A Window Shift

The window shift operation (Figure 6) is implemented by adding offsets of half of the window sizes to the voxelized coordinates and then running the same window partition algorithm. We have proposed to limit the window shift operation per scale for efficiency of processing sparse inputs. We wondered what happens if we add more window shifts to the model. Will it impact model accuracy? We added one more window shift to the scale 1 and scale 2 respectively. It has slowed down our training by 10%. Surprisingly, it slightly decreased the model accuracy as shown in Table 6. Our hypothesis is that more shifts make the model harder to train when we do not need to rely on window shifts to increase receptive field.

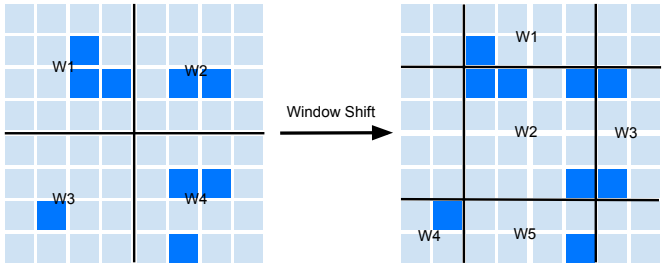

Fig. 6. Sparse window shift. The dark blue cells are voxels with points. The light blue cells are voxels without points. Left shows a grid of 8×8 BEV voxels partitioned into 4 non-empty sparse windows with window size of 4×4. After window shift, it results in 5 non-empty sparse windows as shown on the right.

Table 6. Impact of adding more window shifts.

More Window Shifts	Vehicle 3D AP/L1	Pedestrian 3D AP/L1
✗	79.36	82.91
✓	79.17	82.36

B Qualitative Results

Figure 7 visualizes ground truth boxes, detected boxes, and attention scores for layers selected from different scales for the 15th frame in scene 8907419590259234067_1960_000_1980_000 selected from the Waymo Open Dataset validation set. The selected layers are the stride 1, 2 afters multi-scale feature fusion, and stride 1, 2, 4, 16 from the main backbone. We use all foreground

points as the query points. The predicted boxes almost overlap perfectly with the ground truth boxes. The attention score pattern shown in these subplots indicates that different information is captured in different layers and scales. Interestingly, we have found that most of the attention scores are either 0 or 1 for foreground query points. We hope that these findings can inspire more research in the future.

C Future Work: More Tasks

Waymo Open Dataset [39] has recently added semantic segmentation labels for about 14% of the frames per scene for all of the 1150 scenes. We have extended the SWFormer detection network to perform joint semantic segmentation and detection. Figure 8 illustrates the joint detection and semantic segmentation network architecture. We concatenate the per-point feature from the voxel embedding net before per-voxel max pooling and its corresponding voxel feature from a selected scale after multi-scale feature fusion to predict the per-point semantic segmentation logits. Without much tuning, we have obtained reasonable semantic segmentation results as shown in Table 7 and Figure 9. We plan to further improve this model and extend it to more autonomous driving related tasks.

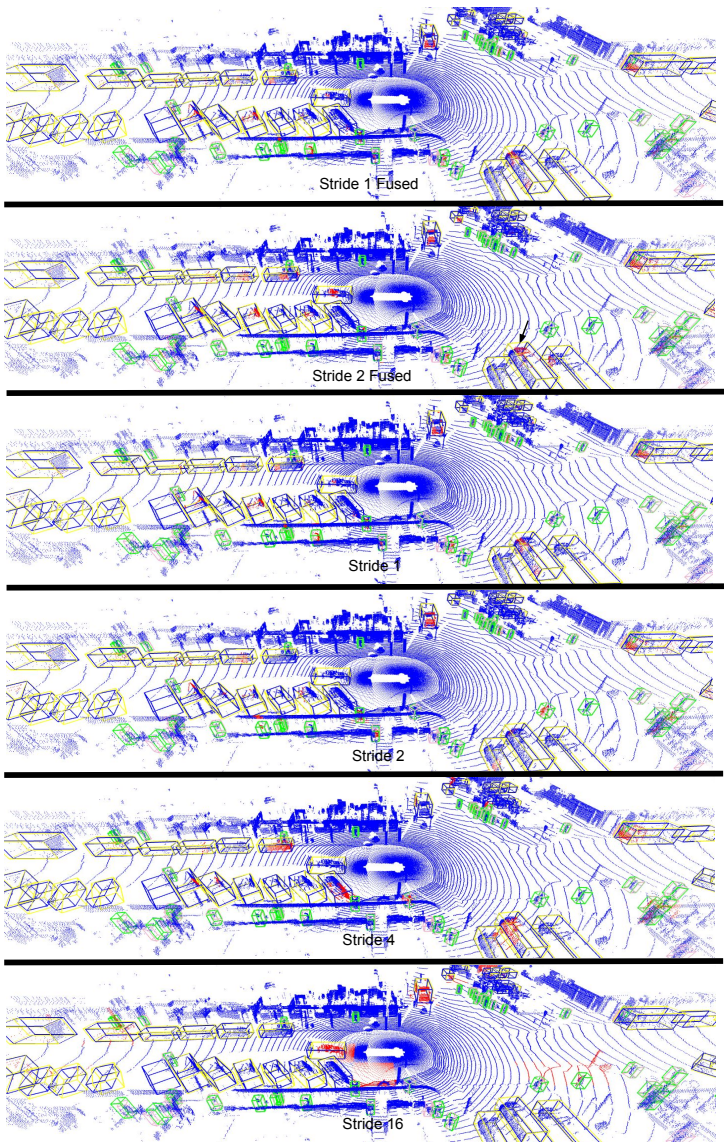

Fig. 7. Attention scores and model prediction visualization. Blue box: ground truth vehicle. Yellow box: vehicle prediction. Green box: ground truth pedestrian. Purple box: pedestrian detection. Points are colored with the bwr colormap (0: blue, 1: red), where red points mean attention scores close to 1. As red points are distributed differently in each subfigure, it is clear that different layers are attending to different locations.

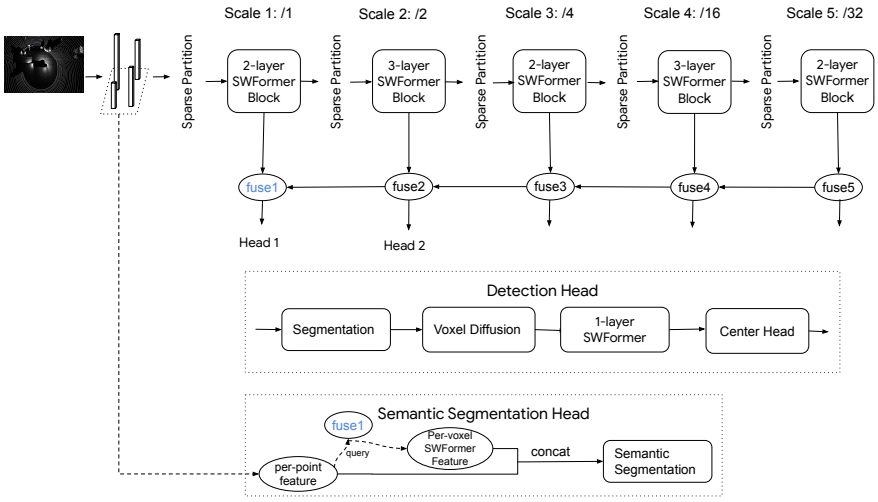

Fig. 8. Overview of the updated neural architecture for joint 3D detection and semantic segmentation. On top of Figure 1, it adds an extra segmentation head for the additional segmentation task.

Table 7. Joint detection and semantic segmentation results on Waymo Open Dataset validation set and test set.

Class Name	Validation IOU	Test IOU
Bicycle	36.76	38.15
Bicyclist	51.43	51.77
Building	75.18	65.75
Bus	65.45	39.50
Car	75.05	72.29
Construction Cone	48.34	21.37
Curb	55.54	48.46
Lane Marker	43.97	30.73
Motorcycle	56.68	58.37
Motorcyclist	1.48	0.57
Other Ground	34.34	37.52
Other Vehicle	23.95	25.43
Pedestrian	60.87	61.08
Pole	55.50	51.65
Road	78.46	68.06
Sidewalk	59.67	59.77
Sign	53.70	43.60
Traffic Light	22.74	22.30
Tree Trunk	54.74	50.64
Truck	48.73	55.86
Vegetation	79.78	68.08
Walkable	65.87	59.08
mIOU	52.19	46.82

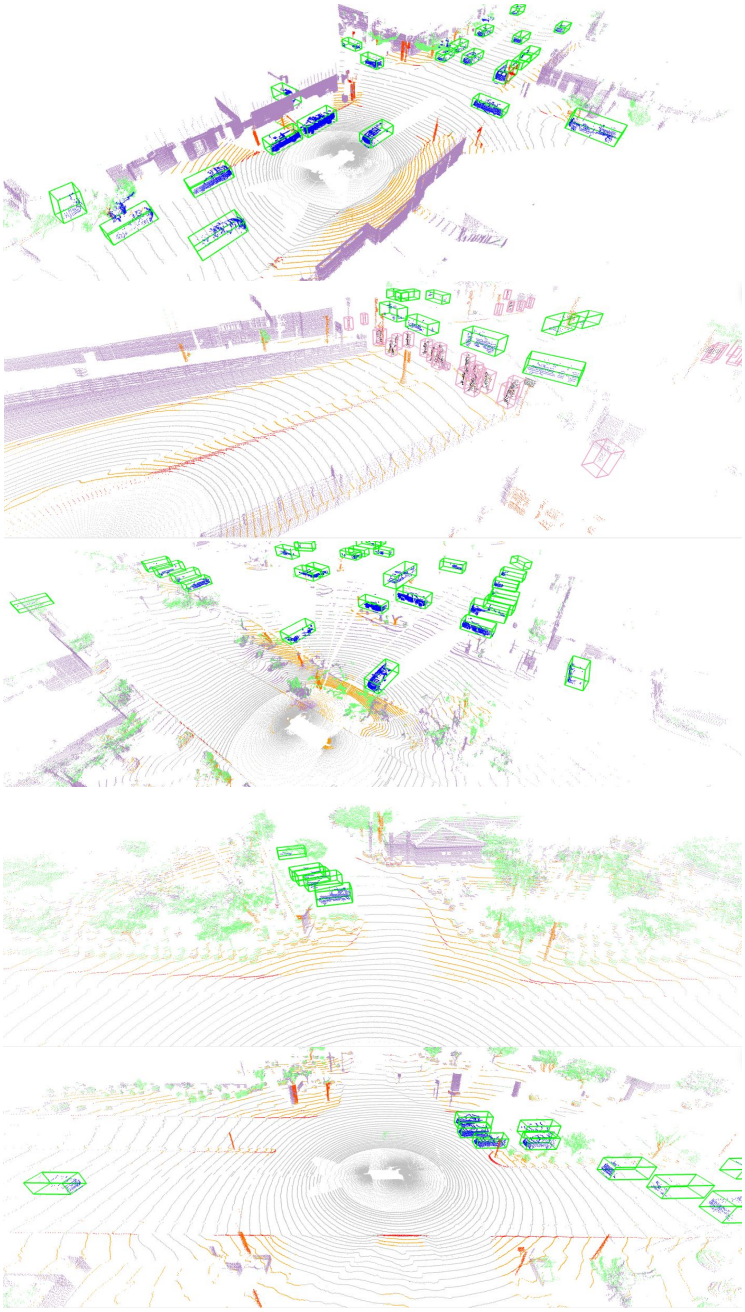

Fig. 9. Joint detection and semantic segmentation qualitative results. Green boxes: vehicle. Lavender boxes: pedestrian. Lavender points: building. Grey points: road. Orange points: sidewalk. Blue points: vehicle. Black points: pedestrian. Red points: pole/sign/tree trunk. Green points: vegetation.