A Spatiotemporal Approach to Tri-Perspective Representation for 3D Semantic Occupancy Prediction

[1,2]Sathira Silva [1]Savindu Wannigama [3]Gihan Jayatilaka (presenter) [2]Muhammad Haris Khan [1]Roshan Ragel

[1]University of Peradeniya, Peradeniya 20400, Sri Lanka
[2]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[3]University of Maryland, College Park, MD 20742, USA

Contents

- Background & Problem Statement
- Contributions
- Architecture:
  - Virtual View Transformation (VVT)
  - Spatial Cross Attention (SCA)
  - Temporal Cross View Hybrid Attention (TCVHA)
- Experiments:
  - 3D SOP
  - LiDAR Segmentation

Background & Problem Statement

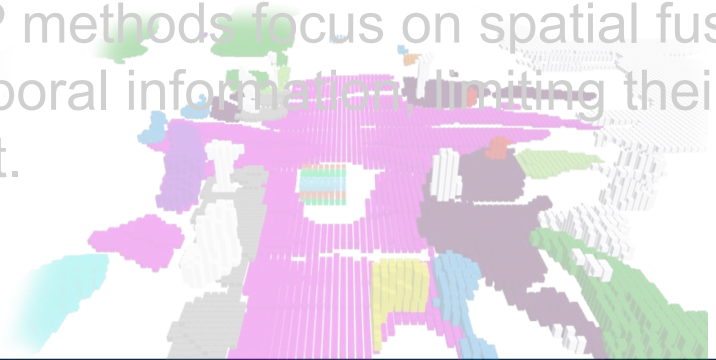
- 3D Semantic Occupancy Prediction (SOP) aims to predict per-voxel semantic labels for a 3D scene, enabling a dense and structured understanding of the environment for applications like autonomous driving and robotics.

- Existing 3D SOP methods focus on spatial fusion while overlooking temporal information, limiting their ability to leverage historical context.

Background & Problem Statement

- 3D Semantic Occupancy Prediction (SOP) aims to predict per-voxel semantic labels for a 3D scene, enabling a dense and structured understanding of the environment for applications like autonomous driving and robotics.

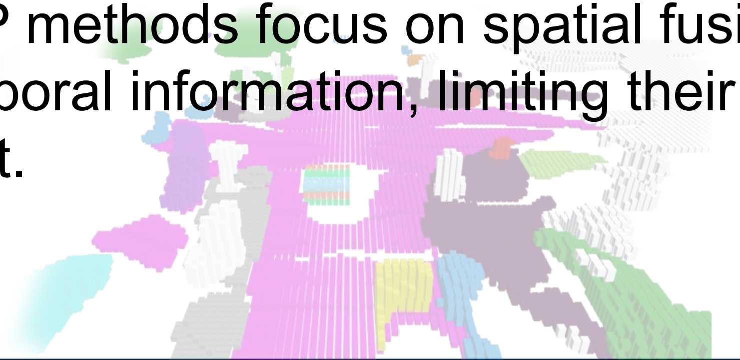
- Existing 3D SOP methods focus on spatial fusion while overlooking temporal information, limiting their ability to leverage historical context.



Background & Problem Statement

- 3D Semantic Occupancy Prediction (SOP) aims to predict per-voxel semantic labels for a 3D scene, enabling a dense and structured understanding of the environment for applications like autonomous driving and robotics.

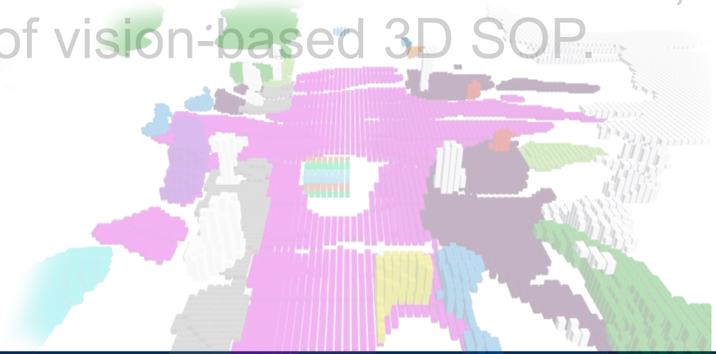
- Existing 3D SOP methods focus on spatial fusion while overlooking temporal information, limiting their ability to leverage historical context.



Contributions

- We introduce S2TPVFormer, which features a novel temporal fusion workflow for TPV representation and utilizes CVHA to enhance spatiotemporal information sharing across planes.

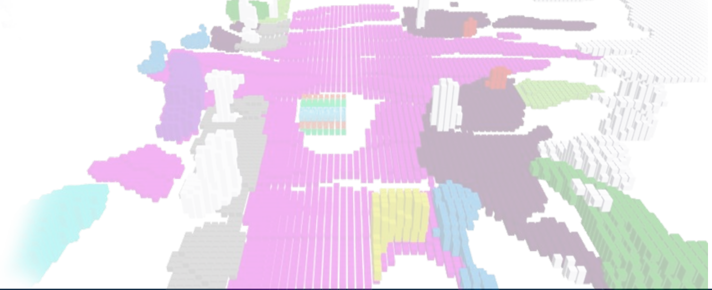
- S2TPVFormer achieves a **+4.1% mIOU improvement over TPVFormer** on the nuScenes validation set, showcasing the strong potential of vision-based 3D SOP.



Contributions

- We introduce S2TPVFormer, which features a novel temporal fusion workflow for TPV representation and utilizes CVHA to enhance spatiotemporal information sharing across planes.

- S2TPVFormer achieves a **+4.1% mIOU improvement over TPVFormer** on the nuScenes validation set, showcasing the strong potential of vision-based 3D SOP.



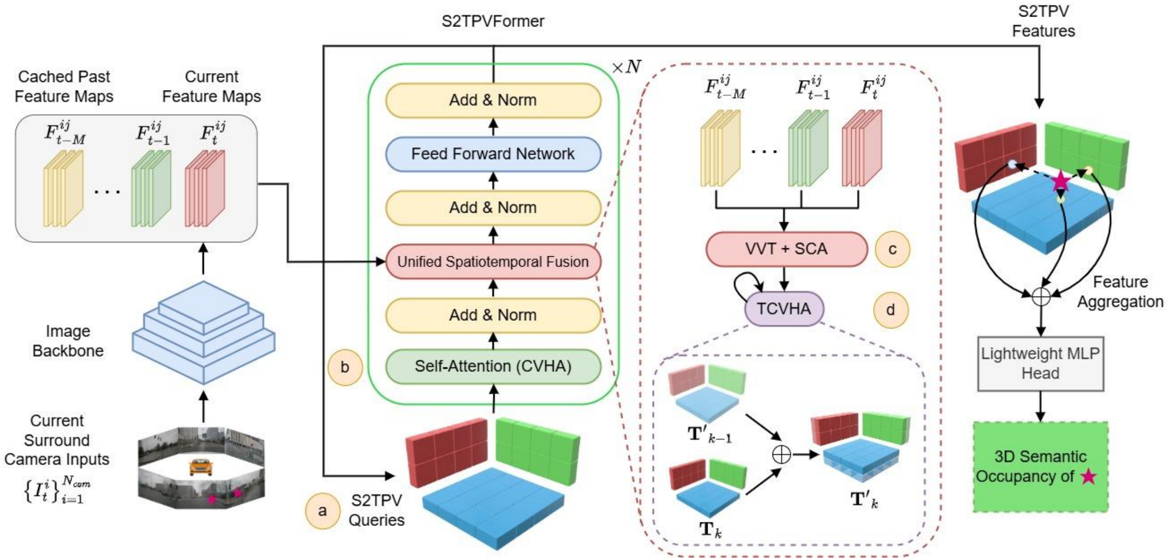
Contributions

- We introduce S2TPVFormer, which features a novel temporal fusion workflow for TPV representation and utilizes CVHA to enhance spatiotemporal information sharing across planes.

- S2TPVFormer achieves a **+4.1% mIOU improvement over TPVFormer** on the nuScenes validation set, showcasing the strong potential of vision-based 3D SOP.

Architecture

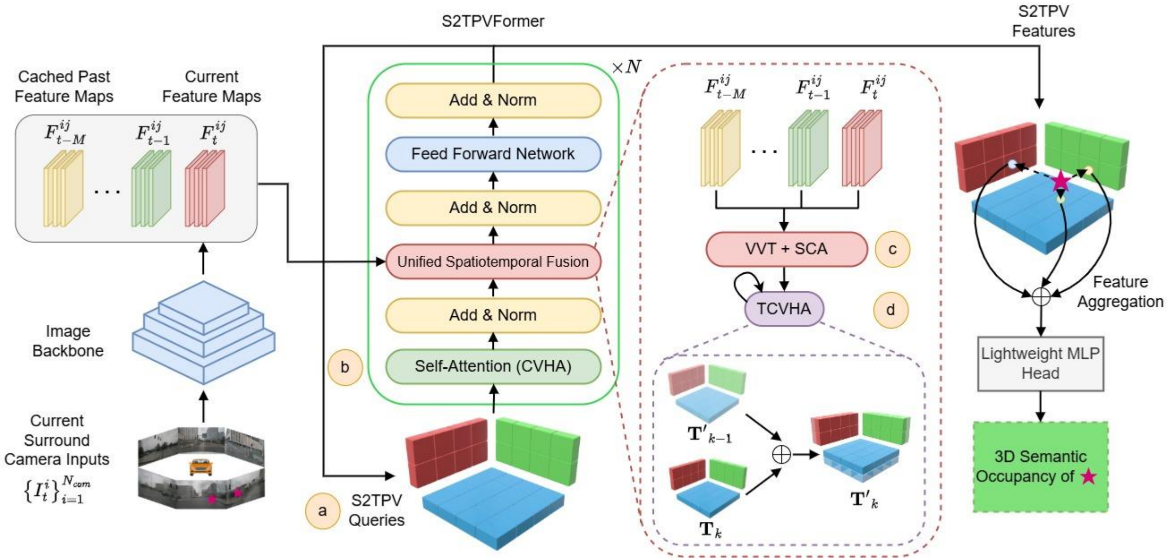
Virtual View Transformation (VVT)



- **Purpose:** Enables viewing camera features as if they were captured in the current time step.
- **How It Works:** Reconstructs missing or misaligned visual information from past views.

Architecture

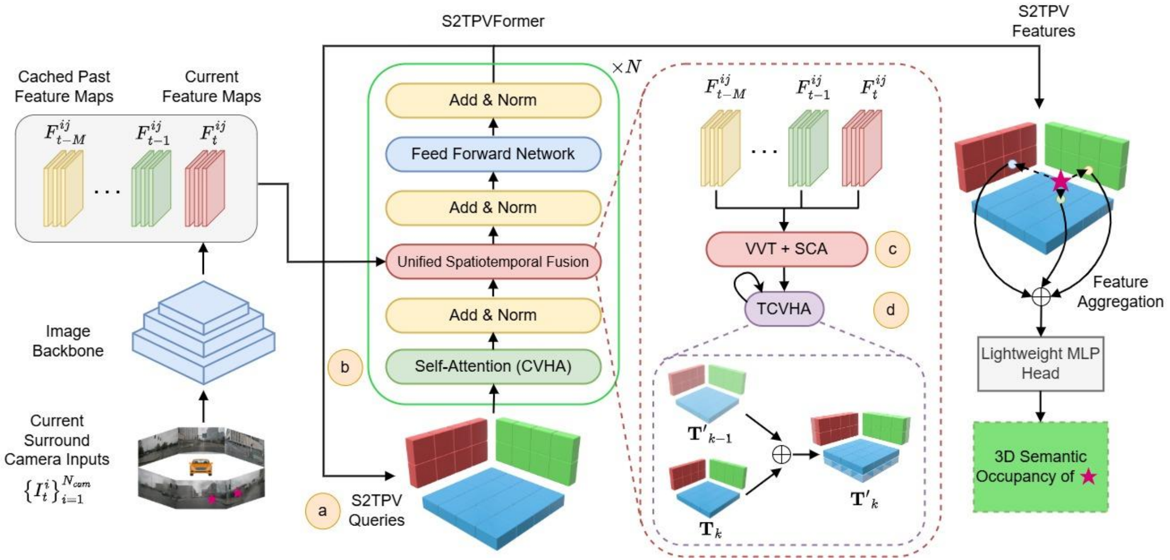
Spatial Cross Attention (SCA)



- **Purpose:** Fuses virtual camera view features onto S2TPV queries at each time step.
- **How It Works:** Extracts spatial features from virtual camera views, aligns and integrates these features with current S2TPV queries.

Architecture

Temporal Cross View Hybrid Attention (TCVHA)



- **Purpose:** Merges virtual spatial TPV features across multiple time steps.
- **How It Works:** Establishes cross-time dependencies and refines spatial-temporal feature fusion for better scene understanding.

Experiments

Comparative Results for 3D SOP

Method	mIoU (%)	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	other flat	sidewalk	terrain	manmade	vegetation
TPVFormer	52.0	59.6	**26.3**	77.6	74.1	30.9	47.5	**41.8**	20.2	44.9	67.8	86.3	54.5	55.5	54.6	47.5	44.0
S2TPVFormer (Base)	**56.1**	**60.1**	16.5	**85.9**	**74.3**	**42.2**	**51.5**	37.0	**21.2**	**49.4**	**74.2**	**86.4**	**56.3**	**57.9**	**55.0**	**65.4**	**65.0**
S2TPVFormer (Small)	43.4	54.3	17.2	66.0	69.5	28.2	22.8	32.1	15.1	31.7	59.6	82.4	49.9	47.8	47.4	34.9	36.0

3D SOP results on the nuScenes validation set

+4.1% improvement to mIoU accuracy compared to SOTA

Experiments

Comparative Results LiDAR Segmentation

Method	Input Modality	mIoU (%)	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	other flat	sidewalk	terrain	manmade	vegetation
MINet	LiDAR	56.3	54.6	8.2	62.1	76.6	23.0	58.7	37.6	34.9	61.5	46.9	93.3	56.4	63.8	64.8	79.3	78.3
LidarMultiNet	LiDAR	81.4	80.4	48.4	94.3	90.0	71.5	87.2	85.2	80.4	86.9	74.8	97.8	67.3	80.7	76.5	92.1	89.6
UniVision	LiDAR	72.3	72.1	34.0	85.5	89.5	59.3	75.5	69.3	65.8	84.2	71.4	96.1	67.4	71.9	65	77.9	71.7
PanoOcc	LiDAR	71.4	82.5	32.3	88.1	83.7	46.1	76.5	67.6	53.6	82.9	69.5	96.0	66.3	72.3	66.3	80.5	77.3
OccFormer	LiDAR	70.8	72.8	29.9	87.9	85.6	57.1	74.9	63.2	53.5	83	67.6	94.8	61.9	70.0	66.0	84.0	80.5
TPVFormer-Small†	Camera	59.2	65.6	15.7	75.1	80.0	45.8	43.1	44.3	26.8	72.8	55.9	92.3	53.7	61.0	59.2	79.7	75.6
TPVFormer-Base†	Camera	69.4	74.0	27.5	86.3	85.5	60.7	68.0	62.1	49.1	81.9	68.4	94.1	59.5	66.5	63.5	83.8	79.9
S2TPVFormer (Base)	Camera	60.4	61.2	18.2	80.6	78.1	55.2	57.6	41.5	26.4	76.1	61.3	89.8	49.4	56.6	58.0	79.3	76.4

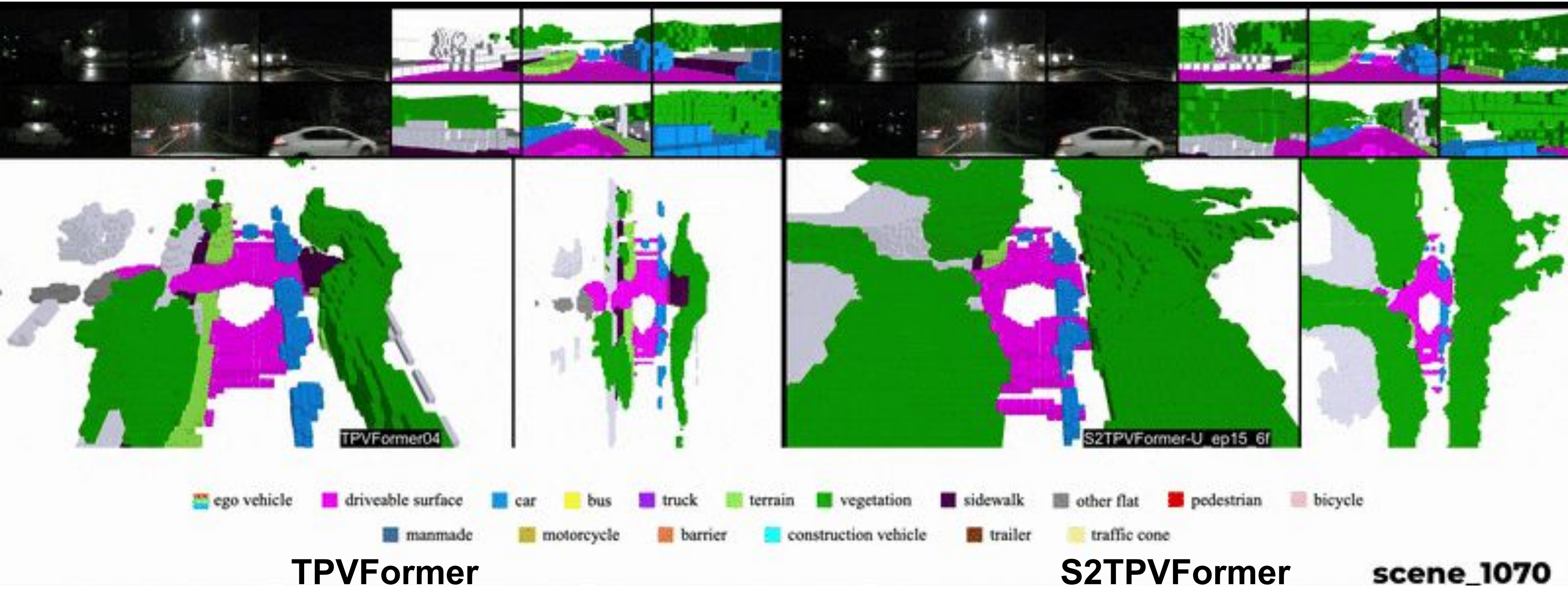
LidarSeg results on the nuScenes test set.

Thank You!

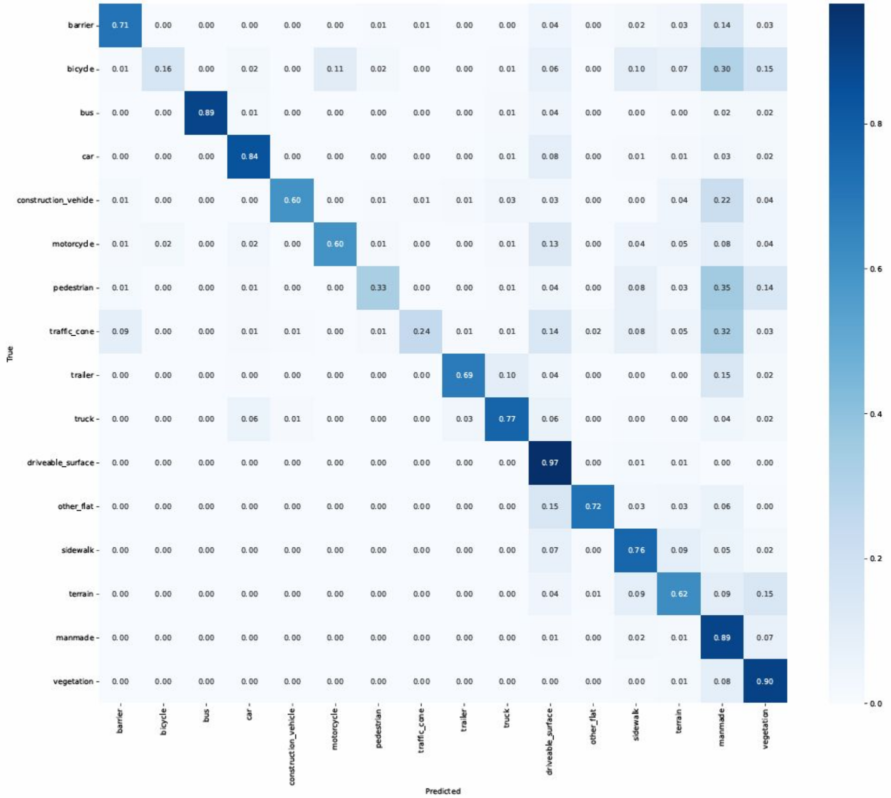
⬆️ Link to our project page

Visualization

- nuScenes: *1070, 0905, 0904, 0562*

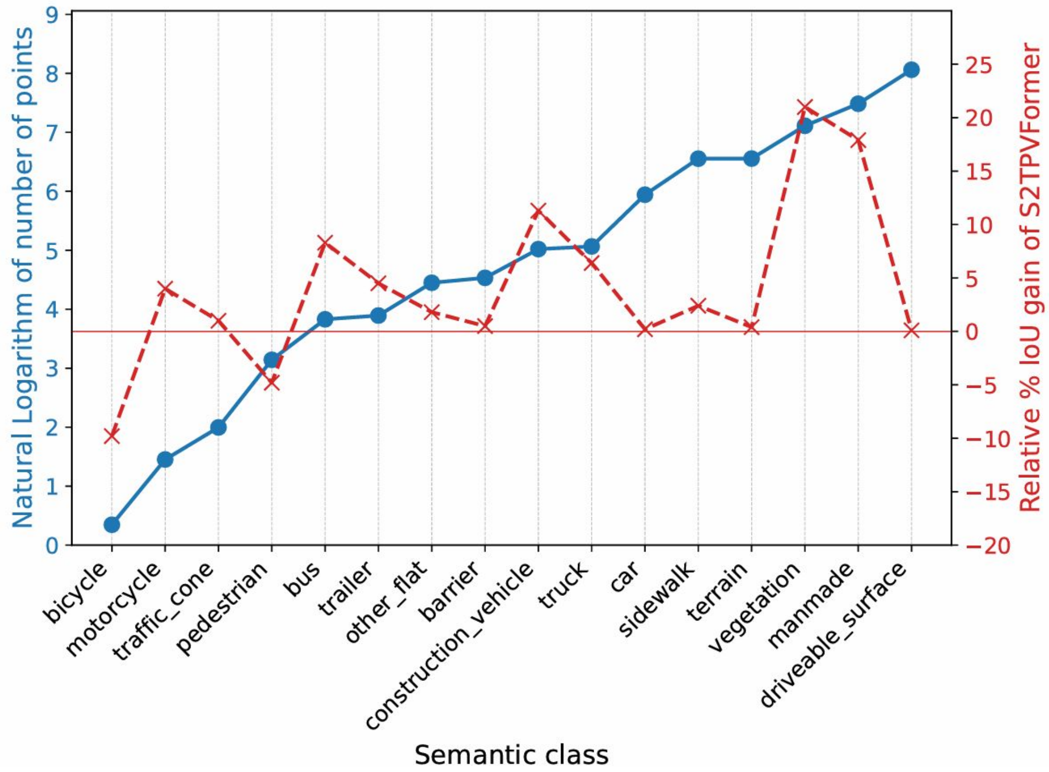


Prediction Summary



- This figure presents the confusion matrix of the S2TPVFormer (base) model's predictions.
- This confusion matrix corresponds to the same predictions analyzed in our paper, where we detail the per-class IoUs and the mean IoU for 3D Semantic Occupancy Prediction (SOP) on the nuScenes validation dataset.

Relative mIoU Gain



- This figure presents a dual-axis representation, where
  - the blue axis and its corresponding graph show the distribution of the natural logarithm of the **number of per-class ground truth points** in the training dataset.
  - Conversely, the red axis and its graph show the **per-class IoU gain** achieved by S2TPVFormer in comparison to TPVFormer.