Rewis3d: Reconstruction Improves Weakly-Supervised Semantic Segmentation

Supplementary Material

In this supplement to our work on Rewis3d – Reconstruction Improves Weakly-Supervised Semantic Segmentation, we provide further results, derivations, and implementation details, as indexed below. We **particularly encourage** the reader to see the added information on 3D segmentation performance in App. A and the additional discussion on the merits of reconstructed 3D over LiDAR in App. E.

A. Rewis3d Improves 3D Segmentation

As shown in Table A1, cross-modal learning significantly benefits 3D segmentation. Our CMC framework improves 3D mIoU by +3.7 on Waymo and +4.1 on NYUv2 compared to the 3D-only EMA baseline, demonstrating effective bidirectional knowledge transfer. Due to the lack of ground truth labels for the reconstructed point clouds, we evaluate against unprojected 2D segmentation masks. While this introduces some uncertainty in the reference labels, the validity of relative performance comparisons remains intact. Thus, we demonstrate reliable improvements in the 3D modality, which are visually corroborated in Fig. A1.

Table A1. **3D Semantic Segmentation Performance (mIoU %).** Cross-modal consistency also improves 3D segmentation through bidirectional knowledge transfer from 2D. The column Δ vs EMA denotes the absolute performance improvement in percentage points (pp) of our method compared to the 3D-only Mean Teacher baseline (EMA).

Method	**Waymo**		**KITTI-360**		**NYUv2**	
	mIoU	Δ vs EMA	**mIoU**	Δ vs EMA	**mIoU**	Δ vs EMA
EMA (3D only)	41.8	–	44.3	–	24.4	–
Ours (Recon)	**45.5**	**+3.7**	**44.9**	**+0.6**	**28.5**	**+4.1**

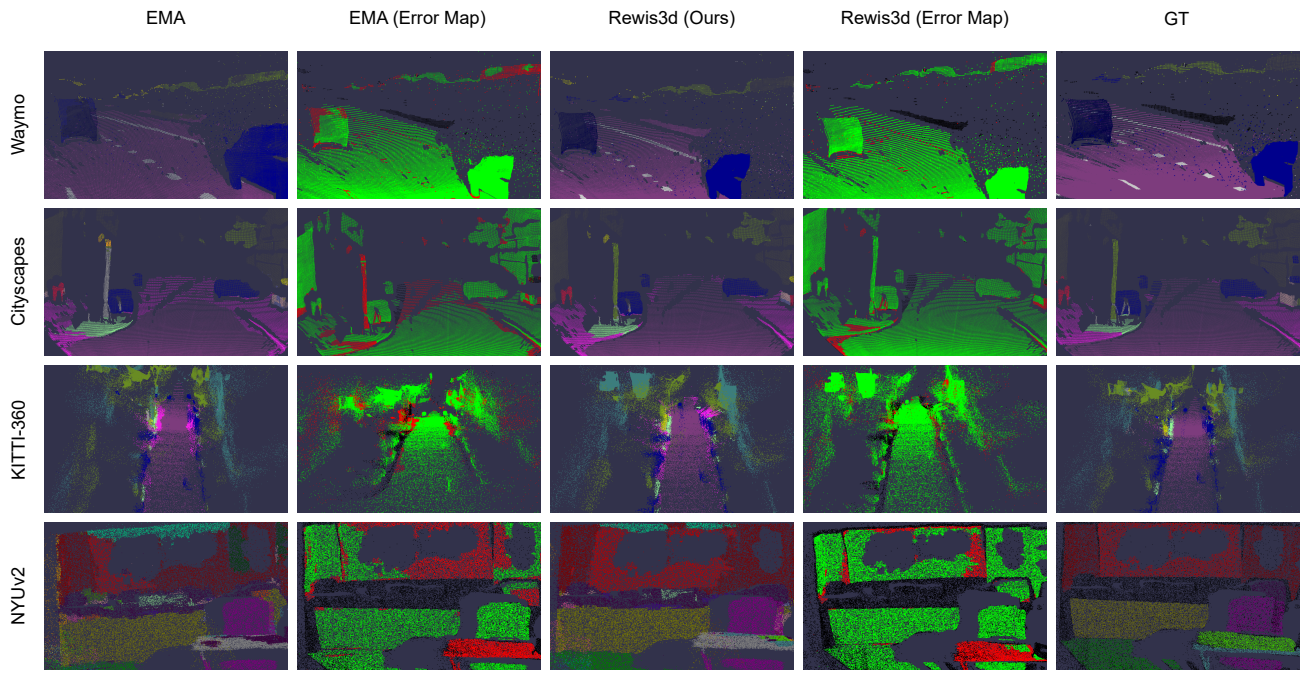

Figure A1. **3D Segmentation and Error Maps.** Rewis3d improves also improves the 3D segmentation noticeably. The bidirectional knowledge transfer through CMC especially improves the 3D segmentation quality with respect to errors emanating from misclassifications of objects. This highlights how segmentation models in 2D and 3D domain possess different advantages. While the structure of the 3D space lends itself to more accurate separation of objects, the correct class assignment appears to be easier to learn in 2D.

B. Further Qualitative Results

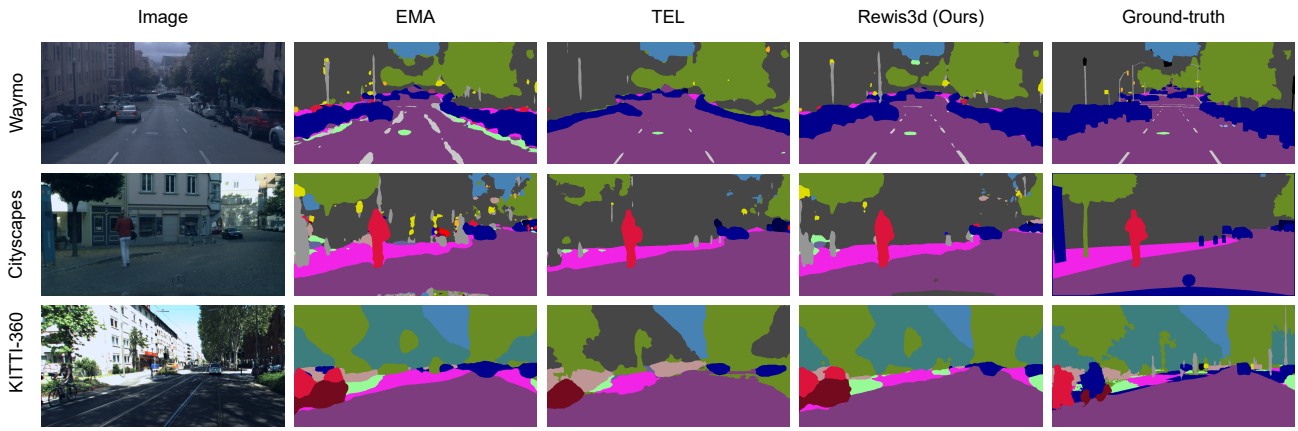

Figure B1. **Qualitative comparison with point supervision.** Even with minimal supervision (one point per object), Rewis3d successfully propagates labels to the full object extent. Note the improved segmentation of the vehicle and road markers in Waymo (top row) and the clearer delineation of the sidewalk in KITTI-360 (bottom row) compared to the EMA and TEL baselines. Additionally, we observe that the KITTI-360 ground truth contains occasional labeling errors (e.g., the bicycle); consequently, our model's visually correct predictions in these regions may be penalized during quantitative evaluation against the imperfect ground truth.

Figure B2. **Additional Qualitative Results with Scribble Supervision.** We provide further qualitative examples of Rewis3d predictions on the Waymo (top), Cityscapes (middle), and KITTI-360 (bottom) datasets. These visualizations demonstrate the model's capability to generate spatially coherent segmentations with precise boundaries across diverse outdoor urban environments, solely trained with sparse scribble supervision.

C. Quantitative Results

This section presents detailed class-wise evaluations to complement the experiments discussed in the main paper. Table C1 details the individual performance of Rewis3d on Waymo using scribble and point labels. Similarly, comprehensive results for NYUv2 and KITTI-360 are provided in Table C2 and Table C3, respectively. Finally, Table C4 compares the class-wise performance of Rewis3d against competing methods across the three types of weak labels available for Cityscapes. The tables for the three outdoor datasets further contain results of using EoMT as the segmentation model with Rewis3d to demonstrate that our method is orthogonal to the improvements gained from relying on the strong foundational priors that EoMT leverages on its own.

Table C1. **Comparison of 2D mIoU Results and Relative Scores on Waymo Dataset** Bold numbers indicate the best and underlined values the second-best performance among non-fully supervised approaches.

Method	mIoU	SS/FS (%)	bicycle	bird	building	bus	car	construction cone pole	cyclist	ground	ground animal	lane marker	motorcycle	motorcyclist	other large vehicle	pedestrian	pedestrian object	pole	road	road marker	sidewalk	sign	sky	traffic light	trailer	truck	vegetation
Fully Supervised	59.1	-	79.3	10.0	93.4	16.0	93.9	65.4	57.9	74.6	0.0	64.3	73.7	0.0	16.0	83.0	9.0	69.7	94.8	64.9	84.3	73.7	95.3	74.0	28.1	66.3	89.2
Point annotations																											
EMA	43.4	73.4	58.1	**0.0**	84.8	15.7	80.7	39.6	10.9	50.3	**0.0**	30.9	54.0	**0.0**	**10.8**	58.7	7.0	43.9	85.3	43.1	60.8	49.4	92.0	54.3	22.4	54.3	78.8
SASFormer	37.4	63.2	52.9	**0.0**	80.7	16.4	75.5	11.9	15.4	49.8	**0.0**	32.7	47.1	**0.0**	0.1	52.9	7.0	23.9	89.9	29.9	57.4	44.0	**93.8**	32.0	1.7	40.8	78.3
TEL	35.4	59.9	0.0	**0.0**	78.5	**41.9**	54.3	21.2	9.4	45.7	**0.0**	**53.9**	28.5	**0.0**	2.3	49.3	0.0	34.8	86.4	39.7	55.1	45.4	92.9	41.2	1.9	29.8	73.3
Ours (Recon)	**48.8**	**82.6**	**59.9**	**0.0**	**88.0**	6.1	**89.5**	**44.3**	**23.7**	**64.1**	**0.0**	53.6	**60.2**	**0.0**	7.6	**66.5**	**7.3**	**49.9**	**91.9**	**54.4**	**72.9**	**57.7**	91.0	**55.9**	**28.8**	**64.6**	**81.2**
Scribble annotations																											
EMA	49.4	83.6	71.9	**0.0**	87.8	15.4	84.8	46.9	42.9	59.6	0.0	39.1	65.4	**0.0**	14.7	68.7	5.6	52.7	89.1	47.3	65.6	56.3	92.8	56.6	35.0	57.8	81.6
SASFormer	37.8	64.0	48.8	**0.0**	81.1	23.4	81.5	12.8	8.4	51.7	0.0	34.3	40.6	**0.0**	0.0	55.2	3.8	25.0	91.1	32.3	58.5	44.5	**93.7**	33.2	0.9	46.1	78.2
TEL	42.4	71.8	53.1	**0.0**	82.4	**35.9**	67.3	23.1	29.4	52.0	0.0	53.2	46.9	**0.0**	0.0	63.1	2.8	41.6	88.3	37.7	61.6	54.5	93.4	50.9	1.0	45.8	76.4
Ours (Real 3D)	47.8	80.9	62.2	**0.0**	88.5	14.2	87.5	35.9	**44.2**	64.6	0.0	33.2	63.2	**0.0**	5.7	70.8	**9.1**	48.3	89.1	39.1	71.2	52.3	90.1	45.9	**46.3**	52.7	81.3
Ours (Recon)	**53.3**	**90.2**	**72.2**	**0.0**	**88.9**	30.1	**90.0**	**51.5**	37.5	**68.3**	**0.1**	**55.4**	**67.2**	**0.0**	**17.5**	**73.6**	7.5	**55.5**	**92.7**	**58.3**	**76.0**	**62.8**	91.0	**59.0**	42.1	**60.2**	**82.0**
Scribble annotations & EoMT backbone																											
EMA	53.13	-	65.44	0.00	**88.77**	15.97	**90.19**	**51.47**	55.24	**62.41**	**0.71**	**43.00**	60.37	0.00	10.41	**75.11**	**27.79**	**57.93**	**90.74**	**48.33**	72.10	**62.61**	**91.94**	**58.49**	45.03	73.68	**80.64**
Ours (Recon)	**54.04**	-	**67.18**	**5.19**	88.05	**32.40**	90.15	45.22	**65.26**	61.93	0.00	39.75	**62.83**	0.00	**21.85**	70.34	25.58	57.67	89.58	44.90	**72.27**	60.24	90.09	57.14	**45.70**	**77.89**	79.81

Table C2. **Detailed Class-wise IoU Comparison on NYUv2 Dataset** Bold numbers indicate the best and underlined values the second-best performance among non-fully supervised approaches.

Method	mIoU	SS/FS (%)	background	bag	bathtub	bed	blinds	books	bookshelf	box	cabinet	ceiling	chair	clothes	counter	curtain	desk	door	dresser	floor	floor mat	lamp	mirror	nightstand	otherfurniture	otherprop	otherstructure	paper	person	picture	pillow	refrigerator	shelves	sink	sofa	table	television	toilet	towel	wall	whiteboard	window
Fully Supervised	51.1	-	80.4	12.5	48.0	63.8	51.0	35.4	68.3	19.0	71.7	64.1	54.4	22.8	61.4	62.3	25.0	49.2	55.8	64.5	27.8	45.5	47.8	51.5	20.6	41.4	37.2	38.3	81.8	66.7	45.9	64.1	11.2	61.1	45.2	43.0	68.4	80.7	46.0	85.6	78.4	46.1
Scribble annotations																																										
EMA	42.9	84.0	66.6	11.0	39.3	52.8	31.0	31.9	53.6	11.8	63.8	47.7	49.0	21.9	59.0	47.1	**20.5**	**48.5**	45.0	57.3	**34.3**	31.3	36.1	42.8	15.1	31.7	28.6	31.3	66.5	56.7	36.4	53.3	11.5	50.0	38.6	36.8	57.9	69.5	36.3	80.2	77.1	37.0
SASFormer	45.2	88.5	71.2	15.6	**50.8**	59.3	40.3	34.0	56.0	13.6	**69.0**	51.8	50.0	**23.2**	59.5	52.1	18.0	42.5	40.8	56.9	32.5	34.8	34.3	38.7	14.3	34.2	29.8	33.0	**75.6**	53.4	40.3	**59.7**	**12.4**	**57.5**	36.9	35.5	**61.4**	69.9	40.1	79.2	72.8	38.4
TEL	39.1	76.5	69.0	**16.3**	28.9	47.2	18.6	30.4	56.7	10.4	57.5	49.3	39.7	16.0	53.4	36.1	10.7	40.7	33.7	54.2	15.8	33.5	32.0	28.4	**28.3**	33.6	16.3	35.4	69.9	51.1	35.3	51.6	6.7	47.5	31.3	27.3	41.5	63.4	25.8	76.6	74.4	38.6
Ours (Real 3D)	44.7	87.6	72.7	9.9	47.4	55.9	39.7	31.4	57.3	14.0	62.8	54.1	45.6	21.6	**60.7**	**58.3**	17.7	45.9	**50.5**	**61.8**	28.1	35.5	**41.3**	44.1	14.4	34.1	**34.3**	33.8	73.9	59.7	36.4	54.1	9.8	55.7	37.4	**42.9**	51.6	68.3	40.0	80.9	70.1	38.8
Ours (Recon)	**46.1**	**90.2**	**73.3**	9.1	40.1	**60.2**	**44.9**	**36.3**	**63.1**	**14.2**	67.2	**58.9**	**51.0**	21.9	60.6	54.5	18.5	47.7	41.4	61.6	32.2	**43.5**	37.3	**45.0**	16.1	**38.4**	33.7	**35.6**	74.3	**65.2**	**46.5**	56.5	11.7	55.3	**41.6**	41.6	60.1	**74.7**	**43.6**	**82.2**	**77.5**	**44.1**

Table C3. **Comparison of mIoU Results and Relative Scores on KITTI-360 Dataset** Bold numbers indicate the best and underlined values the second-best performance among non-fully supervised approaches.

Method	mIoU	SS/FS (%)	bicycle	building	car	fence	motorcycle	person	pole	road	sidewalk	sky	terrain	traffic light	traffic sign	truck	vegetation	wall
Fully Supervised	68.4	-	49.1	89.0	94.5	53.4	60.6	66.3	43.4	96.4	85.5	94.3	77.7	0.0	47.9	78.6	89.8	67.8
Point annotations																		
EMA	52.2	76.3	33.1	78.2	78.6	42.0	38.7	43.3	22.8	84.4	59.9	84.7	56.7	**0.0**	33.6	51.5	80.0	48.1
SASFormer	27.0	39.5	21.0	54.6	22.2	28.2	3.8	6.7	0.1	65.6	27.5	53.2	26.0	**0.0**	8.2	23.5	60.9	30.5
TEL	48.9	71.5	24.8	79.4	70.6	39.5	37.0	37.9	16.0	81.1	54.1	**86.9**	56.8	**0.0**	34.8	42.9	76.3	43.7
Ours (Recon)	**58.2**	**85.1**	**43.7**	**83.9**	**89.7**	**49.5**	**48.8**	**46.4**	**25.3**	**91.0**	**70.8**	72.2	**61.8**	**0.0**	**37.9**	**63.9**	**83.8**	**63.1**
Scribble annotations																		
EMA	60.3	88.1	43.4	83.5	86.7	46.5	**59.7**	58.2	35.3	85.8	65.4	87.3	57.6	**0.0**	39.7	**74.2**	82.4	58.5
SASFormer	46.4	67.8	35.8	72.7	78.3	42.1	23.3	21.7	4.5	74.4	52.1	76.9	61.4	**0.0**	17.2	55.6	78.9	47.8
TEL	59.2	86.5	**50.5**	82.6	85.9	45.7	51.6	53.1	31.4	89.1	69.7	88.2	**67.7**	**0.0**	40.6	50.8	83.5	57.4
Ours (Real 3D)	61.7	90.2	48.2	84.3	89.4	49.5	53.5	56.8	36.1	91.5	74.6	79.8	63.4	**0.0**	40.3	72.8	83.8	62.9
Ours (Recon)	**63.4**	**92.7**	41.1	**85.9**	**89.8**	**50.6**	56.3	**61.6**	**36.3**	**94.0**	**78.6**	**90.3**	66.4	**0.0**	**41.0**	72.6	**85.8**	**64.7**
Scribble annotations & EoMT backbone																		
EMA	62.58	-	47.23	83.63	89.17	47.51	**62.86**	59.52	**36.29**	91.19	71.78	86.77	61.16	0.00	42.42	**78.58**	82.95	60.31
Ours (Recon)	**63.94**	-	**47.29**	**85.76**	**89.31**	**49.27**	57.66	**61.79**	34.49	**94.05**	**78.94**	**88.25**	**66.41**	0.00	**43.07**	76.62	**85.65**	**64.47**

Table C4. **Comparison of mIoU Results and Relative Scores on Cityscapes Dataset** Bold numbers indicate the best and underlined values the second-best performance among non-fully supervised approaches.

Method	mIoU	SS/FS (%)	bicycle	building	bus	car	fence	motorcycle	person	pole	rider	road	sidewalk	sky	terrain	traffic light	traffic sign	train	truck	vegetation	wall
Fully Supervised	77.6	-	74.5	92.4	86.4	94.5	60.1	62.7	78.9	58.9	60.4	98.2	85.4	94.8	65.7	65.5	74.8	80.4	82.0	92.3	66.3
Point annotations																					
EMA	50.5	65.1	49.0	74.6	57.8	82.9	29.1	25.0	51.8	25.9	30.3	94.1	64.9	78.4	37.7	16.9	30.0	52.3	49.4	77.2	32.7
SASFormer	42.7	55.0	35.5	67.5	53.1	70.1	33.8	17.8	38.3	12.1	10.4	91.8	41.1	75.8	18.6	10.8	25.4	45.8	54.2	69.4	41.5
TEL	53.0	68.3	**59.0**	81.2	39.7	80.6	**36.7**	**31.3**	**62.9**	**32.8**	**44.2**	93.2	62.7	80.5	41.7	**36.5**	**57.8**	30.9	23.9	**84.1**	26.8
Ours (Recon)	**56.5**	**72.8**	54.0	**82.9**	**69.6**	**85.3**	33.1	31.1	55.3	32.4	31.8	**94.6**	**67.1**	**84.2**	**43.1**	25.5	40.6	**62.1**	**55.3**	81.6	**43.1**
Scribble annotations																					
EMA	61.2	78.9	63.6	82.2	73.9	83.4	41.7	48.6	64.5	40.0	50.9	87.1	50.6	80.2	45.3	40.4	51.4	61.3	68.7	80.7	48.3
SASFormer	55.6	71.7	53.9	82.1	68.4	74.8	41.4	23.4	52.7	26.5	28.5	92.3	47.0	88.0	49.9	35.6	49.2	56.2	55.8	84.4	46.4
TEL	64.4	83.0	**69.6**	86.7	62.6	65.0	46.7	47.0	**71.2**	**48.7**	51.9	91.3	58.6	**92.5**	**53.0**	**59.1**	**68.7**	60.1	54.2	**87.6**	48.8
Ours (Recon)	**68.1**	**87.8**	65.4	**87.5**	**78.9**	**89.0**	**48.2**	**49.6**	69.8	47.4	**55.2**	**94.6**	**66.9**	90.6	52.0	51.4	61.6	**68.7**	**75.8**	86.4	**55.0**
Coarse annotations																					
EMA	66.5	85.7	63.6	87.7	73.6	87.8	49.8	50.7	64.6	43.1	47.0	**95.9**	**71.7**	91.9	52.6	50.2	60.1	59.8	72.6	87.5	53.0
SASFormer	42.8	55.2	41.1	68.6	54.5	74.3	33.8	15.0	40.8	7.8	16.4	92.7	37.4	79.6	14.2	9.1	26.0	36.4	56.2	70.2	39.2
TEL	64.9	83.7	**69.9**	87.3	62.8	73.1	50.9	**56.3**	**70.4**	**49.6**	**54.2**	91.3	55.8	91.1	49.5	**61.9**	**71.2**	59.5	51.0	87.5	39.7
Ours (Recon)	**68.6**	**88.4**	64.7	**89.0**	**80.6**	**88.3**	**51.2**	53.1	67.3	43.1	48.8	95.7	69.9	**92.5**	**55.7**	52.3	62.6	**67.7**	**77.9**	**87.7**	**55.2**
Scribble annotations & EoMT backbone																					
EMA	71.15	-	64.96	86.88	87.56	89.79	52.10	54.00	**72.72**	50.65	57.58	**95.71**	**73.52**	88.77	53.06	53.93	62.45	79.96	85.03	85.49	57.67
Ours (Recon)	**73.50**	-	**69.81**	**88.86**	**88.81**	**91.00**	**58.77**	**62.64**	72.18	**52.17**	**58.59**	95.59	71.86	**90.80**	**56.04**	**55.68**	**66.45**	**82.45**	**87.17**	**86.66**	**60.93**

D. Sampling Strategies for Reconstruction

As discussed in the main paper, efficiently processing the massive point clouds generated from video sequences (often exceeding 60 million points) presents a challenge. Simply downsampling the global point cloud randomly is ineffective for our Cross-Modal Consistency (CMC) loss, as it yields too few correspondences (approx. 167 points) within the target image's field of view, as shown in Fig. D1. Conversely, retaining points exclusively from the target view maximizes correspondences but discards the geometric context necessary for the 3D network to learn robust features, resulting in fragmented 3D shapes (see "Correspondences Only" in Fig. D1). To resolve this, we employ a *View-Aware Sampling Strategy* that constructs a hybrid point cloud: 60% of points are sampled from the current camera view to ensure dense 2D-3D alignment for supervision, while the remaining 40% are sampled from the surrounding scene to provide structural context. This balanced approach enables both effective cross-modal transfer and accurate 3D segmentation.

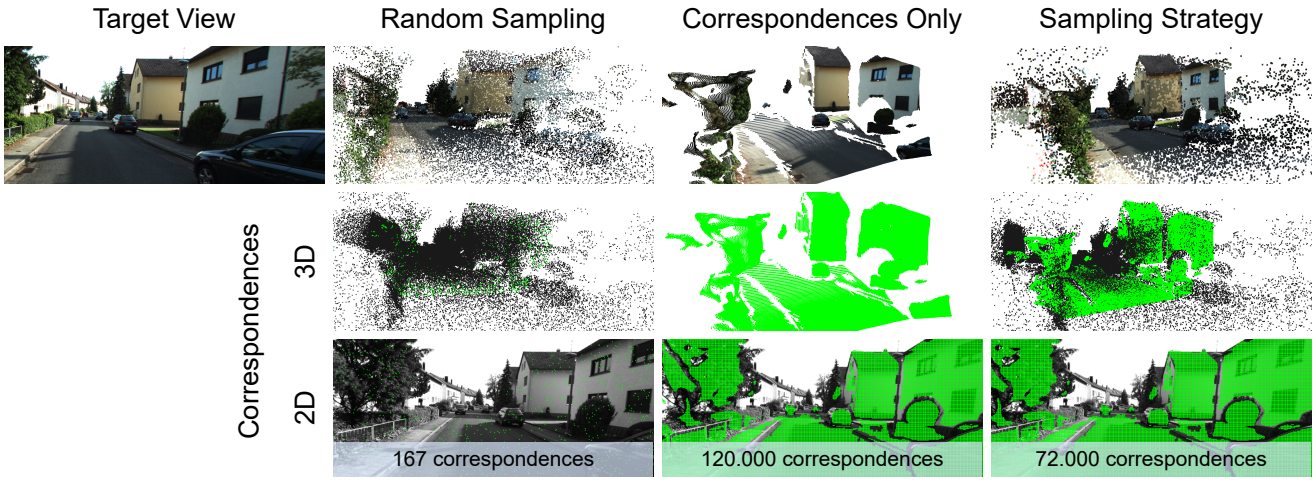

Figure D1. **Visualization of sampling strategies for cross-modal consistency.** The rows display the 3D point cloud (middle) and its projection onto the 2D target view (bottom). Green points indicate valid 2D-3D correspondences for the target image, while black points represent the remaining scene geometry. Left: Random Sampling (120k points global) results in sparse correspondences (avg. 167) in the view, insufficient for dense consistency learning. Center: Sampling strictly from correspondences (100% target view) maximizes alignment but results in a fragmented 3D scene (note the missing car geometry), preventing the 3D network from learning global context. Right: Our View-Aware Sampling Strategy (60% target view, 40% spatial context) ensures dense correspondences ($\sim$72K) while maintaining the holistic scene structure required for robust 3D segmentation.

E. Analysis of Real vs. Reconstructed Supervision

In Sec. 4.3 of the main paper, we observed the counterintuitive result that supervision from reconstructed point clouds ("Ours (Recon)") outperformed ground-truth LiDAR ("Ours (Real 3D)") on the Waymo dataset (53.3% vs. 51.8%). We hypothesized that this performance gap stems from two factors inherent to the raw LiDAR data: (1) high sparsity leading to fewer valid correspondences, and (2) the lack of a reconstruction confidence score, which forces the use of single-term weighting rather than our proposed Dual Confidence mechanism.

To validate this hypothesis, we conducted a controlled experiment where we artificially degraded our Reconstructed data to mimic the characteristics of the Real LiDAR sensor. Specifically, we:

1. Restricted the reconstruction to single-scan density
2. Reduced the amount of correspondences to the average amount of correspondences available in the real 3D data
3. Disabled the reconstruction confidence weighting component, relying only on prediction confidence (matching the Real 3D setup).

The results are presented in Table E1. When the reconstructed data is restricted to the same sparsity and weighting constraints as the Real LiDAR ("Recon (Simulated LiDAR)"), the performance drops to 51.7%, virtually matching the Real 3D performance of 51.8%. This confirms that the superior performance of Rewis3d stems specifically from the density of the multi-view reconstruction and the noise-suppression capability of the reconstruction confidence, rather than an artifact of the data source itself.

Table E1. **Validation of Real vs. Reconstructed Gap (Waymo).** We simulate the characteristics of Real LiDAR (sparsity and lack of confidence scores) using our Reconstructed data. When matching the constraints of Real LiDAR, our performance aligns with the Real 3D baseline, confirming that the gains of Rewis3d come from the increased density and dual-confidence weighting enabled by the reconstruction.

Source	**Density**	**Confidence Weighting**	**mIoU (%)**
Ours (Real 3D)	Sparse (LiDAR)	Single (Pred. Only)	51.8
Ours (Recon)	Sparse (Simulated)	Single (Pred. Only)	51.7
Ours (Recon)	**Dense (Multi-view)**	**Dual (Pred. + Rec.)**	**53.3**

F. Label Generation

The scribble labels employed in the majority of our evaluations were generated using Scribbles for All [3]. While we utilized the published labels for KITTI-360 and Cityscapes, the annotations for Waymo and NYUv2 were generated specifically for this work. Table F1 details the configuration parameters adopted for the label generation process.

Table F1. **Scribble Generation Configurations for Waymo Open Dataset and NYUv2**

Parameter	**Waymo Open Dataset Value**	**NYUv2 Dataset Value**
height distortion	0.9	1
min binary erosion	2	2
max binary erosion	40	20
it extra erosion	5	2
min erosion area share	0.002	0.003
max erosion area share	0.05	0.15
background px value	[0, 0, 0]	[255, 255, 255]
background input values	[0]	[0]
ignore values	[0]	[255]
patience	20	20
error tolerance px	0	0
min blob area	1500	80
line thickness	5	3
scribble scale	1.0	1.0

G. Implementation Details & Hyperparameters

This section provides a comprehensive overview of the hyperparameters and augmentation strategies employed for the 2D and 3D branches of our framework. Table G1 specifies the optimization settings, which are aligned with standard practices for SegFormer and Point Transformer V3 to ensure fair comparisons. Additionally, Tables G2 and G3 itemize the specific data augmentations applied. Notably, we apply stronger augmentations to the student models relative to the teachers—a fundamental aspect of the Mean Teacher paradigm designed to enforce consistency and robustness.

Table G1. **Overview of core training hyperparameters for 2D and 3D branches**

	2D Branch		**3D Branch**
Config	**Value**	**Config**	**Value**
Optimizer	AdamW	Optimizer	AdamW
Scheduler	PolyLR	Scheduler	OneCycleLR
Criteria	CrossEntropy Student Teacher Loss	Criteria	CrossEntropy Student Teacher Loss
Learning Rate	0.00005	Learning Rate	0.001
Weight Decay	1e-08	Weight Decay	0.005
Batch Size	12	Batch Size	12
Datasets	KITTI-360 / Waymo / NYUv2 /	Datasets	KITTI-360 / Waymo / NYUv2 /
Epochs	Cityscapes: 50 (200 NYUv2)	Epochs	Cityscapes: 50 (200 NYUv2)

Table G2. **Overview of Applied Data Augmentations in 2D.** The first two columns list the augmentations and their corresponding parameters. The third and fourth columns indicate whether each augmentation is applied to the inputs of both the student and teacher models, or to only one of them.

Augmentation	**Parameters**	**Student**	**Teacher**
Random Horizontal Flip	p: 0.5	✓	✓
Scale and Distort	scale: [0.5, 1.2], distort: [0.9, 1.1]	✓	✓
Random Crop	KITTI-360: (376, 512) Waymo: (640, 640) NYUv2: (480, 480)	✓	✓
Gaussian Blur	kernel size: 7; p: 0.5	✓	
Augmix	severity: 2; p: 0.5	✓	
Cutout (1)	square size: 90; p: 1.0	✓	
Cutout (2)	square size: 90; p: 0.5	✓	

Table G3. **Overview of Applied Data Augmentations in 3D.** The first two columns list the augmentations and their corresponding parameters. The third and fourth columns indicate whether each augmentation is applied to the inputs of both the student and teacher models, or to only one of them.

Augmentation	**Parameters**	**Student**	**Teacher**
Random Rotation	axis: z, angle [-1, 1]; p: 0.5	✓	✓
Random Scale	scale: [0.9, 1.1]	✓	✓
Random Flip	p: 0.5	✓	✓
Random Jitter	sigma: 0.005; clip: 0.02	✓	
PointClip	[100, 40, 100]	✓	

H. Colormaps

Figure H1. **Semantic class color coding for all evaluated datasets.** We visualize results using consistent color mappings across each dataset: Cityscapes (19 classes), KITTI-360 (19 classes), NYUv2 (40 classes), and Waymo Open Dataset (25 classes). Color assignments follow the official dataset conventions where available. Note that while some classes share similar names across datasets (e.g., 'road', 'car', 'person'), their definitions and label protocols may differ between datasets.