Unleashing Network Potentials for Semantic Scene Completion

Supplementary Material

This supplementary includes ablation results of AMMNet with different image encoders (Sec. S1), comparison results of using different feature fusion strategies (Sec. S2) and different schemes for alleviating overfitting (Sec. S3), and more visualization results (Sec. S4).

S1. Ablation with Different Image Encoder

This supplementary is for Sec. 5.5 of the main paper. To validate the generalization ability of AMMNet, we supplement the ablation study by substituting the SegFormer-B2 [40] image encoder with the ResNet50 [16] or DeepLabv3 [3]. As Table S1 shows: 1) adopting the DeepLabv3 encoder pre-trained on image segmentation task achieves the best performance, suggesting that enhanced semantic understanding can better facilitate scene completion and prediction of correct semantic categories. The stronger SegFormer-B2 encoder also improves performance over ResNet50, especially the semantic metric SSC-mIoU by 2.7% in the baseline model. This aligns with the fact that superior visual features facilitate semantic prediction; 2) AMMNet maintains consistent performance gains over baseline regardless of the choice of image encoder. Specifically, using a ResNet50, AMMNet improves SC-IoU by 3.7% and SSC-mIoU by 2.3%. With a SegFormer-B2 encoder, the gains are even higher, reaching 4.0% and 2.4% respectively. Notably, AMMNet also achieves considerable improvements of 2.9% on SC-IoU and 1.5% on SSC-mIoU with the DeepLabv3 encoder pre-trained on image segmentation task. The consistent gains verify that AMMNet's robust effectiveness stems from a better unleashing of the network potentials, rather than reliance on specific encoders.

Methods	Image Encoder	SC-IoU	SSC-mIoU
Baseline	ResNet50	70.3%	43.4%
AMMNet		74.0% (↑ 3.7%)	45.7% (↑ 2.3%)
Baseline	Segformer -B2	71.6%	46.1%
AMMNet		75.6% (↑ 4.0%)	48.5% (↑ 2.4%)
Baseline	DeepLabv3	73.4%	54.6%
AMMNet		76.3% (↑ 2.9%)	56.1% (↑ 1.5%)

Table S1. The ablation study of using different image encoder, including ResNet50 [16], Segformer-B2 [40]), and DeepLabv3 [3], in our AMMNet on the test set of NYU [31].

S2. Alternative Fusion Strategies

This supplementary is for Sec. 5.5 of the main paper. To validate the efficacy of cross-modal modulation, we compare it with several widely-adopted alternatives for fusing multi-modal representations including addition, concatenation, refinement with SENet [18], refinement with CBAM [39], and soft selection [34]. Experiments are conducted by replacing all three modulation modules in AMMNet with each scheme. Due to the enormous computational overhead of 3D tasks, we do not consider transformer-based attention methods.

As Table S2 shows, simple fusion schemes like direct addition or concatenation prove insufficient for optimally exploiting cross-modal representations. Incorporating refinement as SENet [18] provides a slight 0.3% SSC-mIoU improvement over the addition. Another alternative, dynamically selecting modalities via soft gating [34], provides 0.6% SSC-mIoU gain over baseline addition. Our proposed cross-modal modulation obtains further noticeable performance gains, elevating SSC-mIoU by 0.6% and SC-IoU by 0.8% over incorporating soft selection [34]. Experiments validate cross-modal modulation facilitates more holistic fusion to better discover synergistic cross-model potential.

Method	Fusion Type	SC-IoU	SSC-mIoU
-	Add	75.2%	47.3%
-	Concat	75.0%	46.8%
SENet [18]	Refine	74.2%	47.6%
CBAM [39]	Refine	74.6%	47.4%
HighWay [34]	SoftSelect	74.8%	47.9%
Ours	Modulation	**75.6%**	**48.5%**

Table S2. Performance comparison of different feature fusion schemes by replacing all three modulation modules in AMMNet with each scheme respectively.

S3. Schemes to Alleviate Overfitting

This supplementary is for Sec. 5.5 of the main paper. To examine the isolated impact of different schemes in

Method	Removed Scheme	SC-IoU	SSC-mIoU
AMMNet†	None	74.4%	47.7%
AMMNet†	Dropout	74.7% (↑ 0.3%)	47.5% (↓ 0.2%)
AMMNet†	Label Smooth	73.4% (↓ 1.0%)	47.3% (↓ 0.4%)
AMMNet†	Data Augment(3D)	74.3% (↓ 0.4%)	47.4% (↓ 0.1%)
AMMNet†	Data Augment(2D)	74.8% (↑ 0.1%)	47.0% (↓ 0.7%)
AMMNet†	$\mathcal{L}_{(D,G)}$	71.6% (↓ 2.8%)	46.1% (↓ 1.6%)

Table S3. The ablation study of different schemes for alleviating overfitting based on AMMNet† on the test set of NYU [31].

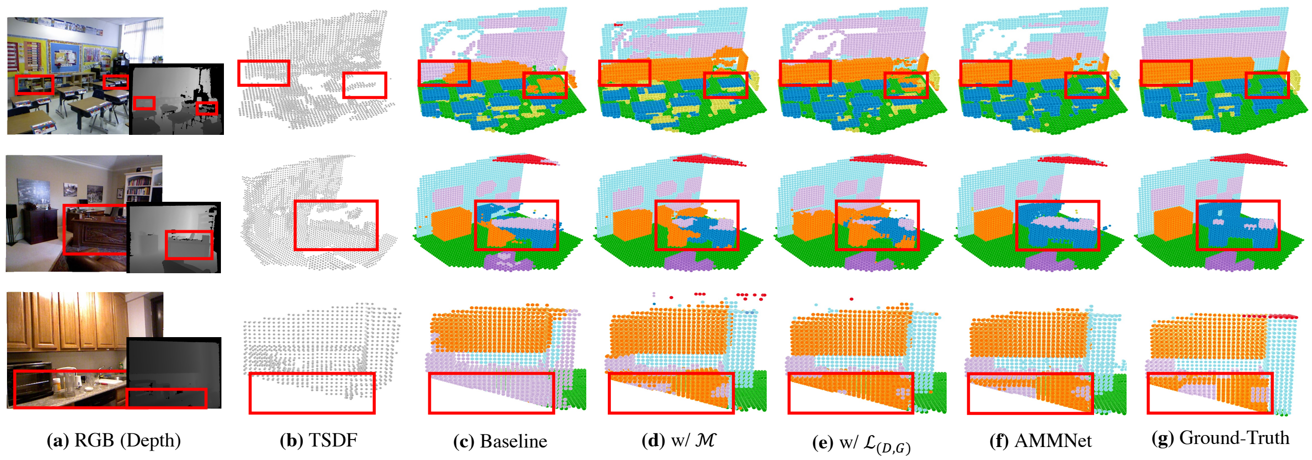

Figure S1. Visualization results for ablation study based on the test set of NYU [31]. The proposed cross-modal modulation $\mathcal{M}$ (in (d)) and adversarial training scheme $\mathcal{L}_{(D,G)}$ (in (e)) improve the baseline with better volumetric occupancy and semantics. Combining both (in (f)) achieves the best results.

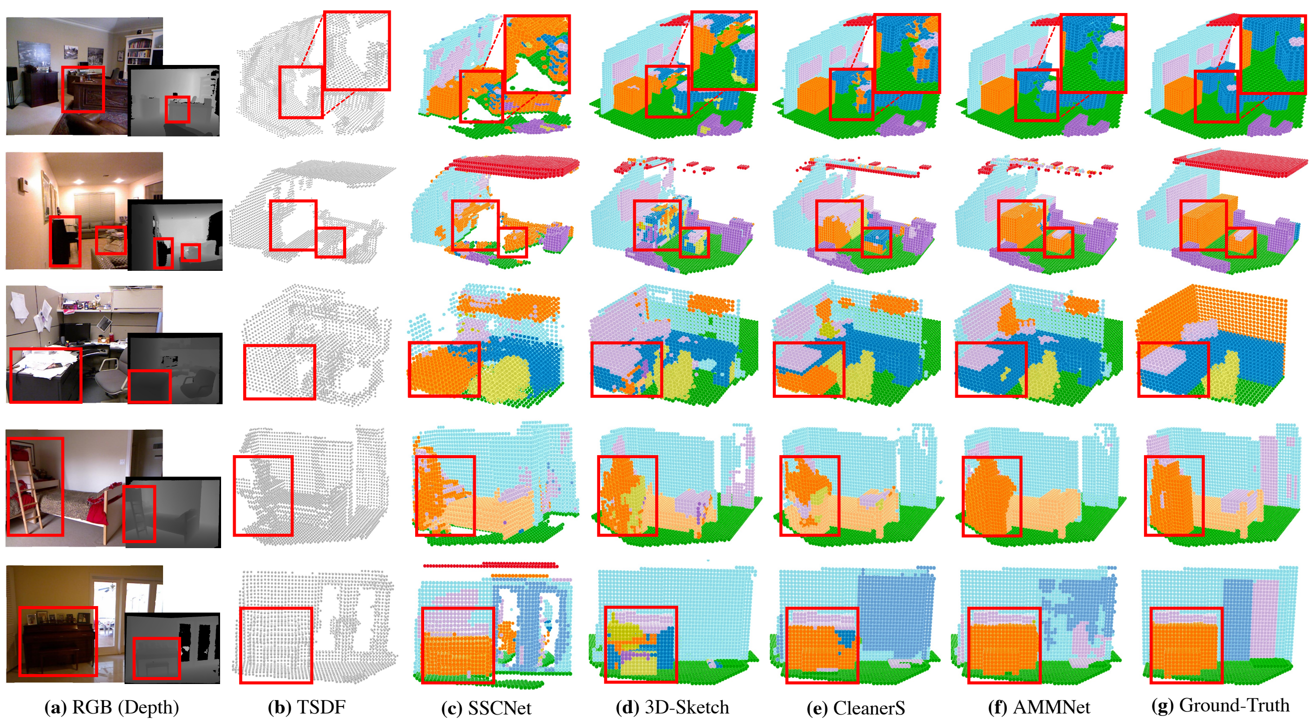

Figure S2. More qualitative comparisons on challenging indoor scenes from the test set of NYU [31] with state-of-the-art methods, including SSCNet [32], 3D-Sketch [4], and CleanerS [37].

alleviating overfitting, we ablate different modules from the AMMNet†, where all cross-modal modulation modules are removed. Schemes analyzed include dropout, label smoothing, 2D/3D data augmentation, and our adversarial training scheme $\mathcal{L}_{(D,G)}$. As Table S3 reports, removing most regularizers causes minor performance drops, confirming their auxiliary effects. Specifically, simple schemes like dropout exhibit weaker regularization power, as evaluated by the minor 0.2% SSC-mIoU drop when excluded. Meanwhile, complementary strategies like label smoothing [35] (0.4% SSC-mIoU drop) and 2D/3D augmentation (0.1%/0.7% SSC-mIoU reduction) help prevent learned biases and memorization. Findings confirm the necessity of strong regularization guided by domain insights.

Notably, excluding our $\mathcal{L}_{(D,G)}$ degrades results substantially by 2.8% in SC-IoU and 1.6% SSC-mIoU. This verifies the vital role of our custom adversarial training scheme in alleviating overfitting. We advise blending it with existing methods like augmentation and label smoothing [35] to maximize performance.

S4. More Visualization Results

This supplementary is for Sec. 5.4 and Sec. 5.3 of the main paper. As Figure S1 shows, incorporating the proposed cross-modal modulation $\mathcal{M}$ (in (d)) improves semantic perception over the baseline (in (c)), correcting erroneous predictions. Building on this, additionally introducing adversarial training (our AMMNet in (f)) further unleashes model potentials, attaining high-fidelity outputs better approximating the ground truth voxels (in (g)). In Figure S2, we supplement more visual examples compared to state-of-the-art methods.

References

[1] Mohammad Mahdi Bejani and Mehdi Ghatee. A systematic review on overfitting control in shallow and deep neural networks. *Artificial Intelligence Review*, pages 1–48, 2021. 3

[2] Yingjie Cai, Xuesong Chen, Chao Zhang, Kwan-Yee Lin, Xiaogang Wang, and Hongsheng Li. Semantic scene completion via integrating instances and scene in-the-loop. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 324–333, 2021. 5, 6

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 5, 6, 1

[4] Xiaokang Chen, Kwan-Yee Lin, Chen Qian, Gang Zeng, and Hongsheng Li. 3d sketch-aware semantic scene completion via semi-supervised structure prior. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4193–4202, 2020. 1, 3, 4, 5, 6, 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 6

[6] Haotian Dong, Enhui Ma, Lubo Wang, Miaohui Wang, Wuyuan Xie, Qing Guo, Ping Li, Lingyu Liang, Kairui Yang, and Di Lin. Cvsformer: Cross-view synthesis transformer for semantic scene completion. In *IEEE International Conference on Computer Vision (ICCV)*, pages 8874–8883, 2023. 5, 6

[7] Aloisio Dourado, Teofilo E De Campos, Hansung Kim, and Adrian Hilton. Edgenet: Semantic scene completion from a single rgb-d image. In *International conference on pattern recognition (ICPR)*, pages 503–510, 2021. 2

[8] Aloisio Dourado, Frederico Guth, and Teofilo de Campos. Data augmented 3d semantic scene completion with 2d segmentation priors. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3781–3790, 2022. 6

[9] Chenzhuang Du, Tingle Li, Yichen Liu, Zixin Wen, Tianyu Hua, Yue Wang, and Hang Zhao. Improving multi-modal learning with uni-modal teachers. *arXiv preprint arXiv:2106.11059*, 2021. 3

[10] Dajun Du, Kang Li, and Minrui Fei. A fast multi-output rbf neural network construction method. *Neurocomputing*, 73 (10-12):2196–2202, 2010. 3

[11] Michael Firman, Oisin Mac Aodha, Simon Julier, and Gabriel J Brostow. Structured prediction of unobserved voxels from a single depth image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5431–5440, 2016. 2, 5, 6, 7

[12] Mudasir A Ganaie, Minghui Hu, AK Malik, M Tanveer, and PN Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, 2022. 3

[13] Martin Garbade, Yueh-Tung Chen, Johann Sawatzky, and Juergen Gall. Two stream 3d semantic scene completion. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 0–0, 2019. 5

[14] Christian Garbin, Xingquan Zhu, and Oge Marques. Dropout vs. batch normalization: an empirical study of their impact to deep learning. *Multimedia Tools and Applications*, 79: 12777–12815, 2020. 3

[15] Yuxiao Guo and Xin Tong. View-volume network for semantic scene completion from a single depth image. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 726–732, 2018. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[17] Di Hu, Xuelong Li, et al. Temporal multimodal learning in audiovisual speech recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3574–3582, 2016. 3

[18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018. 1

[19] Jie Li, Yu Liu, Dong Gong, Qinfeng Shi, Xia Yuan, Chunxia Zhao, and Ian Reid. Rgbd based dimensional decomposition residual network for 3d semantic scene completion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7693–7702, 2019. 1, 3, 5, 6

[20] Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. Anisotropic convolutional networks for 3d semantic scene completion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3351–3359, 2020. 3, 5, 6

[21] Jie Li, Qi Song, Xiaohu Yan, Yongquan Chen, and Rui Huang. From front to rear: 3d semantic scene completion through planar convolution and attention-based network. *IEEE Transactions on Multimedia*, 25:8294–8307, 2023. 5, 6

[22] Shice Liu, Yu Hu, Yiming Zeng, Qiankun Tang, Beibei Jin, Yinhe Han, and Xiaowei Li. See and think: Disentangling semantic scene completion. In *Advances in Neural Informa-*

tion Processing Systems (NeurIPS), pages 263–274, 2018. 1, 3

[23] Yu Liu, Jie Li, Qingsen Yan, Xia Yuan, Chunxia Zhao, Ian Reid, and Cesar Cadena. 3d gated recurrent fusion for semantic scene completion. *arXiv preprint arXiv:2002.07269*, 2020. 1

[24] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2016. 6

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2018. 6

[26] Jian-Hao Luo, Jianxin Wu, and Weiyao Lin. Thinet: A filter level pruning method for deep neural network compression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5058–5066, 2017. 3

[27] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR*, pages 8238–8247, 2022. 3

[28] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017. 3

[29] Christoph B Rist, David Emmerichs, Markus Enzweiler, and Dariu M Gavrila. Semantic scene completion using local deep implicit functions on lidar data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7205–7218, 2021. 3

[30] Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. Text data augmentation for deep learning. *Journal of big Data*, 8:1–34, 2021. 3

[31] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision (ECCV)*, pages 746–760, 2012. 1, 2, 5, 6, 7, 8

[32] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1746–1754, 2017. 2, 5, 6

[33] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 3

[34] Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 1

[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 5, 2, 3

[36] Jiaxiang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng. Not all voxels are equal: Semantic scene completion from the point-voxel perspective. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 2352–2360, 2022. 3

[37] Fengyun Wang, Dong Zhang, Hanwang Zhang, Jinhui Tang, and Qianru Sun. Semantic scene completion with cleaner self. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 867–877, 2023. 1, 2, 3, 4, 5, 6, 7

[38] Xuzhi Wang, Di Lin, and Liang Wan. Ffnet: Frequency fusion network for semantic scene completion. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 2550–2557, 2022. 5, 6

[39] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 1

[40] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 12077–12090, 2021. 6, 1

[41] Pingping Zhang, Wei Liu, Yinjie Lei, Huchuan Lu, and Xiaoyun Yang. Cascaded context pyramid for full-resolution 3d semantic scene completion. In *IEEE International Conference on Computer Vision (ICCV)*, pages 7801–7810, 2019. 5, 6

[42] Min Zhong and Gang Zeng. Semantic point completion network for 3d semantic scene completion. In *European Conference on Artificial Intelligence (ECAI)*, pages 2824–2831, 2020. 2

[43] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2016. 3