Ours)*	**37.80**	**9.28**	**43.64**	**22.30**	**44.72**	**50.07**	**23.68**	**25.23**	**22.77**	**19.66**	**30.43**	**38.73**	**82.05**	**42.61**	**51.68**	**55.84**	**41.91**	**38.05**
Tiny	BEVFormer [22]	32.02	4.86	39.79	7.17	42.46	47.10	18.46	13.18	17.76	12.46	28.74	33.19	78.64	35.36	45.27	47.29	38.93	33.61
	PanoOcc [45]	33.99	6.97	39.60	18.80	40.67	45.63	18.19	21.43	19.10	16.53	25.99	35.15	80.60	38.44	49.02	52.11	36.81	32.87
	ProtoOcc (Ours)*	**35.68**	**8.33**	**40.55**	**19.84**	**42.95**	**48.08**	**20.31**	**22.78**	**21.21**	**17.00**	**28.22**	**36.60**	**81.42**	**41.32**	**50.16**	**53.82**	**38.63**	**35.35**

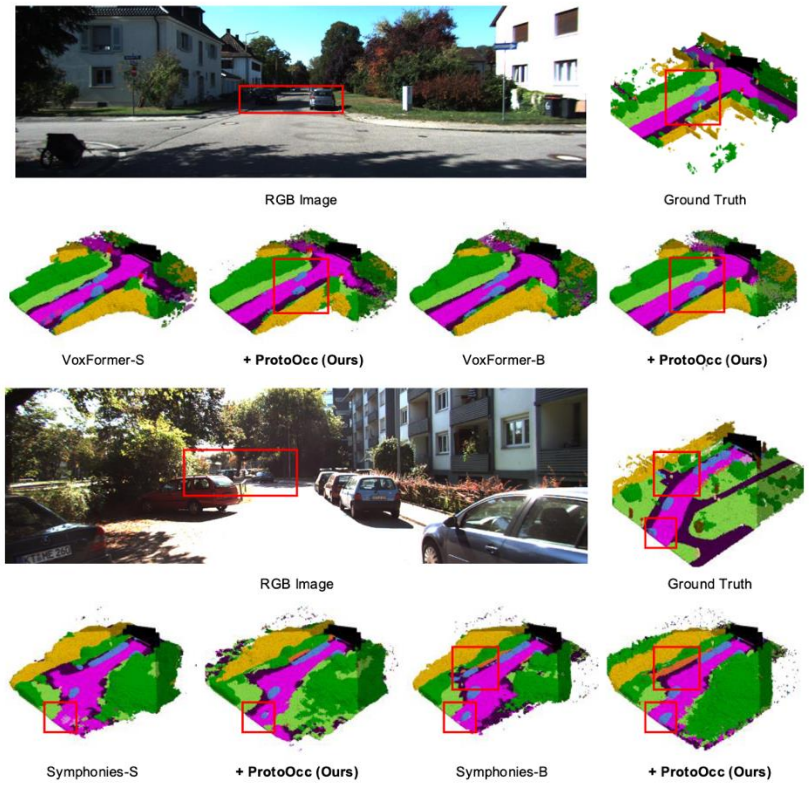
Model	Pub.	IoU	mIoU
MonoScene [3]	CVPR 22	36.86	11.08
TPVFormer [12]	CVPR 23	35.61	11.36
OccFormer [51]	ICCV 23	36.50	13.46
HASSC [43]	CVPR 24	44.82	13.48
VoxFormer-S [21]	CVPR 23	43.10	11.51
+ ProtoOcc	-	43.55 (+0.35)	12.39 (+0.88)
VoxFormer-B [21]	CVPR 23	44.02	12.35
+ ProtoOcc	-	**44.90** (+0.85)	13.57 (+1.22)
Symphonies-S [14]	CVPR 24	41.67	13.64
+ ProtoOcc	-	43.02 (+1.35)	14.50 (+0.86)
Symphonies-B [14]	CVPR 24	41.85	14.38
+ ProtoOcc	-	42.12 (+0.27)	**14.83** (+0.45)

Experiments

Qualitative Results

- Our proposed ProtoOcc enables to capture the long-range objects and occluded objects.





Experiments

Driving Video

Experiments

Ablation Study

- We conduct ablation studies to analyze the impact of augmentation and prototype settings.

Table 5. Effects of different augmentations and their consistency regularization (C.R.).

	Exp. #	1	2	3	4	5	6	7	8	9	10	11
Augmentation	Random Dropout	-	✓	-	-	-	✓	✓	-	-	✓	-
	Gaussian Noise	-	-	✓	-	-	-	-	✓	✓	✓	-
	Transpose	-	-	-	✓	-	✓	-	✓	-	-	✓
	Flips	-	-	-	-	✓	-	✓	-	✓	-	✓
mIoU	w/o C.R.	35.78	36.99	36.46	36.49	36.70	36.95	36.78	36.47	36.72	36.66	36.59
	w/ C.R.	N/A	**37.19**	**36.80**	**36.86**	**37.18**	**37.21**	**37.16**	**36.90**	**36.82**	**37.25**	**36.78**

Table 6. Sensitivity on the number of 2D prototypes M.

2D prototype M	1440 ($r = 2$)	350 ($r = 4$)	144 ($r = 6$)	91 ($r = 8$)
# Pseudo Mask		299		
mIoU	36.90	37.80	36.95	37.07

Table 7. Effect of pseudo mask quality.

2D prototype M	Mask Generator	# Pseudo Mask	mIoU
350	SEEDS [38]	91	37.14
	Segment Anything [16]	92	37.29

Table 8. Impact of the granularity level.

2D prototype M	Mask Generator	# Pseudo Mask	mIoU
91	SEEDS [38]	299	37.07
	Segment Anything [16]	92	37.21
350	SEEDS [38]	299	37.80
	Segment Anything [16]	92	37.29

Conclusion

3D Occupancy Prediction with Low-Resolution Queries via Prototype-aware View Transformation

Gyeongrok Oh[1,*], Sungjune Kim[1,*], Heeju Ko[1], Hyung-gun Chi[2], Jinkyu Kim[1], Dongwook Lee[3], Daehyun Ji[3], Sungjoon Choi[1], Sujin Jang[3,†], Sangpil Kim[1,†]

[1]Korea University [2]Purdue University [3]AI Center, DS Division, Samsung Electronics

Abstract

The resolution of voxel queries significantly influences the quality of view transformation in camera-based 3D occupancy prediction. However, computational constraints and the practical necessity for real-time deployment require smaller query resolutions, which inevitably leads to an information loss. Therefore, it is essential to encode and preserve rich visual details within limited query sizes while ensuring a comprehensive representation of 3D occupancy. To this end, we introduce ProtoOcc, a novel occupancy network that leverages prototypes of clustered image segments in view transformation to enhance low-resolution context. In particular, the mapping of 2D prototypes onto 3D voxel queries encodes high-level visual geometries and complements the loss of spatial information from reduced query resolutions. Additionally, we design a multi-perspective decoding strategy to efficiently disentangle the densely compressed visual cues into a high-dimensional 3D occupancy scene. Experimental results on both Occ3D and SemanticKITTI benchmarks demonstrate the effectiveness of the proposed method, showing clear improvements over the baselines. More importantly, ProtoOcc achieves competitive performance against the baselines even with 75% reduced voxel resolution. Project page: https://kuai-lab.github.io/cvpr2025protoocc.

1. Introduction

3D occupancy prediction (*3DOP*) is the task of determining which parts of a 3D space are occupied by objects and identifying their class categories. In particular, view transformation from 2D to 3D space is an essential step for successful camera-based occupancy prediction. Prior works employ various 3D space representation strategies, such as bird's-eye-view (BEV) plane, tri-perspective-view (TPV) planes, and 3D voxel cube, which aim to map multi-view camera information into corresponding grid cells in a unified 3D space. Among these, 3D voxels are the most common representation strategy in the 3DOP task [11, 12, 37, 40, 45, 47, 49, 51], as they naturally encode visual information into a structured semantic 3D space. Therefore, the effectiveness of 3D voxelization (*i.e.*, voxel queries) plays a crucial role in determining the prediction performance in camera-based 3DOP.

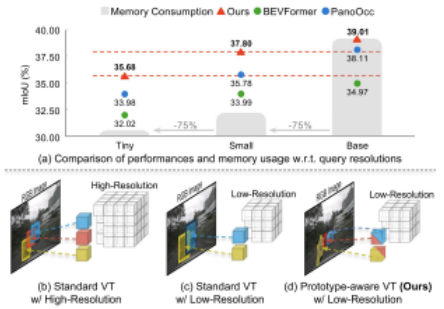


Figure 1. (a) Our *ProtoOcc* can perform comparably to higher-resolution counterparts while using 75% less memory. (b-c) Reducing query resolutions in standard view transformation (VT) is required for faster inference, but brings geometrical ambiguity. (d) Our prototype-aware VT can capture high-level geometric details while preserving computational efficiency.

Trade-offs between computation and performance have long been a challenge in various deep learning applications [6, 8, 17, 26]. Learning voxel queries also encounters this dilemma, especially in real-time and safety-critical vision systems such as autonomous driving and robotics. As illustrated in Figure 1, utilizing voxel queries with high-resolution may produce reliable performances, but requires intensive computation. Accelerating the inference inevitably necessitates smaller query sizes. However, it results in performance degradation primarily due to its inability to capture precise high-level details within limited spatial storage,

*Equal contributions
†Corresponding author (s.steve.jang@samsung.com, spk7@korea.ac.kr)

Thank you for listening