Split	Building Scene Ids
Evaluation	UwV83HsGsw3, X7HyMhZNoso, Z6MFQCViBuw, e9zR4mvMWw7, q9vSo1VnCiC, rPc6DW4iMge, rqfALeAoiTq, uNb9QFRL6hY, wc2JMjhGNzB, x8F5xyUWy9e, yqstnuAEVhm
Testing	VFuaQ6m2Qom, VLzqgDo317F, ZMojNkEp431, jh4fc5c5qoQ, jtcxE69GiFV, pRbA3pwrgk9, pa4otMbVnkk, D7G3Y4RVNrH, dhjEzFoUFzH, GdvgFV5R1Z5, gYvKGZ5eRqb, YmJkqBEsHnH,
Training	/* all other scenes excluded from evaluation & testing splits */

Table 1. Dataset split for Matterport3D [2] segmentation.

A. Experimentation details

A.1. Matterport3D dataset

To divide the 10800 panoramic equirectangular images in the Matterport3D [2] dataset, we create standard training, evaluation, and test splits. The 90 building-scale scenarios, which included a range of scene types like residences, offices, and churches, were divided into an 80-10-10 split. For all our segmentation experiments using the 40 object categories, we use these training, validation, and test splits.

B. Qualitative analysis

B.1. Multi-modal panoramic semantic segmentation

Figure 2 and Figure 1, which come from the Stanford2D3DS [1] evaluation set and the Structured3D [7] test set, respectively, show further qualitative comparisons between various fusion combinations for our proposed framework. In Fig. 2 (a) and (b), our tri-model (**RGB-D-N**) is able to give better segmentation results in the categories denoted by the black dashed rectangles, such as the *Door*, *Window*, and *Bookshelf*, while the baseline (**RGB**-only) model struggles to recognize these significantly distorted objects. The **RGB**-only baseline models wrongly segment the *Door* in figure Fig. 1 (c) as a part of the *Wall*. Our tri-model (**RGB-D-N**) in this case achieves the correct segmentation results with greater accuracy than **RGB-D** techniques. The same conditions apply to the *Cabinet* in Fig. 1 (a) and the support between the *Bed* and *Cabinet* in Fig. 1 (b). Compared to other approaches, In Fig. 1 (d), along with the precise geometry shapes for objects placed inside the *Cabinet* structure, a better segmentation result from our multi-modal (**RGB-D-N**) is displayed. However, due to visual ambiguity, the category is incorrectly predicted by all models.

#Inputs	Method	#Params (G)	TFLOPs
Unary	Trans4PASS+ [6]	0.039	0.131
	HoHoNet [5]	0.070	0.125
	PanoFormer [4]	0.020	0.081
	OURS	0.040	0.079
Binary	HoHoNet [5]	0.070	0.126
	PanoFormer [4]	0.020	0.081
	OURS	0.081	0.106
Ternary	*OURS*	0.123	0.133

Table 2. Comparison of computational complexity calculated @ $512 \times 1024 \times 3$ input dimensional.

C. Quantitative analysis

C.1. Computational complexity

For tri-modal (**RGB**-**D**epth-**N**ormals), bi-modal (**RGB**-**D**epth), and uni-modal (**RGB**-Only) panoramic fusion on Stanford2D3DS [1], we compare the computational complexity of our framework with that of existing methods in Tab. 2. As the number of input streams rises, our study indicates that our method's complexity also significantly rises.

C.2. Detailed results in indoor scenarios

More qualitative comparisons based on three-fold cross validation of Stanford2D3DS [1] indoor scenarios are shown in Tab. 3 to support our propose approach. When compared to the current panoramic approaches, our multi-model fusion models segment objects in regularly used categories including ceiling, wall, floor, window, and office furniture better. Our **RGB**-**D**epth-**N**ormals fusion model receives top score mIoU in 8 out of 13 categories. However, this model struggled to segment the *Beam*, *Column*, and *Wall* categories.

Figure 3 shows the advantage of combining multi-modalities, such as **RGB**, **D**epth, and **N**ormals, over the baseline of our technique that uses **RGB** alone to utilize complimentary textual, geometric, and disparity information. With our tri-fusion model (**RGB-D-N**), we generally observe a considerable improvement across all object categories. For the *Pillow* and Mirror categories on Structured3D [7], refer Fig. 3 (a), as well as the *Bathtub* and *Gym Equipment* categories on Matterport3D [2], refer Fig. 3 (b), we saw a considerable rise of mIoU of up to 10% and 15%, respectively. However, the box category on Structured3D [7] and the *Cabinet*, *Plant*, and *Toilet* categories on [2] also had drops of 1% to 4%.

Figure 1. Structured3D [7] segmentation visualizations. Zoom in for better view..

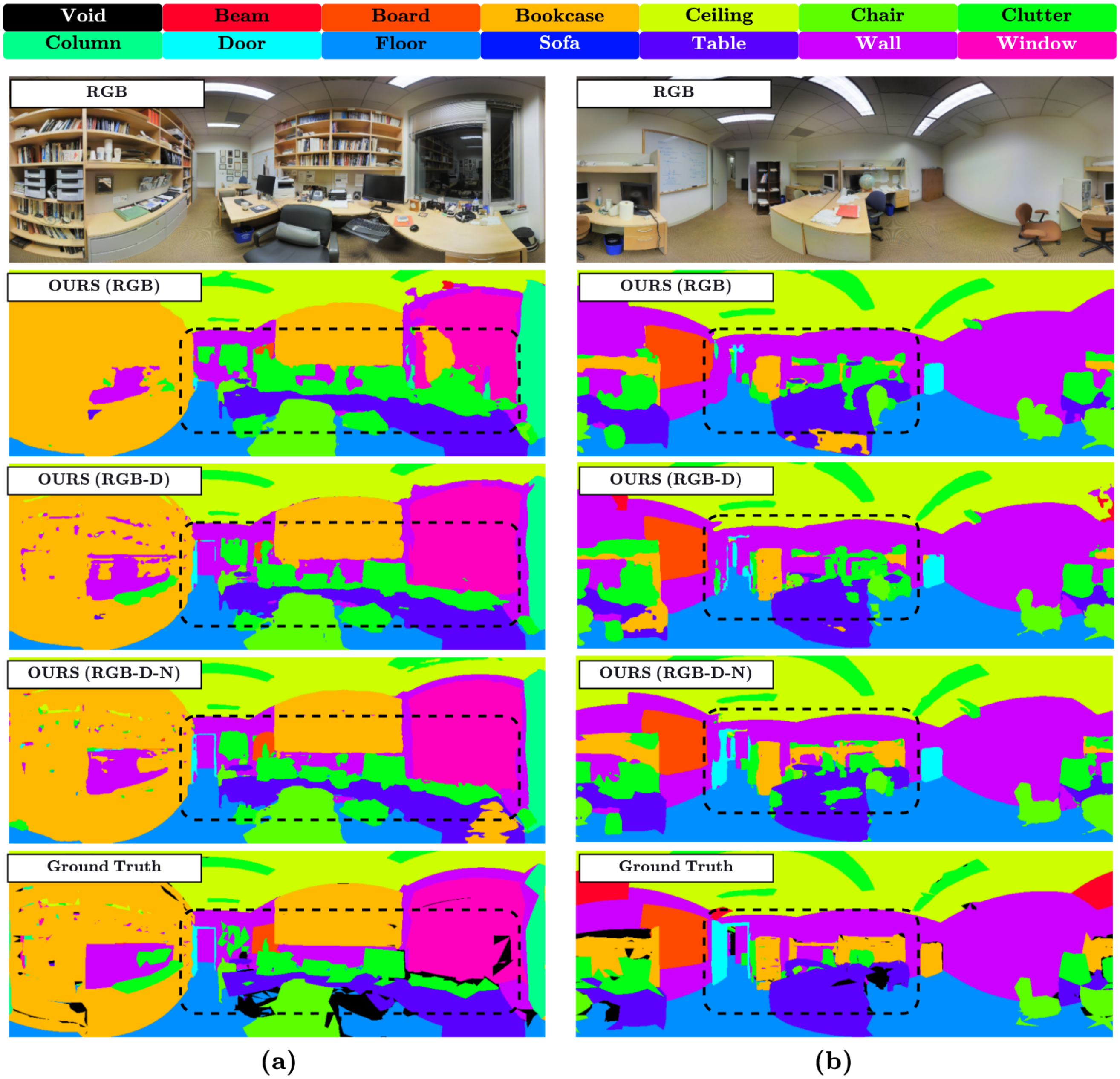


Figure 2. Stanford2D3DS [1] segmentation visualizations. Zoom in for better view.

References

[1] Iro Armeni, Sasha Sax, Amir R. Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *CoRR*, abs/1702.01105, 2017. 1, 3, 4

[2] Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. In *3DV*, pages 667–676. IEEE Computer Society, 2017. 1, 4

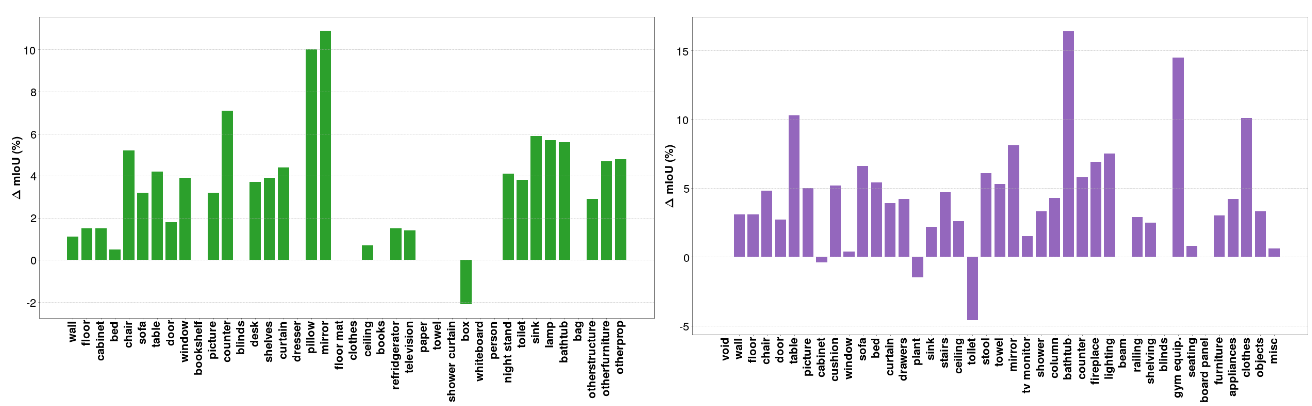
[3] Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. Tangent images for mitigating spherical distortion. In *CVPR*, pages 12423–12431. Computer Vision Foundation / IEEE, 2020. 4

[4] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: Panorama transformer for indoor 360$^{\circ }$ depth estimation. In *ECCV (1)*, volume 13661 of *Lecture Notes in Computer Science*, pages 195–211. Springer, 2022. 1, 4

[5] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *CVPR*, pages 2573–2582. Computer Vision Foundation /

Method	Modal	mIoU	beam	board	bookcase	ceiling	chair	clutter	column	door	floor	sofa	table	wall	window
Trans4PASS+ [6]	RGB	52.0	11.9	63.2	52.4	81.8	55.8	37.4	18.0	59.1	89.1	30.0	55.8	70.3	51.7
HoHoNet [5]		52.0	9.7	61.4	50.8	82.3	54.6	35.1	18.2	61.3	89.6	34.0	54.5	71.7	52.6
PanoFormer [4]		52.3	8.1	62.1	52.6	83.7	53.1	36.9	18.8	64.6	90.3	29.4	57.2	72.7	51.0
CBFC [8]		52.2	–	–	–	–	–	–	–	–	–	–	–	–	–
Tangent [3]		45.6	–	–	–	–	–	–	–	–	–	–	–	–	–
OURS		52.9	4.9	63.9	55.1	83.1	59.1	40.2	15.4	57.7	90.5	33.8	56.8	70.9	55.8
HoHoNet [5]	RGB-D	56.7	11.0	63.7	55.2	88.9	63.5	45.2	19.8	67.5	96.2	37.4	59.6	74.3	55.1
PanoFormer [4]		57.0	**15.4**	59.0	54.9	89.7	66.1	45.9	20.1	72.1	97.2	32.3	62.5	74.8	51.5
CBFC [8]		56.7	–	–	–	–	–	–	–	–	–	–	–	–	–
Tangent [3]		52.5	–	–	–	–	–	–	–	–	–	–	–	–	–
OURS		55.5	7.9	64.6	56.1	85.9	69.3	41.6	17.5	58.4	96.0	39.1	61.4	71.9	51.6
OURS	RGB-H	**60.6**	10.8	67.9	59.0	**91.0**	74.3	53.1	23.9	68.1	97.8	43.3	65.8	76.0	56.9
OURS	RGB-N	58.2	10.8	62.5	57.6	88.6	71.0	46.5	20.2	66.4	97.4	39.2	64.1	74.5	58.4
OURS	RGB-D-H	60.0	8.0	67.3	58.2	90.6	71.8	49.5	**25.0**	64.7	97.8	46.8	65.9	75.1	59.4
OURS	RGB-D-N	59.4	5.7	**77.6**	**65.7**	90.4	**76.0**	**54.2**	4.6	**81.9**	97.9	**53.6**	**71.9**	67.3	**69.0**
OURS	RGB-N-H	60.2	7.8	67.9	59.3	90.5	73.2	50.8	22.8	64.9	**98.1**	44.5	67.7	**76.3**	59.3

Table 3. Per-class results (%) on the 3-fold validation of the Stanford2D3DS [1] benchmark.



(a) Structured3D [7]

(b) Matterport3D [2]

Figure 3. Per-class mIoU (%) gain of OURS (**RGB**-**D**epth-**N**ormals) multi-modal panoramic semantic segmentation over baseline **RGB**-only (OURS) from Structure3D (*left*) and Matterport3D (*right*) test splits. Zoom in for better view.

IEEE, 2021. 1, 4

[6] Jiaming Zhang, Kailun Yang, Hao Shi, Simon Reiß, Kunyu Peng, Chaoxiang Ma, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Behind every domain there is a shift: Adapting distortion-aware vision transformers for panoramic semantic segmentation. *CoRR*, abs/2207.11860, 2022. 1, 4

[7] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset

for structured 3d modeling. In *ECCV (9)*, volume 12354 of *Lecture Notes in Computer Science*, pages 519–535. Springer, 2020. 1, 2, 4

[8] Zishuo Zheng, Chunyu Lin, Lang Nie, Kang Liao, Zhijie Shen, and Yao Zhao. Complementary bi-directional feature compression for indoor 360° semantic segmentation with self-distillation. In *WACV*, pages 4490–4499. IEEE, 2023. 4