Table 6: The ablation of mIoU improvement of each component on the nuScenes and WOD `val` split when trained only for the segmentation task. XSF and STD stand for cross-space transformer and segmentation transformer decoder.

Baseline (3.1)	STD	Multi-frame	XSF	nuScenes	WOD
✓				76.6	70.3
✓	✓			78.3 (+1.7)	70.6 (+0.3)
✓	✓	✓		80.8 (+4.2)	71.2 (+0.9)
✓	✓	✓	✓	**81.7** (+5.1)	**71.3** (+1.0)

Table 8: Panoptic segmentation result on nuScenes `val` split.

	stage	PQ	SQ	RQ	mIoU
LidarMultiNet [66]	2-stage	**81.8**	**90.8**	89.7	83.6
LiDARFormer	1-stage	**81.8**	90.7	**89.9**	**84.1**

Table 9: Design choice of the segmentation decoder on the nuScenes `val` split.

LiDARFormer seg only result without XSF (mIoU)	80.8
w/o voxel to class attention	80.4 (-0.4)
w/o class to voxel attention	80.1 (-0.7)
w/o dynamic kernel	80.3 (-0.5)
w/o class embedding initialization	80.5 (-0.3)

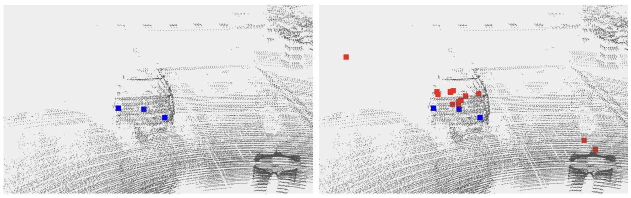
formance, particularly for the detection task. We also evaluate the panoptic segmentation performance of our multi-task network in Table 8. Even without a second stage dedicated to panoptic segmentation, our model achieves competitive results compared to the previous best method, LidarMultiNet. This demonstrates the ability of our multi-task transformer decoder to generate more compatible results for both tasks.

Table 7: The ablation of the improvement of shared transformer decoder on the nuScenes `val` split when jointly trained with detection task.

Baseline [66]	XTF Seg	XTF Det	XSF	mIoU	mAP	NDS
✓				81.8	65.2	70.0
✓	✓			82.1 (+0.3)	65.4 (+0.2)	70.2 (+0.2)
✓		✓		82.4 (+0.6)	65.9 (+0.7)	70.3 (+0.3)
✓	✓	✓		82.6 (+0.8)	66.0 (+0.8)	70.2 (+0.2)
✓	✓	✓	✓	**82.7** (+0.9)	**66.6** (+1.4)	**70.8** (+0.8)

Table 10: The ablation of XSF on the nuScenes `val` split. S→D and D→S denote sparse-to-dense (3b) and dense-to-sparse (3a) XSFs.

S→D	D→S	Add Convs	mIoU	mAP	NDS
Segmentation Only					
✓	✓		81.7	-	-
		✓	80.9 (-0.8)	-	-
Multi-task					
✓	✓		82.7	66.6	70.8
	✓	✓	82.8 (+0.1)	66.0 (-0.6)	70.5 (-0.3)



Direct Mapping　　Cross-space Transformer

Figure 5: **Visualization of the learned offsets.** We showcase the features of a car's 3D voxels (blue) and their corresponding deformable offsets (red) that were learned in our XSF module. For a better visual representation, we only highlight the offsets with high attention scores.