Supplementary Materials

Anonymous Author(s)
Affiliation
Address
email

We provide additional materials for our submission. The content is organized as follows:

- Sec. 1 We present the implementation details of our framework.
- Sec. 2 We show the qualitative evaluation of label-free scene understanding and the result analysis.
- Sec. 3 We show the quantitative and quantitative evaluation of label-free scene understanding on the nuImages dataset.
- Sec. 4 We discuss the limitations of our method and present the potential improvement directions.

1 Elaborated implementation details

For the nuScenes dataset [1], the original image size is 900x1600. We resized the image to 224x416 as the input for image encoders. The learning rate is set to 0.1. Regarding the ScanNet dataset [3], the original image size is 480x640. We resized the image to 240x320 as the input for image encoders. The learning rate is also set to 0.1. In Latent Space Consistency Regularization, the output feature dimensions, denoted as K_f, are set to 64 for the linear mapping layers of the 2D function $\theta_f(\cdot)$, the 3D function $\varphi_f(\cdot)$, and the SAM function $\phi_f(\cdot)$. For Prediction Consistency Regularization, the feature dimension for image pixel, point, and text embedding is set to 512. The DeeplabV3 model used in our approach is pre-trained on the ImageNet dataset. Following MaskCLIP [6], we incorporate the class name into 85 hand-crafted prompts and input them into the CLIP text encoder. This process generates multiple text features, which are then averaged to obtain the text embedding. The prompt templates can be found in our source code under "utils/prompt_engineering.py". We provide the source code for the developed version as a reference and plan to release its final version after acceptance.

2 Qualitative Evaluation of Label-free Scene Understanding

We present a comprehensive qualitative evaluation of both 2D and 3D label-free scene understanding on the ScanNet and nuScenes datasets. As depicted in Figures 2, 3, 4, and 5, our method demonstrates remarkable performance, often comparable to human annotations. Notably, our 2D and 3D networks exhibit the ability to "segment anything" by leveraging knowledge distilled from SAM.

3 Quantitative and Qualitative Evaluation on nuImages dataset

The strength of nuScenes lies in its collection of 1000 meticulously curated scenes with 3D annotations, encompassing a wide range of challenging driving scenarios. To complement this offering, nuImages provides an additional 93,000 2D annotated images sourced from a significantly larger dataset. This includes 67,279 images for training, 16,445 for validation, and 9,752 for testing. The dataset consists of 11 shared classes with nuScenes. NuImages offers both past and future camera images, resulting in a comprehensive collection of 1,200,000 camera images. To evaluate the

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Table 1: Comparison (mIoU) with current state-of-the-art label-free methods for semantic segmentation tasks on the ScanNet [3], nuImages and nuScenes [4] dataset.

Methods	Publication	ScanNet		nuImages	nuScenes
		$2D$	$3D$	$2D$	$3D$
MaskCLIP [6]	ECCV 2022	17.3	14.2	14.1	12.8
MaskCLIP++ [6]	ECCV 2022	20.3	21.6	17.3	15.3
OpenScene [5]	CVPR 2023	14.2	16.8	12.4	14.6
CLIP2Scene [2]	CVPR 2023	23.7	25.6	18.6	20.8
Ours	–	**28.4**	**33.5**	**22.1**	**26.8**

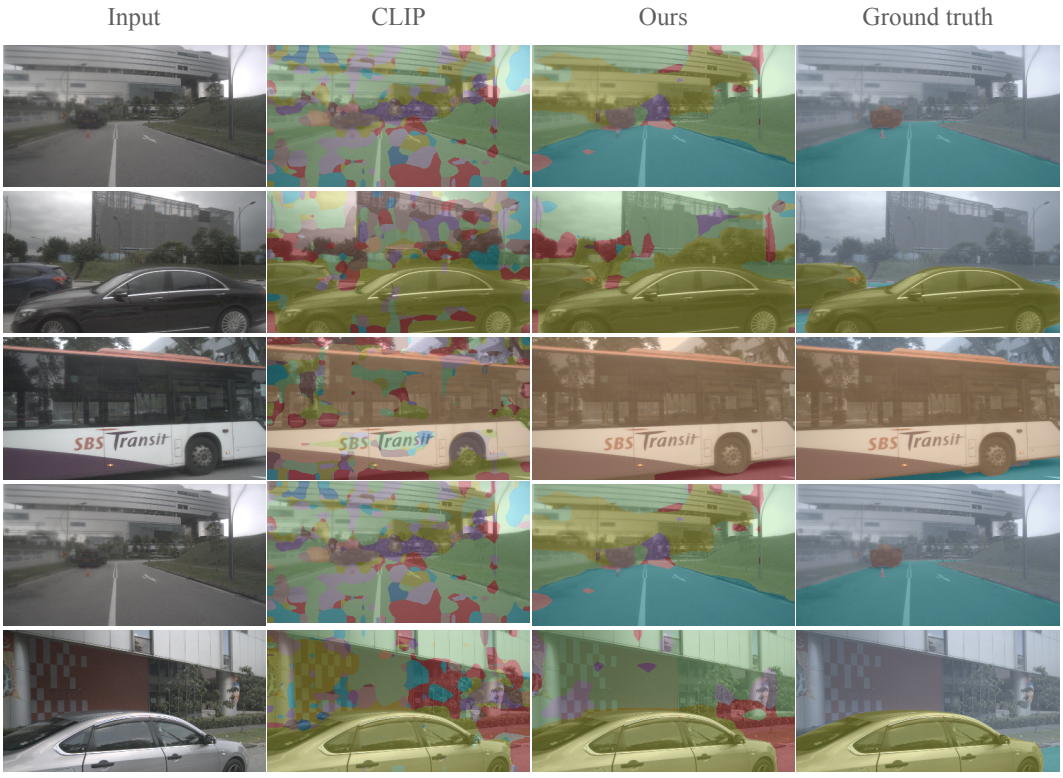

Figure 1: Qualitative results of label-free 2D semantic segmentation on nuImages dataset. From the left to the right are the input, prediction by CLIP, our full method, and ground truth, respectively. Note that the grey colour in ground truth indicates unmarked areas.

performance of our method, we conducted a comparison on a subset of 16,445 validation images. The results, as illustrated in Table 1, demonstrate that our method significantly outperforms other state-of-the-art approaches. As shown in Fig. 1, we also present the qualitative results on nuImages dataset.

4 Limitations and Future Work

While our work has achieved impressive label-free scene understanding performance, it still encounters limitations in certain cases. One notable example is the failure to recognize "derivable surface" as depicted in Fig. 5. This shortcoming can be attributed to the fact that "derivable surface" is not a concept that is relatively common in CLIP, the framework we utilize. To potentially improve the performance, we plan to explore alternative descriptions such as "road" to generate text embeddings. This direction will be a focus of our future research efforts.

Figure 2: Qualitative results of label-free 2D semantic segmentation on ScanNet dataset. From the left to the right are the input, prediction by SAM, prediction by CLIP, ours w/o CNS, our full method, and ground truth, respectively.

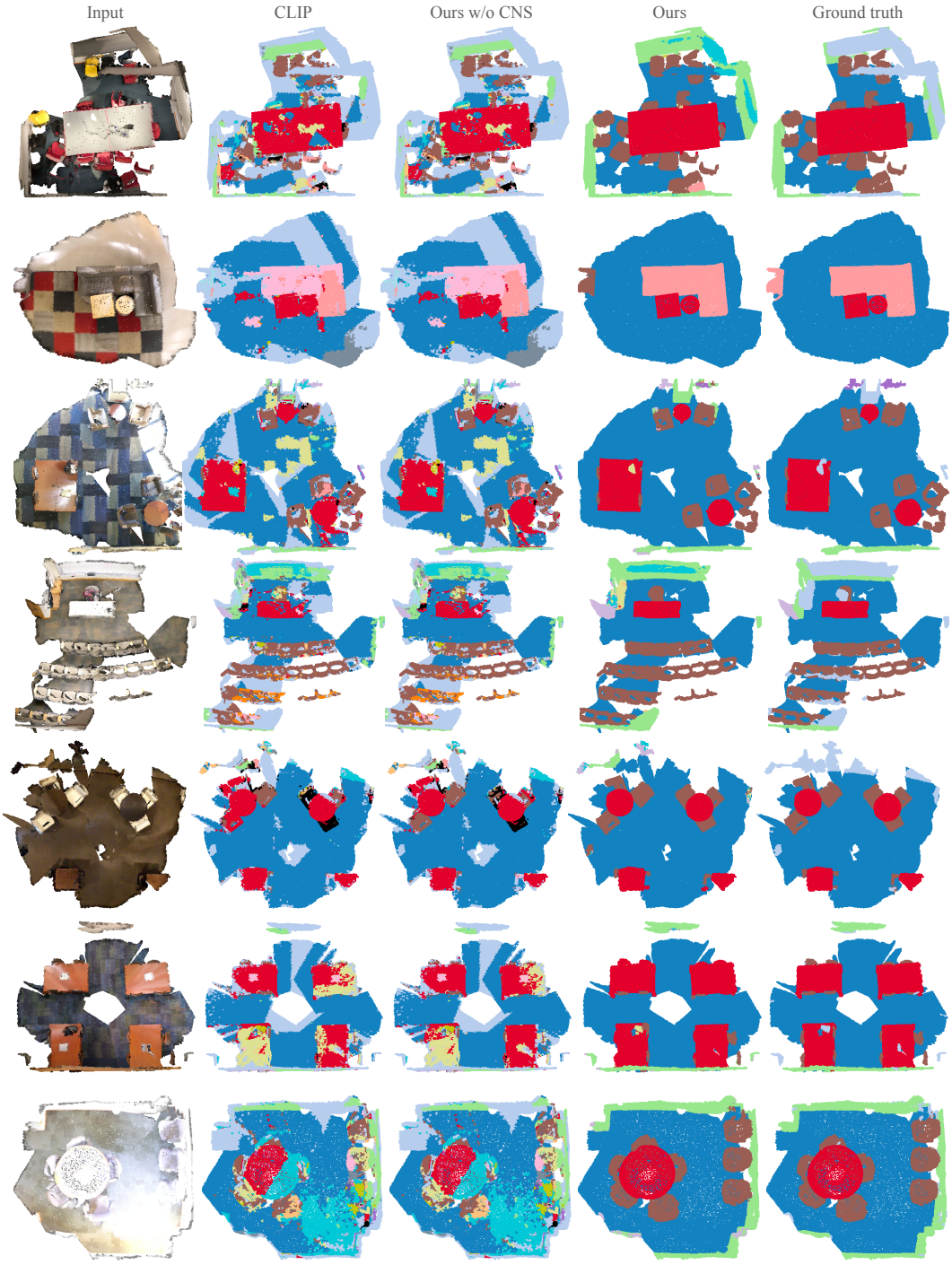

Figure 3: Qualitative results of label-free 3D semantic segmentation on ScanNet dataset. From the left to the right are the input, prediction by CLIP, ours w/o CNS, our full method, and ground truth, respectively. Note that the grey colour in ground truth indicates unmarked areas.

Figure 4: Qualitative results of label-free 2D semantic segmentation on nuScenes dataset. From the left to the right are the input, prediction by CLIP, prediction by CLIP, ours w/o CNS, our full method, respectively.

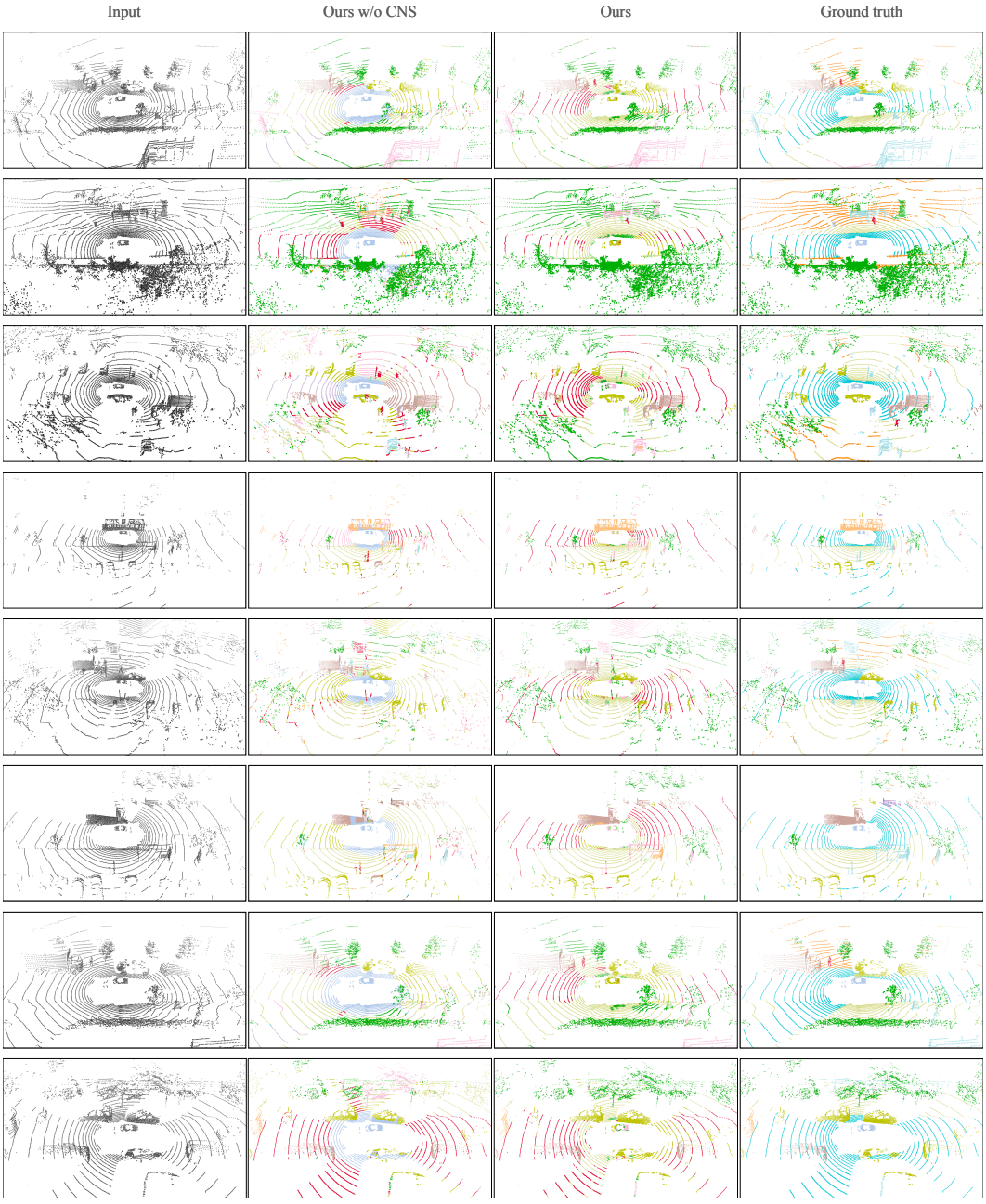

Figure 5: Qualitative results of label-free 3D semantic segmentation on nuScenes dataset. From the left to the right are the input, ours w/o CNS, our full method, and ground truth, respectively.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020.

[2] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. *arXiv preprint arXiv:2301.04926*, 2023.

[3] Angela Dai, Angel Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017.

[4] Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *IEEE Robotics and Automation Letters*, 7:3795–3802, 2022.

[5] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. *arXiv preprint arXiv:2211.15654*, 2022.

[6] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712, 2022.