Sigma: Siamese Mamba Network for Multi-Modal Semantic Segmentation

Zifu Wan[1] Pingping Zhang[2] Yuhao Wang[2] Silong Yong[1]
Simon Stepputtis[1] Katia Sycara[1] Yaqi Xie[1]
[1]Robotics Institute, Carnegie Mellon University, USA
[2]School of Future Technology, Dalian University of Technology, China

Supplementary Material

A. Experimental Details

During training, we perform data augmentation, including random flipping and scaling with random scales $[0.5, 1.75]$, to all datasets. We adopt VMamba [9] pre-trained on ImageNet [11] as the backbone, which includes three versions, namely VMamba-Tiny, VMamba-Small, and VMamba-Base. The detailed settings of the three models are listed in Table A1. We select AdamW optimizer [6] with weight decay 0.01. The original learning rate is set to $6e^{-5}$ and we employ a poly learning rate schedule with 10 warm-up epochs. We use cross-entropy as the loss function. When reporting testing results on NYU Depth V2 [12] and SUN RGB-D [13] datasets, we use multiple scales $\{0.75, 1, 1.25\}$ according to most previous RGB-Depth semantic segmentation methods [8,20]. We use mean Intersection over Union (mIoU) averaged across semantic classes as the evaluation metric to measure the segmentation performance. For each of the datasets, more implementation details are described as follows.

MFNet dataset. The tiny and small backbones are trained on four 3090Ti GPUs and the base backbone is trained on four A6000 GPUs. We use the original image size of 640×480 for training and inference. The batch size is set to 8 for training. A single 3090Ti GPU is used for inferencing all the models.

PST900 dataset. The tiny and small backbones are trained on two A6000 GPUs. We use the original image size of 1280×720 for training and inference. The batch size is set to 4 for training. A single A6000 GPU is used for inferencing all the models.

NYU Depth V2 dataset. Unlike other methods [2,8] to use HHA format of depth images for training, we directly use raw depth images and we found no apparent performance difference between the formats. We take the whole image with the size 640×480 for training and inference. 4 3090Ti GPUs are used to train the tiny and small backbones with batch size 8, and 4 A6000 GPUs are used to train the base model.

SUN-RGBD dataset. Unlike previous methods which use larger resolution input (730×530 [16,20] or 640×640 [1]), we adopt the input resolution of 640×480 and keep the same training settings as NYU Depth V2 dataset. We also use raw depth images instead of HHA format for training.

Backbone	VSS Block Number				Embedded Dimension
	Stage 1	Stage 2	Stage 3	Stage 4	
VMamba-Tiny	2	2	9	2	96
VMamba-Small	2	2	27	2	96
VMamba-Base	2	2	27	2	128

Table A1. Details about three versions of backbone.

B. Daytime and Nighttime Performance

To explore the effectiveness of our method on daytime and nighttime RGB-T images, we use the MFNet [4] dataset and follow CMX [8] to use 205 daytime images and 188 nighttime images in the test set for evaluation. As shown in Table B2, our method delivers better results on both daytime and nighttime results, demonstrating the effectiveness of our proposed method.

C. Ablation Studies

Apart from the ablation studies on the effect of each of our components, we further conduct experiments on the detailed design of the State Space Models. In Table C3, we compare the effect of the state size in State Space Models and the number of CVSS blocks in our Mamba decoder. From the table, we can find that setting the state size to 4 and the decoder layers to [4,4,4] leads to the optimal result.

Method	Modal	Daytime	Nighttime
FRRN [10]	RGB	40.0	37.3
DFN [19]	RGB	38.0	42.3
BiSeNet [18]	RGB	44.8	47.7
SegFormer-B2 [17]	RGB	48.6	49.2
SegFormer-B4 [17]	RGB	49.4	52.4
MFNet [4]	RGB-T	36.1	36.8
FuseNet [5]	RGB-T	41.0	43.9
RTFNet [14]	RGB-T	45.8	54.8
FuseSeg [15]	RGB-T	47.8	54.6
GMNet [21]	RGB-T	49.0	57.7
CMX (MiT-B2) [8]	RGB-T	51.3	57.8
CMX (MiT-B4) [8]	RGB-T	52.5	59.4
Sigma (VMamba-T)	RGB-T	54.1	59.0
Sigma (VMamba-S)	RGB-T	**55.0**	60.0
Sigma (VMamba-B)	RGB-T	54.1	**60.9**

Table B2. Performance comparison on daytime and nighttime MFNet [4] dataset. We use mIoU (%) for evaluation.

#	Encoder	State Size	Decoder Layers	mIoU (∇)
1	VMamba-T	4	[4, 4, 4]	60.5 (0.0)
2	VMamba-T	4	[3, 3, 3]	60.2 (0.3)
3	VMamba-T	4	[2, 2, 2]	59.4 (1.1)
4	VMamba-T	8	[4, 4, 4]	60.3 (0.2)
5	VMamba-T	16	[4, 4, 4]	59.7 (0.8)

Table C3. Ablation studies of decoder layers and the space size of the state space models on the MFNet [4] dataset.

D. Complexity Comparison of CroMB and Self-Attention

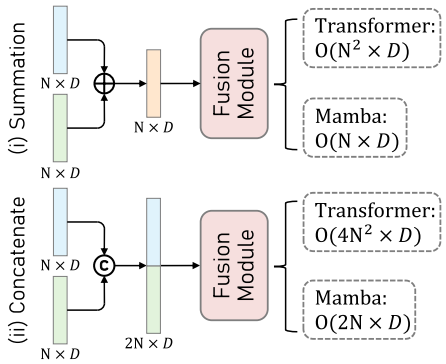

Figure D1. Comparative analysis of complexity across different fusion methods utilizing Transformer and Mamba: Mamba-based fusion approaches significantly reduce complexity by an order of magnitude compared to their Transformer-based counterparts.

As shown in Fig. D1, we compare the theoretical complexity of fusion methods using Mamba and Transformer. This shows the linear scalability advantage of Mamba over quadratic Transformer-based methods. In Fig. D2, we illustrate the qualitative growth in FLOPs as the input sequence length increases. It is evident that our ConM mechanism has much less computation consumption than constituting the

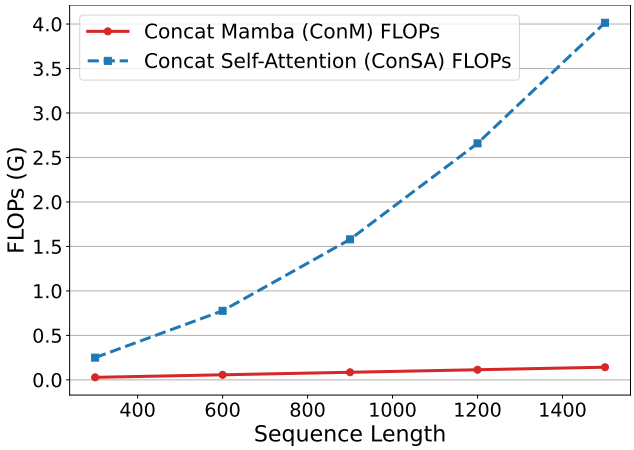

Figure D2. Qualitative computation comparison of Concat Self-Attention (ConSA) and our Concat Mamba (ConM) mechanism.

State Space Model with Self-Attention. This underscores the exceptional efficiency of our proposed ConM in integrating multi-modal features.

Stage	Feature Size			FLOPs (G)	
	Height	Weight	Channel	ConM	ConSA
1	120	160	96	1.82	–
2	60	80	192	1.71	77.89
3	30	40	384	1.65	15.94
4	15	20	768	1.62	8.19

Table D4. Quantitative comparison of computation complexity between Concat Self-Attention (ConSA) and our proposed Concat Mamba (ConM) mechanism.

In Table D4, we compare the floating-point operations per second (FLOPS) of our proposed ConMB and Concat Mamba (ConSA), which employs self-attention instead of SSM. The "Stage" column indicates the four encoding stages, with the input feature size for each fusion block also provided. The findings reveal that ConMB maintains low FLOPs across all stages, whereas the FLOPs for the self-attention mechanism escalate significantly with increases in height and width.

E. Limitations and Future Work

While Sigma has achieved outstanding results in various RGB-X semantic segmentation tasks, two main limitations remain. 1) *Underutilization of Mamba for Longer Sequences:* Mamba's capability to handle extremely long sequences is a significant advantage, particularly beneficial in fusion tasks involving more than two modalities. However, our current exploration primarily focuses on the application of Mamba for two modalities, potentially not fully leveraging its capacity for modeling longer sequences. Future work will aim to investigate Mamba's performance on datasets

with a greater variety of modalities, such as the DELIVER benchmark. This exploration is pivotal for advancing research on enabling autonomous agents to navigate environments using multiple sensors, including RGB, depth, thermal, and LiDAR. 2) *Memory Consumption in the Mamba Encoder:* The Mamba encoder scans image features from four directions, allowing each pixel to assimilate information from its surrounding pixels. This approach, however, quadruples memory usage, posing a challenge for deployment on lightweight edge devices. Future endeavors will seek to incorporate positional information through alternative methods, such as positional encoders, and employ a 1D SSM to diminish computational and memory demands.

F. More Qualitative Results

In Fig. F3 and Fig. F4, we show more qualitative results of our method compared to others.

References

[1] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, pages 348–367. Springer, 2022. 1, 5

[2] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In *European Conference on Computer Vision*, pages 561–577. Springer, 2020. 1, 5

[3] Fuqin Deng, Hua Feng, Mingjian Liang, Hongmin Wang, Yong Yang, Yuan Gao, Junfeng Chen, Junjie Hu, Xiyue Guo, and Tin Lun Lam. Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4467–4473. IEEE, 2021. 4

[4] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. pages 5108–5115, 2017. 1, 2, 4

[5] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian conference on computer vision*, pages 213–228. Springer, 2016. 2

[6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[7] Mingjian Liang, Junjie Hu, Chenyu Bao, Hua Feng, Fuqin Deng, and Tin Lun Lam. Explicit attention-enhanced fusion for rgb-thermal perception tasks. *IEEE Robotics and Automation Letters*, 2023. 4

[8] Huayao Liu, Jiaming Zhang, Kailun Yang, Xinxin Hu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *arXiv preprint arXiv:2203.04838*, 2022. 1, 2, 4, 5

[9] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024. 1

[10] Tobias Pohlen, Alexander Hermans, Markus Mathias, and Bastian Leibe. Full-resolution residual networks for semantic segmentation in street scenes. In *CVPR*, 2017. 2

[11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 1

[12] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 1, 5

[13] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 1

[14] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. 4(3):2576–2583, 2019. 2, 4

[15] Yuxiang Sun, Weixun Zuo, Peng Yun, Hengli Wang, and Ming Liu. Fuseseg: Semantic segmentation of urban scenes based on rgb and thermal data fusion. *IEEE Trans. on Automation Science and Engineering (TASE)*, 2020. 2

[16] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12186–12195, 2022. 1

[17] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 2

[18] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 2

[19] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *CVPR*, 2018. 2

[20] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1136–1147, 2023. 1, 5

[21] Wujie Zhou, Jinfu Liu, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Gmnet: graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation. *IEEE Transactions on Image Processing*, 30:7790–7802, 2021. 2

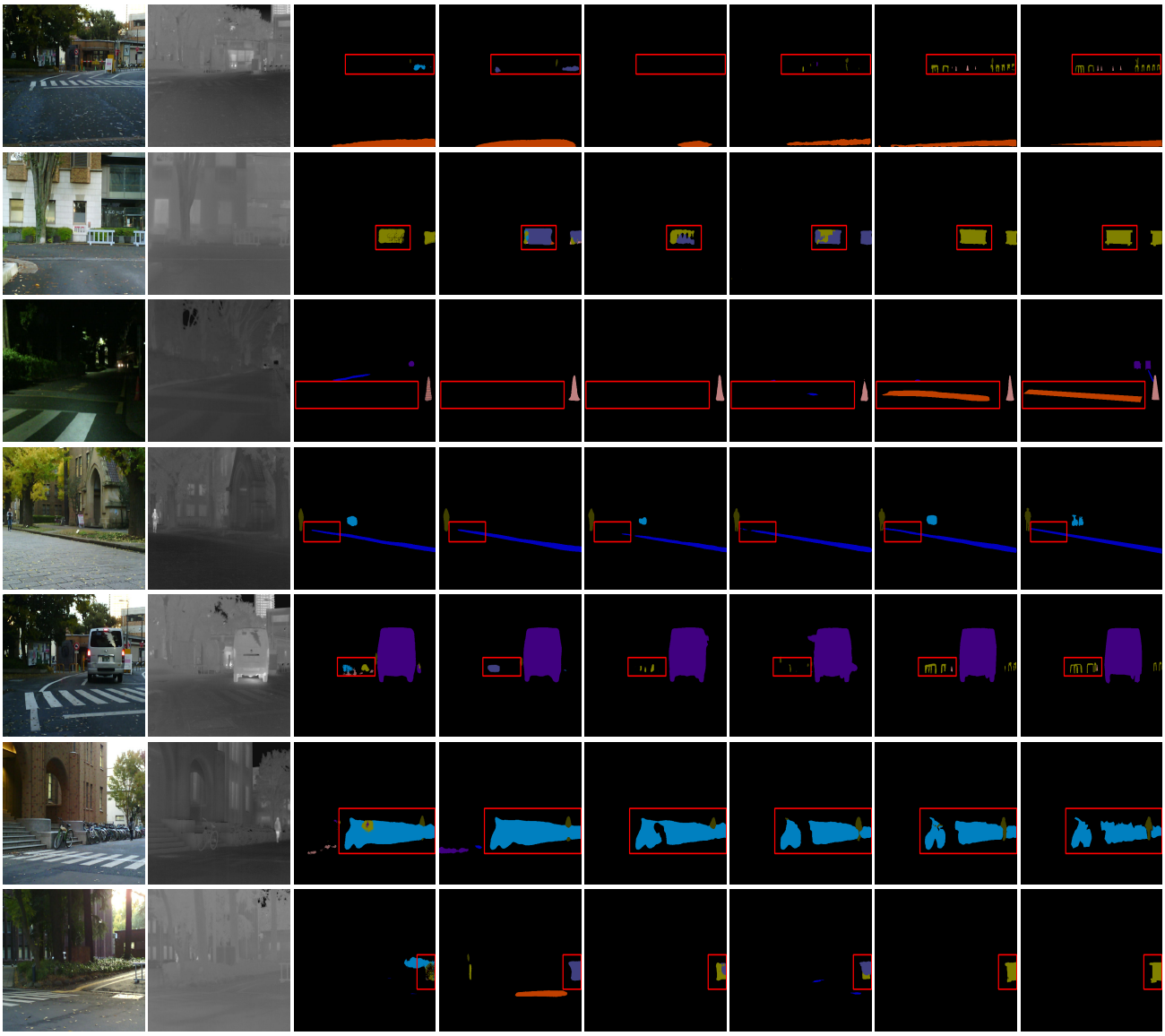

Figure F3. Qualitative comparison on MFNet [4] dataset. More qualitative results can be found in the appendix.

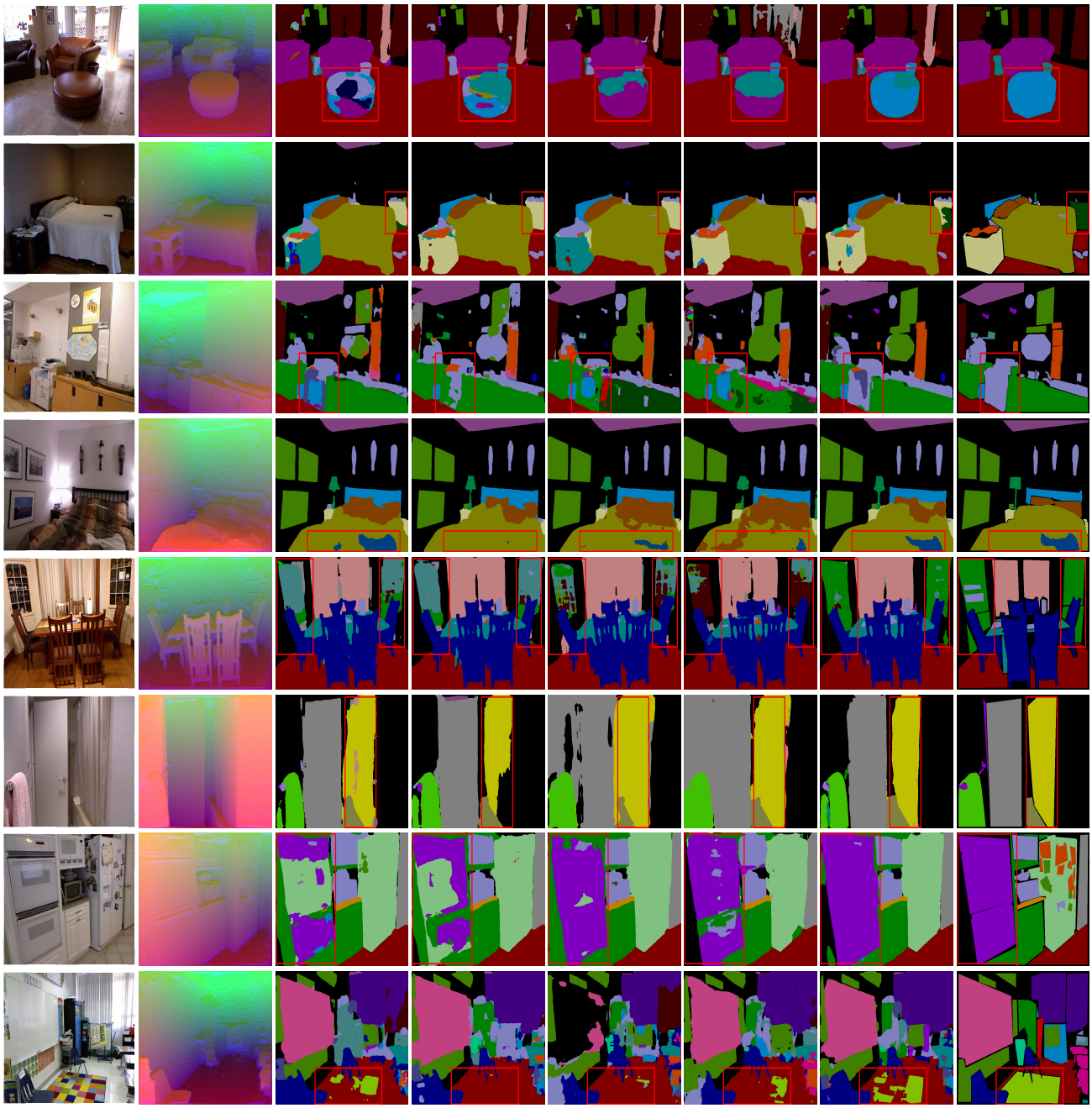

Figure F4. Qualitative comparison on NYU Depth V2 [12] dataset. We use HHA images for better visualization of depth modality. More qualitative results can be found in the appendix.