4D Panoptic Segmentation as Invariant and Equivariant Field Prediction (Appendix)

Minghan Zhu[1,*] Shizhong Han[2] Hong Cai[2] Shubhankar Borse[2]
Maani Ghaffari[1,*] Fatih Porikli[2]

[1]University of Michigan, Ann Arbor [2]Qualcomm AI Research[†]

1. Visualization of Offset Prediction

The offset prediction as an equivariant vector field is a main factor in the significant improvement achieved by our proposed Eq-4D-StOP model. In Fig. 6, we visualize the offset predictions to show this improvement intuitively. We can see that the offset vectors predicted by our equivariant model have more consistent orientations and the end points are closer to the instance center, thus benefiting the object clustering and segmentation.

2. 3D Panoptic Segmentation Performance

While the main focus of this paper is 4D panoptic segmentation, the network structure is also compatible with the 3D panoptic segmentation task by skipping the point cloud aggregation step and only taking a single scan of point cloud as input. In the 3D panoptic segmentation task, we keep the model and training configurations the same as in Sec. 5.2, except for inputting a single frame of point cloud during training and inference. In Tab. 6, we show the performance of our model compared with the baseline. The metrics follow the 2D [2, 5] and 3D [4, 1] panoptic segmentation literature. PQ, the panoptic quality, measures the overall accuracy of panoptic segmentation. $PQ = SQ \times RQ$, where RQ, the recognition quality, measures the ratio of successful instance segmentation with $IoU > 0.5$, and SQ measures the segmentation quality by the average IoU across the successfully segmented instances. The superscripts Th and St refer to the *things* classes and *stuff* classes, as in the 4D metrics. The semantic segmentation accuracy is measured by $mIoU$.

From Tab. 6, we can see that the performance of our Eq-4D-StOP model improves over the non-equivariant baseline in all metrics, which shows that the equivariance property also benefits the 3D panoptic segmentation task. Especially, PQ^{Th} is increased by 4.0 points, showing that the instance segmentation of objects is majorly improved, consistent with our observations in Sec. 5.2 in the 4D segmentation.

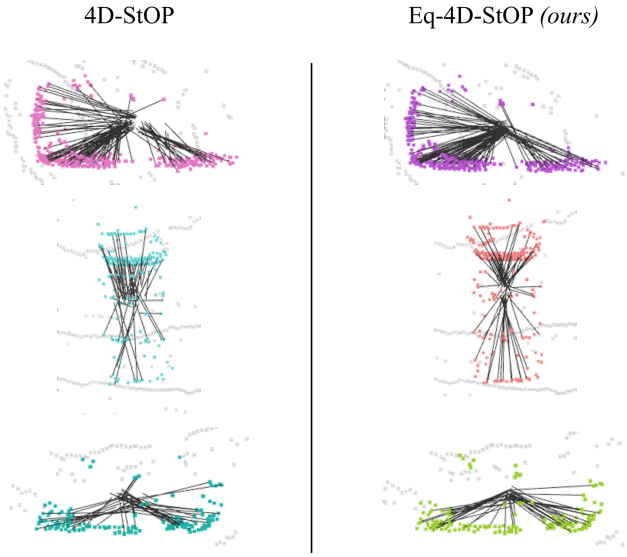

Figure 6. Qualitative comparison of the offset vector prediction (the black line segments) between our method and the baseline. The predictions from our equivariant model are more consistent and the endpoints are more concentrated near the instance centers.

3. Ablation: Rotation Classification for Offset Prediction without Equivariant Features

Besides the benefits brought by the equivariance, there could be another hypothesis for the performance improvement in Eq-4D-StOP: With the rotation classification, it could be easier to regress the offset vector. It can be explained as follows. As discussed in Sec. 4.3.2, for a point $x \in \mathbb{R}^3$ with arbitrary target offset vector $v \in \mathbb{R}^3$, its corresponding orientation is $\theta(v) = \text{atan2}(v_Y, v_X)$. Here we slightly abuse the notation to use θ to represent both the angle and the corresponding rotation matrix. It should

[*]Work done at Qualcomm AI Research during an internship.
[†]Qualcomm AI Research, an initiative of Qualcomm Technologies, Inc.

Method	PQ	$PQ^{\dagger}$	SQ	RQ	PQ^{Th}	SQ^{Th}	RQ^{Th}	PQ^{St}	SQ^{St}	RQ^{St}	$mIoU$
4D-StOP [3]	58.5	64.0	80.3	68.2	62.1	91.0	67.8	56.0	72.5	68.6	64.6
Eq-4D-StOP *(ours)*	61.2	66.2	83.6	70.8	66.1	91.3	71.9	57.5	78.0	70.0	68.0

Table 6. 3D Panoptic segmentation results on SemanticKITTI validation set.

Method	4D-StOP ($c = 256$)	Eq-4D-StOP ($c = 64, n = 4$)	4D-StOP w/ *R-head* ($c = 256, n = 4$)
$LSTQ$	67.1	**69.8**	66.8

Table 7. Experiment of standard KPConv and proposed prediction head for equivariant field prediction (*R-head*) on SemanticKITTI.

not cause ambiguity since all rotations are in $\mathrm{SO}(2)$ in this discussion. The ground truth rotation anchor for vector v is $\theta_{i(v)} \in \mathrm{SO}(2)'$, where $i(v) = \arg\min_i \|\theta_i - \theta(v)\|$. Following Eq. (4) and (6), the learning process is to fit $f(x, \theta_{i(v)})$, the prediction at the $i(v)$'th rotation anchor, to $\theta_{i(v)}^{-1} v$. Intuitively speaking, it means that the offset is always regressed in the local reference frame (rotation anchor) closest to the orientation defined by the offset itself. The variation of v in its closest local reference frame is much smaller than v in the global frame. Specifically, $\theta(\theta_{i(v)}^{-1} v) = \theta_{i(v)}^{-1}\theta(v) \in [-\frac{\pi}{n}, \frac{\pi}{n})$, for $SO(2)' \cong C_n$ with n discretized rotation anchors. In comparison, $\theta(v) \in [-\pi, \pi)$, which implies that the regression of $\theta_{i(v)}^{-1} v$ could be easier.

We test out this hypothesis by experimenting with a network that uses the non-equivariant KPConv [6] backbone and predicts the offset with rotation classification. The ground truth rotation anchors are defined in the same way as the equivariant models, and the target offset to be regressed is also $\theta_{i(v)}^{-1} v$ as discussed above. We call this model 4D-StOP with rotation head (R-head), as in the last column of Tab. 7, which compares the performance with the baseline and our equivariant model. The comparison uses a consistent feature map size and rotation anchor size. The experimental results show that 4D-StOP w/ R-head does not outperform the baseline, indicating that the performance improvement is brought by the equivariant property of the network instead of the smaller variations in the regression targets.

4. Quotient Representation in SO(2) Causes Information Loss

In Sec. 4.2, we introduce that we use the regular representation instead of the quotient representation in our $\mathrm{SO}(2)$ equivariant 4D panoptic segmentation network, because quotient representations cause information loss for abelian groups like $\mathrm{SO}(2)$. Here is a more detailed explanation.

First, we explain what it means to have a quotient representation that does *not* cause information loss, as is the case in E2PN [7]. E2PN is a $\mathrm{SO}(3)$-equivariant network with feature maps in the space $\mathcal{F} = \{f : \mathbb{R}^3 \times S^2 \to V\}$, where $S^2 = \mathrm{SO}(3)/\mathrm{SO}(2)$ is the 2D sphere in 3D space, and also the quotient space of $\mathrm{SO}(3)$ with respect to subgroup $\mathrm{SO}(2)$. As a $\mathrm{SO}(3)$-equivariant network, its feature maps are not defined on $\mathrm{SO}(3)$ but only S^2, which is why it is said to use a *quotient representation* to reduce the feature map size and thus the computational cost. The reason that this quotient representation does not cause information loss is that the group action of $\mathrm{SO}(3)$ on S^2 (i.e., the 3D rotation of a sphere) is *faithful*, which is to say the only rotation in $\mathrm{SO}(3)$ that keeps all points on a sphere unchanged is the identity rotation. It implies that any $\mathrm{SO}(3)$ rotation can be detected from its action on the S^2 feature maps, therefore not losing any information in $\mathrm{SO}(3)$.

Put more formally, we denote the group as G, the subgroup as H, the quotient space as G/H. The group actions of G on G/H is a group homomorphism $\phi : G \to Aut(G/H)$. If the group action is faithful, then the kernel of the homomorphism is $ker(\phi) = \{e\}$, only containing the identity element. By the first isomorphism throrem, $G/ker(\phi) = G \cong Im(\phi)$. That is to say, ϕ is injective. Therefore, there exists an inverse map $\phi^{-1} : Aut(G/H) \supset Im(\phi) \to G, \phi(g) \mapsto g$. We can determine the group element $g \in G$ from the automorphism in the quotient space G/H, thus we say the information of G is fully preserved in G/H.

However, for $\mathrm{SO}(2)$, which is an abelian group, its action on its quotient space is *not* faithful. To see this, we still use the G to denote $\mathrm{SO}(2)$ and H to denote a subgroup of G. An element in the quotient space G/H can be denoted as gH for some $g \in G$. The group action of G on G/H is $g' \mapsto (gH \mapsto g'gH, \forall g \in G)$. Now if we take $g' = h \in H$, then with the abelian property of G, we have $g'gH = hgH = ghH = gH$, meaning that the action of $g' \neq e$ on G/H keeps all elements in G/H unchanged. Therefore, the group action of $\mathrm{SO}(2)$ on its quotient space is not faithful, and $ker(\phi) = H$.

By the first isomorphism throrem, $G/ker(\phi) \cong Im(\phi) \subset G/H$. From G/H, we can only recover elements in $G/ker(\phi) = G/H$ instead of G, therefore the information inside an H-coset is lost.

Here we provide a concrete example in the discretized case. Consider $\mathrm{SO}(2)$ discretized as C_6, i.e., the set composed of 60-degree rotations. If we take a subgroup C_2 (i.e., 180-degree rotations), then the quotient space is $C_6/C_2 =$

C_3. From the quotient features C_3, we will lost discrimination among the C_2-coset. In other words , any rotation angle θ and $\theta + 180°$ corrrespond to the same quotient feature maps in C_3.

Therefore, we use the regular representation instead of the quotient representation. In other words, to enable C_n-equivariance, we use a feature map defined on C_n as well.

5. Nearest-Neighbor Upsampling and 1-by-1 Convolution Are Equivariant

Nearest-neighbor upsampling layer For the nearest-neighbor upsampling layer, denote a coarse-level feature map as $f_{coarse} \in \mathcal{F}$ and a fine-level feature map as $f_{fine} \in \mathcal{F}$. The nearest neighbor upsampling layer gives

$$f_{fine}(x, R) = f_{coarse}(x_{nn}, R) \tag{10}$$

where $x \in X_{fine} \subset \mathbb{R}^3$, in which X_{fine} is the fine point cloud. $x_{nn} \in X_{coarse} \subset \mathbb{R}^3$, where X_{coarse} the coarse point cloud, and x_{nn} is the nearest neighbor of x in the coarse point cloud. Since distance is preserved under rotations, the nearest neighbor for Rx in the rotated coarse point cloud RX_{coarse} is Rx_{nn}. If f_{coarse} is an equivariant feature map, i.e., satisfies Eq. (4), then we have

$$\begin{aligned}[Rf_{fine}](Rx, R') &= [Rf_{coarse}](Rx_{nn}, R') \\ &= f_{coarse}(x_{nn}, R^{-1}R') = f_{fine}(x, R^{-1}R'),\end{aligned} \tag{11}$$

which means $[f_{fine}]$ also satisfies Eq. (4), thus is equivariant.

1-by-1 convolution layer A 1-by-1 convolution is a map $W : V \to V$ which operates on $f(x, R)$ for each x and R individually. Denote an existing equivariant feature map $f_1 \in \mathcal{F}$ and the feature map after the 1-by-1 feature map $f_2 \in \mathcal{F}$, i.e.,

$$f_2(x, R) = W(f_1(x, R)) \tag{12}$$

Then we have

$$\begin{aligned}[Rf_2](Rx, R') &= W([Rf_1](Rx, R')) \\ &= W(f_1(x, R^{-1}R')) = f_2(x, R^{-1}R')\end{aligned} \tag{13}$$

showing that f_2 satisfies Eq. (4), therefore is equivariant.

References

[1] Stefano Gasperini, Mohammad-Ali Nikouei Mahani, Alvaro Marcos-Ramiro, Nassir Navab, and Federico Tombari. Panoster: End-to-end panoptic segmentation of lidar point clouds. *IEEE Robotics and Automation Letters*, 6(2):3216–3223, 2021. 1

[2] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019. 1

[3] Lars Kreuzberg, Idil Esen Zulfikar, Sabarinath Mahadevan, Francis Engelmann, and Bastian Leibe. 4d-stop: Panoptic segmentation of 4d lidar using spatio-temporal object proposal generation and aggregation. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 537–553. Springer, 2023. 2

[4] Jinke Li, Xiao He, Yang Wen, Yuan Gao, Xiaoqiang Cheng, and Dan Zhang. Panoptic-phnet: Towards real-time and high-precision lidar panoptic segmentation via clustering pseudo heatmap. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11809–11818, 2022. 1

[5] Lorenzo Porzi, Samuel Rota Bulo, Aleksander Colovic, and Peter Kontschieder. Seamless scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8277–8286, 2019. 1

[6] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019. 2

[7] Minghan Zhu, Maani Ghaffari, William A Clark, and Huei Peng. E2pn: Efficient se (3)-equivariant point network. *arXiv preprint arXiv:2206.05398*, 2022. 2