3D Open-Vocabulary Panoptic Segmentation with 2D-3D Vision-Language Distillation

Zihao Xiao[1*], Longlong Jing[2], Shangxuan Wu[2], Alex Zihao Zhu[2], Jingwei Ji[2], Chiyu Max Jiang[2], Wei-Chih Hung[2], Thomas Funkhouser[3], Weicheng Kuo[4], Anelia Angelova[4], Yin Zhou[2], and Shiwei Sheng[2*]

[1] Johns Hopkins University, [2] Waymo, [3] Google Research, [4] Google DeepMind

1 PFC Baseline

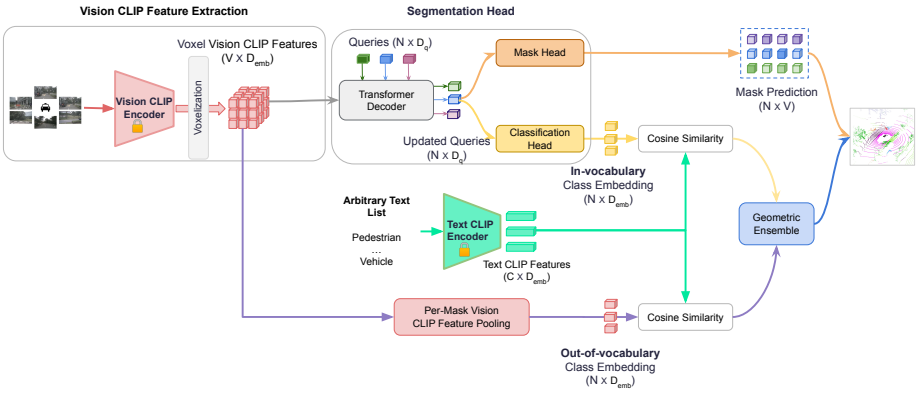


Fig. 1: The overview of PFC baseline. The PFC baseline takes frozen voxel vision CLIP features as inputs. We modified the segmentation head of a closed-set model so that it can predict in-vocabulary class embedding. The classification score is predicted by computing the cosin similarity between the predicted embedding and the text CLIP embedding. During testing, we further apply per-mask vision feature pooling to obtain out-of-vocabulary class embedding. The final per-mask classification logits are the geometric ensemble of in-vocabulary and out-of-vocabulary classification results.

As there is no existing work for the 3D open vocabulary panoptic segmentation task, a natural idea would be to extend the 2D state-of-the-art open vocabulary segmentation to 3D. We start by adapting the essential components from FC-CLIP [7], a 2D open-vocabulary segmentation model that achieves state-of-the-art performances across different datasets, to P3Former [4], a state-of-the-art 3D closed-set panoptic segmentation model. Since the models of FC-CLIP and

∗ Work done while at Waymo

P3Former are very different, we conduct some necessary changes to the architecture of P3Former. We name this baseline PCFormer+FC-CLIP (PFC). The overall architecture of the PFC is shown in Fig. 1.

Vision CLIP feature extraction. FC-CLIP [7] demonstrates that frozen CLIP features can produce promising classification performance on both base and novel classes. In the same spirit, we construct a Vision CLIP feature extractor as follows: a pre-trained V-L segmentation model [1] is applied to extract pixel-wise CLIP features from each camera image. Within each voxel, every LiDAR point is projected into its corresponding camera image based on the intrinsic and extrinsic calibration parameters, in order to index into the corresponding vision CLIP features. The vision CLIP features of all the points belonging to each voxel are then averaged to represent that voxel. The voxel CLIP features will be referred as $F_{vclip} \in \mathbb{R}^{V \times D_{emb}}$, where V is the number of voxels after voxelization and D_{emb} is the dimension of the CLIP features. Note that the Vision CLIP encoder is frozen and it is identical to the one in our proposed method.

Segmentation head. We use one learnable querie q to represent each instance or thing. Queries matched with groudtruth objects are supervised with both classification loss and mask loss. FC-CLIP [7] shows that the mask generation is class-agnostic, and therefore we follow FC-CLIP and only modify the classification head to add a class embedding. Specifically, the class embedding f_{cls} prediction is defined as:

$$v_q = f_{cls}(q) \in \mathbb{R}^{D_{emb}}, \tag{1}$$

where v_q is in the CLIP embedding space. The predicted class logits are then computed from the cosine similarity between the predicted class embedding and the text embedding of every category name from the evaluation set using a frozen CLIP model. The classification logits are defined as:

$$s_{v_q} = \frac{1}{T}[\cos(v_q, t_1), \cos(v_q, t_2), \ldots, \cos(v_q, t_C)] \tag{2}$$

where $t_i \in \mathbb{R}^{D_{emb}}$, $i \in \{1, 2, \ldots, C\}$ is the text embedding, C is the number of categories (C_B in training and C_B+C_N in testing), and T is a learnable temperature term that controls the concentration of the distribution. Following FC-CLIP, we name this trainable classifier the **in-vocabulary** classifier. The loss function, then, is $L = w_\alpha * L_{cls} + w_\beta * L_{mask}$, where L_{cls} and L_{mask} are the softmax cross-entropy classification loss and mask loss, respectively. w_α and w_β are weights for classification loss and mask loss, respectively. Note that, for classification, we apply a softmax cross-entropy loss instead of focal loss because of the following ensembling process.

Geometric ensemble. Previous open-vocabulary works [1–3, 5, 7] show that a trainable in-vocabulary classifier fails to make good predictions for novel classes. During testing, we follow [5, 7], and construct an **out-of-vocabulary** classifier that utilizes voxel Vision CLIP features to get an embedding for each query q by mask pooling the Vision CLIP features:

$$w_q = \frac{1}{|M_q|} \sum_p \mathbb{1}(p \in M_q) F_{vclip}(p) \tag{3}$$

, where M_q is the set of points, p, belonging to the mask for query, q. The out-of-vocabulary classification logits s_{w_q}can be computed as

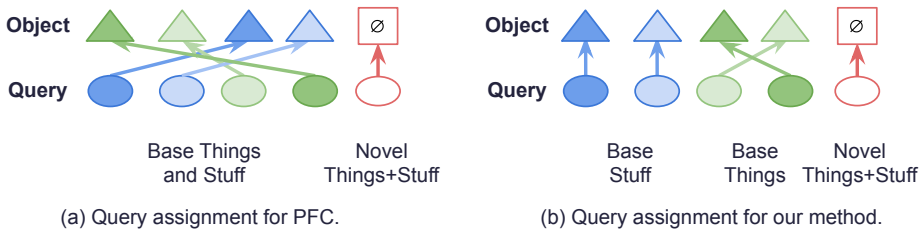
$$s_{w_q} = \frac{1}{T}[\cos(v_q, t_1), \cos(v_q, t_2), \ldots, \cos(v_q, t_C)] \tag{4}$$

where the temperature term T is the same is the one in Eq. (2). Note that the out-of-vocabulary classifier is frozen and is only applied during testing. The final classification score is computed as the geometric ensemble of the in-vocabulary classifier and out-of-vocabulary classifier for every class, i:

$$s_{g_q}(i) = \begin{cases} p_{v_q}(i)^{1-\alpha} p_{w_q}(i)^{\alpha} & \text{if } i \in C_B \\ p_{v_q}(i)^{1-\beta} p_{w_q}(i)^{\beta} & \text{if } i \in C_N \end{cases} \tag{5}$$

where $p_{v_q} = \text{softmax}(s_{v_q})$, $p_{w_q} = \text{softmax}(s_{w_q})$ are the derived probabilities and $\alpha, \beta \in [0, 1]$ are hyperparameters to control the contributions of in-vocabulary classifier and out-of-vocabulary classifier. In practice, we try multiple pairs of α, β and report the result of the best pair. We have found that $\alpha = 0$ and $\beta = 1$ generates the best results for the PFC baseline in all different base/novel splits, which indicates that the baseline solely relies on out-of-vocabulary classifier to make predictions for novel classes.

2 Query Assignment



(a) Query assignment for PFC. (b) Query assignment for our method.

Fig. 2: Visualization for the two strategies for query assignment.

For both the baseline and our method, a single query is used to represent an individual object. This requires specific query assignment strategies to match predictions with groundtruth base objects during training. FC-CLIP uses one set of learnable queries to make predictions for base and novel classes. Therefore, the same set of queries are matched with base *thing* and *stuff* objects. The unmatched queries are potentially in charge of making predictions for novel *thing* and *stuff* objects, as shown in Fig. 2 (a). In contrast, our method uses two sets of queries. The first query set is used to represent base *things* classes after bipartite matching, while the second, fixed, query set is for base *stuff* classes, as shown in Fig. 2 (b). The separation of base *things* queries and base *stuff* queries makes our model converge faster and improves overall performance.

3 Intuition of Voxel-level Distillation Loss

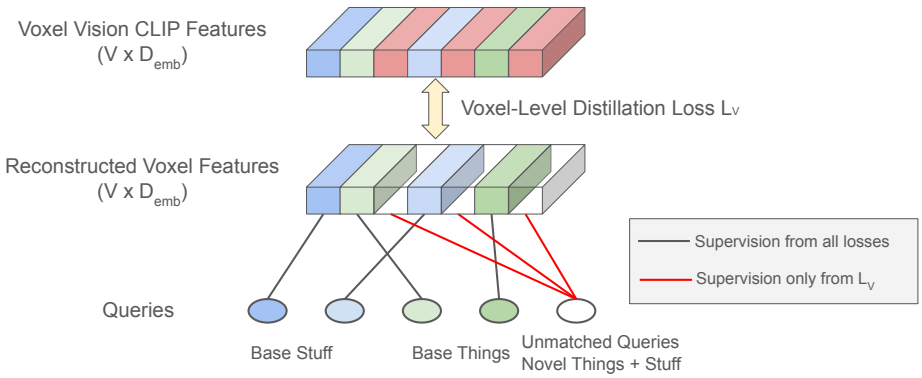


Fig. 3: Intuition of the voxel-level distillation loss (L_V). L_V does not require any labels during training and can be optimized on all queries and voxeles. Therefore, it encourages the unmatched queries to make predictions in the voxels that have no supervision from the base classes.

Fig. 3 illustrates the intuition behind the proposed voxel-level distillation loss L_V. L_V is not dependent on any labels, and therefore, can be applied to all the queries. For voxels belonging to base classes, all loss functions will be enforced during optimization, including the two standard loss functions (per-query mask classification loss L_{cls} and per-query mask prediction loss L_{mask}), the proposed object-level distillation loss L_O and the proposed voxel-level distillation loss L_V. For voxels belonging to queries that do not match to any labels, likely being novel *things* or novel *stuff* objects, they will be mainly supervised by the proposed voxel-level distillation loss L_V. With the help of this loss, the unmatched queries learn to make predictions in the voxels with no supervision from the base classes. In this way, we enforce the supervision on all the queries and voxels and the model can learn to produce meaningful predictions for both base and novel categories.

4 More Experimental Results

Queries in object-level distillation loss. In our loss function design for the object-level distillation loss L_O, we only enforce constraints on queries matched with base classes. One natural question would be: can we apply the constraint on all queries to improve predictions? We conduct an ablation study for this, with results shown in Tab. 1. We consider PQ as the most important metric. When we apply the object-level distillation loss to all queries, the overall performance is slightly worse, especially for the novel *stuff* classes.

More splits. In order to show that our proposed method generalizes well in different scenarios, we conduct experiments on two more random B12/N4 splits.

Table 1: Impact of queries in L_O. We conduct an ablation study comparing applying L_O to matched queries vs all queries. The overall performance is better when we only apply L_O on matched queries.

Queries in L_O	PQ	PQ^{Th}_N	PQ^{St}_N	RQ	RQ^{Th}_N	RQ^{St}_N	SQ	SQ^{Th}_N	SQ^{St}_N	mIoU
Matched Only	**62.0**	49.6	**35.2**	**70.9**	55.6	**46.0**	**87.0**	**89.1**	76.7	60.1
All	61.0	**49.9**	25.4	70.0	**56.3**	34.4	86.3	74.3	**88.7**	**60.5**

Table 2: Performance of panoptic segmentation on nuScenes with a different split. We compare the performance with a different split with 4 novel classes (B12/N4). The novel *things* classes are construction vehicle and traffic cone. The novel *stuff* classes are other-flat and man-made. Our method consistently outperforms the PFC baseline across almost all the metrics by a large margin.

Model	Type	Supervision	PQ	PQ^{Th}_N	PQ^{St}_N	RQ	RQ^{Th}_N	RQ^{St}_N	SQ	SQ^{Th}_N	SQ^{St}_N	mIoU
P3Former [4]	closed-set	full	75.8	76.4	86.9	83.8	84.8	98.3	90.1	89.8	88.4	78.2
PFC	open-voc	partial	49.9	22.5	14.0	59.6	**26.9**	21.9	85.6	82.9	61.0	53.8
Ours	open-voc	partial	**55.4**	**23.2**	**24.7**	**62.9**	26.0	**29.8**	**85.6**	**87.6**	**69.3**	**55.0**

As shown in Tab. 2 and Tab. 3, our method surpasses the PFC baseline in almost all metrics across all the splits, demonstrating the capability of our proposed method.

Performance on novel *stuff* classes. The performance of PFC baseline is almost 0 on novel *stuff* classes. To verify whether it is due to poor mask predictions for the novel stuff calss, we conduct an oracle experiment by max-pooling vision CLIP features with ground truth masks and then use its similarity with CLIP text features to determine its category, and the results are shown in Tab. 4. We achieve 53 RQ using ground truth mask, which demonstrate that the bad performance of PFC baseline is indeed due to poor mask quality. Also, the low RQ number shows that the prediction task on novel stuff class is very challenging.

5 Discussion

Class-agnostic mask generator. As shown in FC-CLIP [7], the mask head is class-agnostic if we do not apply any penalty to unmatched queries. We follow the same strategy in our paper. The metrics SQ^{Th}_N and SQ^{St}_N in all experiments indicate that the mask predictions for both *things* and *stuff* are reasonable.

Comparison with RegionPLC. RegionPLC [6] proposes to take advantage of regional visual prompts to create dense captions. After point-discriminative contrastive learning, the model can be used for semantic segmentation or instance segmentation. There are two main differences between RegionPLC and our method: 1. RegionPLC addresses the problem of semantic segmentation or instance segmentation individually, while our model addresses semantic segmentation and instance segmentation in the same model. 2. RegionPLC focuses on

Table 3: Performance of panoptic segmentation on nuScenes with another different split. We compare the performance with a different split with 4 novel classes (B12/N4). The novel *things* classes are barrier, bus and truck. The novel *stuff* class is drivable surface. Our method consistently outperforms the PFC baseline across almost all the metrics by a large margin.

Model	Type	Supervision	PQ	PQ^{Th}_{N}	PQ^{St}_{N}	RQ	RQ^{Th}_{N}	RQ^{St}_{N}	SQ	SQ^{Th}_{N}	SQ^{St}_{N}	mIoU
P3Former [4]	closed-set	full	75.8	71.1	96.2	83.8	78.8	99.9	90.1	90.1	96.3	78.2
PFC	open-voc	partial	43.1	16.4	2.1	51.6	20.0	3.0	79.8	83.8	**69.7**	43.5
Ours	open-voc	partial	**53.1**	**31.0**	**35.1**	**63.0**	**35.2**	**53.8**	**82.3**	**87.7**	65.3	**50.5**

Table 4: Performance on novel *stuff* classes. We compare the performance of PFC, our method and the oracle setting that based on the GT masks on novel *stuff* classes. The spilits are the same as in Tab. 1 of the main paper.

	mIoU	PQ	RQ	SQ
PFC	4.38	0.5	0.83	60.44
Ours	45.14	35.25	45.97	76.69
Oracle (GT Masks)	51.39	52.61	53.00	99.26

getting point-level discriminative features, while our model takes the pretrained CLIP features as input and aims to build model architecture and design loss functions. In our method, we do not compare with RegionPLC because the experiment settings are different and there is no public code to reproduce the contrastive learning process. However, we do think there is great potential in combining RegionPLC and our method. One idea would be to replace the vision CLIP features in our model with the features derived from RegionPLC.

6 Visualization

We present the visualization of PFC baseline, our method and groundtruth in Fig. 4 and Fig. 5. Note that we only visualize the points that are visible in frontal camera views in Fig. 5.

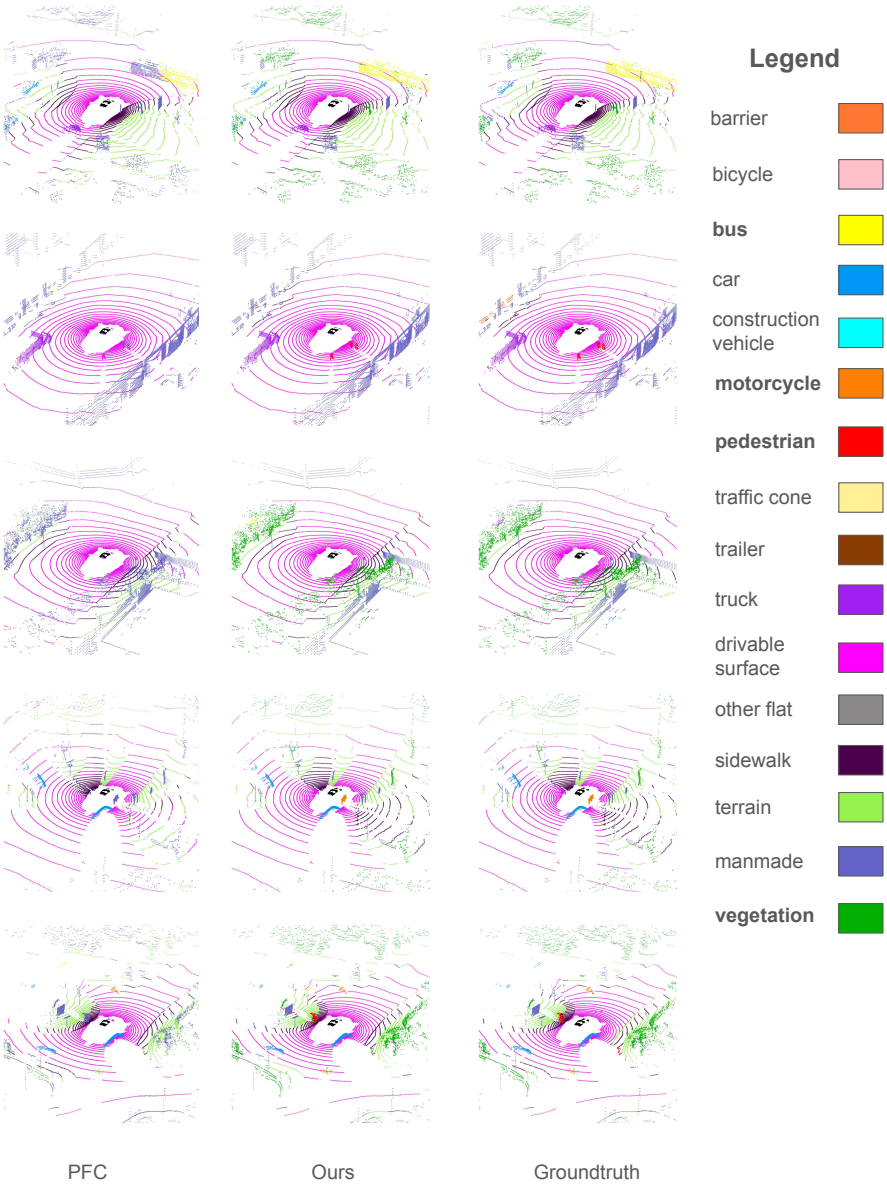


Fig. 4: Qualitative Results in nuScenes Dataset. We present the comparison among PFC, our method and the groundtrth. The novel objects are marked in **bold** in the legend.

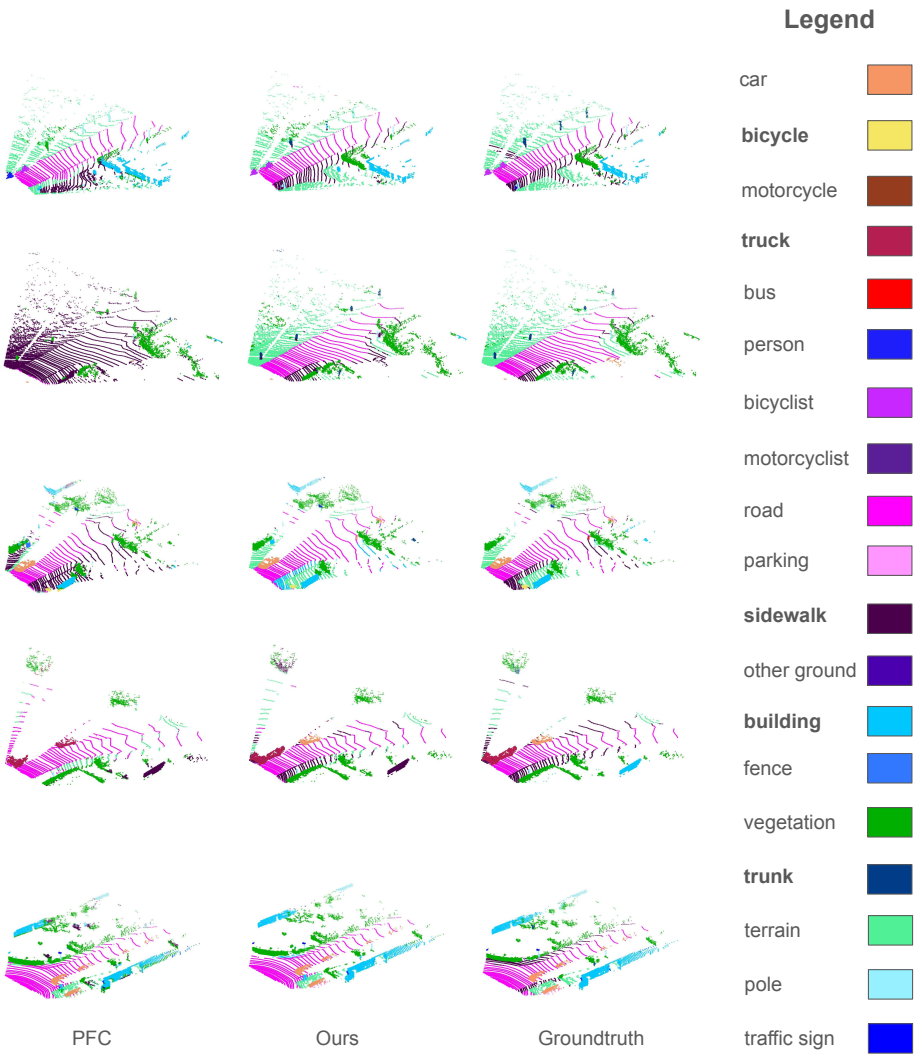


Fig. 5: Qualitative Results in SemanticKITTI Dataset. We present the comparison among PFC, our method and the groundtrth. The novel objects are marked in **bold** in the legend.

References

1. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: ECCV (2022)
2. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. ICLR (2022)
3. Kuo, W., Cui, Y., Gu, X., Piergiovanni, A., Angelova, A.: F-vlm: Open-vocabulary object detection upon frozen vision and language models. In: ICLR (2023)
4. Xiao, Z., Zhang, W., Wang, T., Loy, C.C., Lin, D., Pang, J.: Position-guided point cloud panoptic segmentation transformer. arXiv preprint (2023)
5. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: CVPR (2023)
6. Yang, J., Ding, R., Wang, Z., Qi, X.: Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. In: CVPR (2024)
7. Yu, Q., He, J., Deng, X., Shen, X., Chen, L.C.: Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In: NeurIPS (2023)