We thank all the reviewers for their efforts and thoughtful feedbacks. All reviewers agreed that we proposed a novel and interesting Cross-view Consistency Network (CVCNet) to leverage the advantages of both range view (RV) and Bird's-eye-view (BEV) in 3D detection. In the paper, we presented the state-of-the-art performance on the 3D detection benchmark. Moreover, we provided in-depth analysis and clear ablation studies to validate our contributions. We will address the issues raised by reviewers in the final version.

Experiments on extra datasets and comparisons (R1, R2) - NuScenes and Waymo are relatively new and large 3D detection datasets (50x larger than KITTI w.r.t #scenes). A lot of previous state-of-the-art algorithms, such as AVOD and PointRCNN, did not evaluate their performance on NuScenes and Waymo. In the main paper, we already include all the published results on NuScenes. Additionally, we conduct the experiments on Waymo and the performance is shown in Table 1. Our approach outperforms all one stage detectors by a large margin in overall mAP for vehicle detection.

Table 1: Vehicle Detection mAP for One-stage Detectors on Waymo OD Validation Set

Method	LEVEL 1 3D IoU=0.7			
	Overall	0-30m	30-50m	50m-∞
StarNet (3)	53.70	-	-	-
PointPillars	56.62	81.01	51.75	27.94
PPBA (1)+ PointPillars	62.44	-	-	-
MVF	62.93	86.30	60.02	36.02
AFDet (2)	63.69	87.38	62.19	29.27
CVCNet (ours)	**68.43**	**87.55**	**63.53**	**42**

Our algorithm runs at 8 FPS with a single V100 GPU on Waymo Open Dataset. MVF reported 15 FPS but they did not specify the machine they used or if they optimized to speed up. Since MVF did not release the experimental details and code, it is difficult to make comparisons with it on inference speed and number of parameters. Since MVF uses separate backbones, to show some insights on speed and number of parameters, we present our method with separate backbones for BEV and RV in Table 2. The experiments are conducted on NuScenes validation set. With comparable performances, adding one more backbone will add extra 240 ms runtime (267% more) per frame and 30 MB of paramters (18% more) in our proposed CVCNet.

Table 2: Performance with Separate or Shared Backbones on NuScenes Val Set

Backbone	FPS	#parameters	car	truck	bus	trailer	construction vehicle	pedestrian	motorcycle	bike	traffic cone	barrier	mAP
separate	3	201MB	83.1	**50.2**	59.2	33.7	16.0	81.0	**57.1**	**34.6**	60.9	**66.7**	54.2
shared	11	171MB	**83.2**	50.0	**62.0**	**34.5**	**20.2**	**81.2**	54.4	33.9	**61.1**	65.5	**54.6**

Text clarity (R3) - Thank you for your comments about text clarity. We will carefully revise our paper in the final version. We will make all the equations, as well as the figure in the paper clear and easy to understand. Meanwhile, we should say the other reviewers didn't mention any problems with the writing.

Sorry we don't fully understand your comment "calling cross-view transformers the mapping functions used in the constraint term is confusing". Did you mean L161 "map RV features to BEV space"? We did not write it as a constraint term. This is how we match voxels between two views.

We do see some researchers call output scores features (4) but we will clarify this in the final version.

With our Hybrid-Cylindrical-Spherical Voxelization, a voxel in one view corresponds to a column of voxels in the other view. This property is similar to one of the properties of Hough Transform, ie. a point in one domain correponds to a line in another domain. Our transformers are inspired by voting in Hough Transform. We understand it may not be easy to grasp this property without visualizations. We are considering making dynamic figures for people to get a better idea of it. Epipolar geometry is related in the sense of multi-view correspondence. However, epipolar geometry does not apply to our work because it assumes pinhole camera model while BEV is from orthographic projection.

[1] Cheng, S., Leng, Z., Cubuk, E.D., Zoph, B., Bai, C., Ngiam, J., Song, Y., Caine, B., Vasudevan, V., Li, C., et al.: Improving 3d object detection through progressive population based augmentation. arXiv preprint arXiv:2004.00831 (2020)

[2] Ge, R., Ding, Z., Hu, Y., Wang, Y., Chen, S., Huang, L., Li, Y.: Afdet: Anchor free one stage 3d object detection. arXiv preprint arXiv:2006.12671 (2020)

[3] Ngiam, J., Caine, B., Han, W., Yang, B., Chai, Y., Sun, P., Zhou, Y., Yi, X., Alsharif, O., Nguyen, P., et al.: Starnet: Targeted computation for object detection in point clouds. arXiv preprint arXiv:1908.11069 (2019)

[4] Vora, S., Lang, A.H., Helou, B., Beijbom, O.: Pointpainting: Sequential fusion for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4604–4612 (2020)