Multi-modal Frequency Decomposition Network for Semantic Scene Completion

Supplementary Material

A. Detailed Architecture

We present the detailed architecture of MFDNet in Figure 1. MFDNet is a lightweight network that integrates frequency processing with limited layers of convolution and downsampling, achieving a balance between modality alignment and detail retainment.

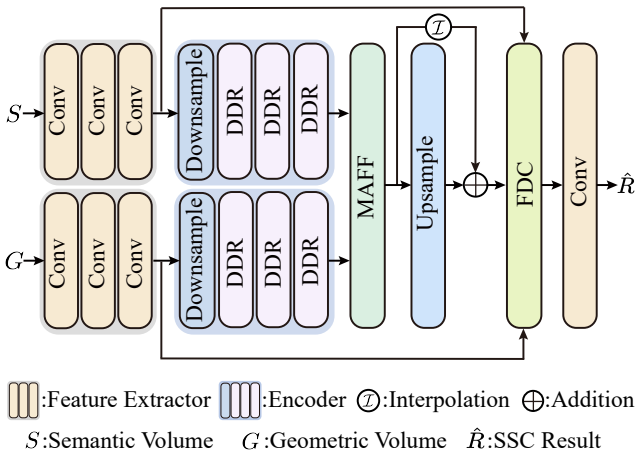


Figure 1. Detailed architecture of MFDNet.

B. Implementation Details

We implement MFDNet with PyTorch and train it on RTX3090 GPUs with a mini-batch of 16 for 500 epochs. We accept SGD as the optimizer with an initial learning rate of 0.05 and a momentum of 0.9. We employ Poly scheduler to adjust the learning rate with power of 0.9. And we pretrain Deeplabv3 [1] as the 2D semantic segmentation network to enable a fair comparison with [2] and [7].

We set the scaling factor α in MAFF to 2.0, and β in FDC to 2.0. In MAFF, we set the numbers of bands k to 3, and the radii of filters $\{\gamma_0, \gamma_1\}$ to $\{0.15, 0.35\}$. In FDC, we set the radius of filter γ_S in the semantic branch to 0.35 and the radius of filter γ_G in the geometric branch to 0.25. Besides, we set the main loss weight $\lambda_{SSC} = 1$ and the auxiliary loss weights $\lambda_S = \lambda_G = 0.25$.

C. Datasets and Metrics

We validate the effectiveness of MFDNet on NYUv2 [5] and NYUCAD [3] datasets. Both of them consist of 1449 RGB-D image pairs, 795 for training and 654 for test. The voxels in 3D annotations are divided into 1 class for free and 11 classes with semantics. The main difference between them is that the depth images in NYUv2 are captured with real sensors and the depth images in NYUCAD are produced with 3D models. The resolution of the voxel-based scene is $60 \times 36 \times 60$.

And we evaluate MFDNet according to SSCNet [6] with Precision, Recall and Intersection over Union (IoU) for scene completion, IoU of each semantic class and mean of them (mIoU) for semantic scene completion.

D. Comparison of Efficiency and Accuracy

To analyze the complexity of our approach, we compare MFDNet with recent state-of-the-art SSC methods in terms of parameters, FLOPs, and inference time in Table 1. MFDNet employs the fewest parameters while achieving the best semantic scene completion accuracy. The FLOPs slightly increase because FDC operates at the full resolution, where it identifies the details lost during downsampling and convolution and adaptively compensates for them. This design introduces a minor increase in FLOPs but substantially strengthens multi-modal feature alignment, thereby contributing to improved SSC performance. At the same time, MFDNet achieves a moderate inference time. Overall, by maintaining the lowest parameter complexity, MFDNet effectively balances computational efficiency and accuracy.

Table 1. Comparison of Efficiency and Accuracy. Params are measured in millions (M), FLOPs are measured in billions (G), and Time denotes the inference time per sample in milliseconds (ms).

Method	Params	FLOPs	Time	IoU	mIoU
FFNet [8]	97.5	330.5	67	71.8	44.4
CVSformer [2]	439.7	**260.5**	187	73.7	52.6
SG-SSC [4]	86.6	278.4	**37**	74.3	54.6
AMMNet [7]	22.2	288.8	89	76.3	56.1
MFDNet	**10.1**	487.8	117	**77.1**	**57.0**

E. Supplementary Ablation Study

E.1. Ablation on Framework Design

We evaluate the effectiveness of the framework, including the impact of the encoder, downsampling rate, and the interpolation across the decoder, as shown in Table 2. All results in this table are obtained without the MAFF and FDC. In the first and second columns, we compare the performance of single- and dual-encoder. With the same other settings, the dual-encoder achieves better performance than the single-encoder, with only a modest increase in inference time and memory consumption. This indicates that dual-encoder can better capture the multi-modal features individually.

Table 2. Impact of different network components. "Int." indicates the interpolation across the decoder. All the experiments are conducted without MAFF and FDC.

Encoder		Down			Int.	IoU	mIoU	Time	Mem.
single	dual	1	1/2	1/4		(%)	(%)	(ms)	(GB)
✓		✓				74.1	53.7	95.76	5.27
✓		✓			✓	75.1	54.1	99.19	5.27
✓			✓			74.9	53.6	**80.82**	**4.85**
✓			✓		✓	74.8	54.0	81.39	4.96
✓				✓		73.2	52.1	81.77	4.92
✓				✓	✓	73.4	52.4	83.23	5.03
	✓	✓				73.9	53.9	115.37	5.40
	✓	✓			✓	75.1	54.3	118.73	5.40
	✓		✓			73.5	53.7	82.25	4.88
	✓		✓		✓	**75.2**	**54.5**	82.48	5.00
	✓			✓		73.5	53.5	82.74	4.94
	✓			✓	✓	74.0	54.0	84.34	5.05

In the third to fifth columns, we compare the performance of different downsampling rates, where the rate 1 means that the network performs on the full resolution, 1/2 means that downsampling the features once with downsampling rate 1/2, and 1/4 means that downsampling the features twice and each downsampling layer has the downsampling rate of 1/2. And when the rate is set to 1/4 we set skip connection across the appended downsampling and upsampling layers. The performance of 1/2 downsampling rate is better than 1/4, as one more downsampling layer and more convolutions cause more detail loss, particularly when the full resolution is as small as $60 \times 36 \times 60$. Besides, the performance of network without downsampling is similar to the 1/2 downsampling rate but needs more computational resources, so we adopt the rate of 1/2. The sixth column compares the networks with and without the interpolation across the decoder, showing that interpolation enhances performance by facilitating multi-scale aggregation in MFDNet.

E.2. Internal Design of MAFF

In Table 3, we compare the performance of the network with different numbers of frequency bands, which is defined as k in MAFF. And we set progressively subdivided filter radii in these settings. As k increases from 1 to 3, the performance of the network improves significantly and the best performance is (77.1% IoU, 57.0% mIoU) when k is 3. This indicates that decomposing the feature into multiple frequency bands effectively optimizes the modeling of intra-modal dependencies and promotes the learning of inter-modal alignment. When k is increased from 4 to 5, the performance remains similar to the setting of 3, suggesting that a moderate level of decomposition is sufficient to capture the misaligned information. Further increasing the number of bands does not provide additional information and instead increases resource consumption.

Table 3. Number of frequency bands in MAFF.

Frequency Band		IoU	mIoU
Number	Radii of Filter	(%)	(%)
1	-	76.4	56.1
2	0.35	76.7	56.7
3	0.15,0.35	**77.1**	**57.0**
4	0.15,0.35,0.50	77.1	56.9
5	0.15,0.35,0.50,0.75	76.9	57.0

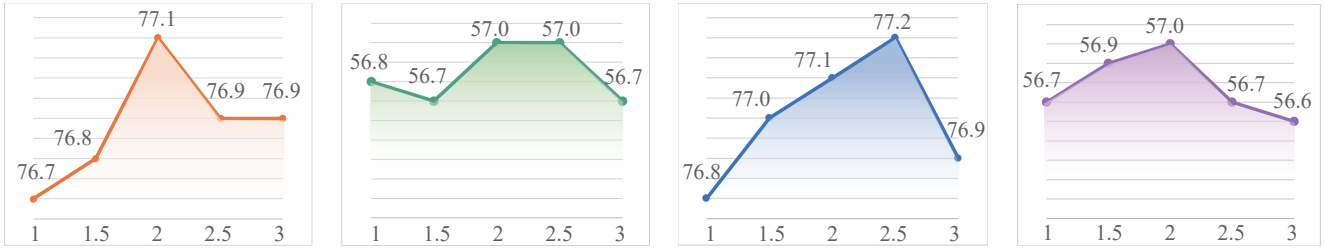
Table 4. Internal ablation study on the compensation modality in FDC.

Semantic Fea.	Geometric Fea.	IoU(%)	mIoU(%)
		75.8	55.7
✓		75.1	55.6
	✓	76.5	56.1
✓	✓	**77.1**	**57.0**

E.3. Internal Design of FDC

We conduct experiments on the compensation modality in FDC in Table 4 because different modalities contribute differently to the network. Without applying any compensation modality to the coarse feature R_I in the first row, we achieve (75.8% IoU and 55.7% mIoU). In the second row, we add high frequency cues in semantic modality to the network and get (75.1% IoU, 55.6% mIoU). This indicates that compensating R_I with semantic details only, while neglecting geometric details, fails to achieve fine-detail alignment and instead disrupts the overall global alignment in R_I. In the third row, we add geometric modality to the network and get (76.5% IoU, 56.1% mIoU). The improvement can be largely attributed to the inclusion of geometric details, which provide clearer spatial constraints and help reduce the difficulty of prediction in semantic scene completion. In the last row, we add both of them to the network and get the best performance of (77.1% IoU, 57.0% mIoU). This indicates a cross-modal complementarity between semantic and geometric representations, which contributes to more accurate and structurally consistent completion.

In Table 5, we conduct experiments on the derivation of the compensated high-frequency components in FDC, where the derivation is obtained from different stages of the network, to explore their capability of providing useful details. In the first row, the coarse feature R_I is not compensated with any high-frequency components. In the second row, the high-frequency component is decomposed from R_I itself and used for compensation. The observed perfor-



(a) IoU- α (b) mIoU- α (c) IoU- β (d) mIoU- β

Figure 2. Internal study of scaling factors α and β.

Table 5. Derivation of the compensated high-frequency components in FDC, where the derivations are obtained from different stages of the network.

Method	IoU(%)	mIoU(%)
-	75.8	55.7
coarse feature R_I	75.6	55.6
fused feature F_M	76.0	55.9
shallow features F_S and F_G	**77.1**	**57.0**

mance degradation suggests that compensating with high-frequency components decomposed from the coarse feature itself provides limited novel information, thereby hindering the learning of effective weights and potentially introducing noise that disrupts the completion process. In the third row, the fused feature F_M generated by MAFF is first upsampled via trilinear interpolation to match the resolution of R_I. Its high-frequency component is then decoupled and used to compensate R_I, resulting in a slight improvement of 0.2% in IoU and 0.2% in mIoU. This indicates that the feature calibrated by MAFF retain some useful detail information. However, due to the loss of fine-grained details caused by downsampling, the effectiveness of this compensation is limited. Finally, the high-frequency components decomposed from the shallow semantic feature F_S and shallow geometric feature F_G are used to compensate R_I, achieving the best performance of (77.1% IoU, 57.0% mIoU). This demonstrates that shallow features preserve abundant local details and provide more effective high-frequency cues for enhancing the coarse feature R_I.

In Table 6, we conduct experiments on the filter radius γ_S of semantic and γ_G of geometric components in FDC. To reduce the burden of hyperparameter tuning, we set γ_S and γ_G within the range [0.2, 0.4] to extract useful high-frequency cues. We get the best performance when γ_S is set to 0.35 and γ_G is set to 0.25. A smaller radius may lead to insufficient valid high-frequency information, whereas an excessively large radius can introduce noise due to irrelevant components.

Table 6. Internal ablation study on the semantic radius γ_S and geometric radius γ_G in FDC.

radius of sem.	radius of geo.	IoU(%)	mIoU(%)
0.35	0.40	76.6	56.7
0.35	0.35	76.8	56.7
0.35	0.30	76.9	56.9
0.35	0.25	77.1	**57.0**
0.35	0.20	**77.2**	56.8
0.40	0.25	76.8	56.8
0.30	0.25	76.9	56.9
0.25	0.25	76.8	56.7
0.20	0.25	76.7	56.6

E.4. Analysis of Scaling Factors

In Figure 2, we conduct experiments on the scaling factors α and β. We get the best performance when α and β are set to 2.0. In this way, the learned weights are limited to 0.0-2.0, which can better meet the necessity of MAFF and FDC, while ensuring the training stability. When the scaling factors are lower than 1.0, all the components will be suppressed, limiting the calibration of misalignment and adaptive fusion of modalities. When the factors are set too large, the maximum values of weights are far greater than required, and this is unfavorable for training with the sigmoid activation function.

E.5. Analysis of Loss Weight Factors

In Table 7, we investigate the influence of the auxiliary semantic loss weight λ_S and geometric loss weight λ_G on the overall performance. In the first row, we remove both auxiliary losses and train the network solely with the SSC loss $\mathcal{L}_{SSC}$. In the subsequent rows, we gradually increase the weights of the auxiliary losses. The best performance is obtained when λ_S and λ_G are set to 0.25, indicating that appropriately weighted auxiliary supervision can effectively facilitate multi-modal feature calibration and alignment. However, further increasing the auxiliary loss weights leads to performance degradation, as the auxiliary objectives begin to overwhelm the SSC loss, leading the model to neglect detail completion at full resolution.

Table 7. Ablation study of the auxiliary loss weight factors λ_S and λ_G.

λ_S	λ_G	IoU(%)	mIoU(%)
0	0	76.6	56.4
0.25	0.25	**77.1**	**57.0**
0.5	0.5	76.3	56.3
0.75	0.75	76.4	56.0
1	1	75.9	55.6

F. Additional Visualization Results

We provide additional ablation results for MAFF and FDC in Figure 3 to further demonstrate the effectiveness of our method, as well as additional visualization results in Figure 4 for comparison with other methods on the NYUv2 test set.

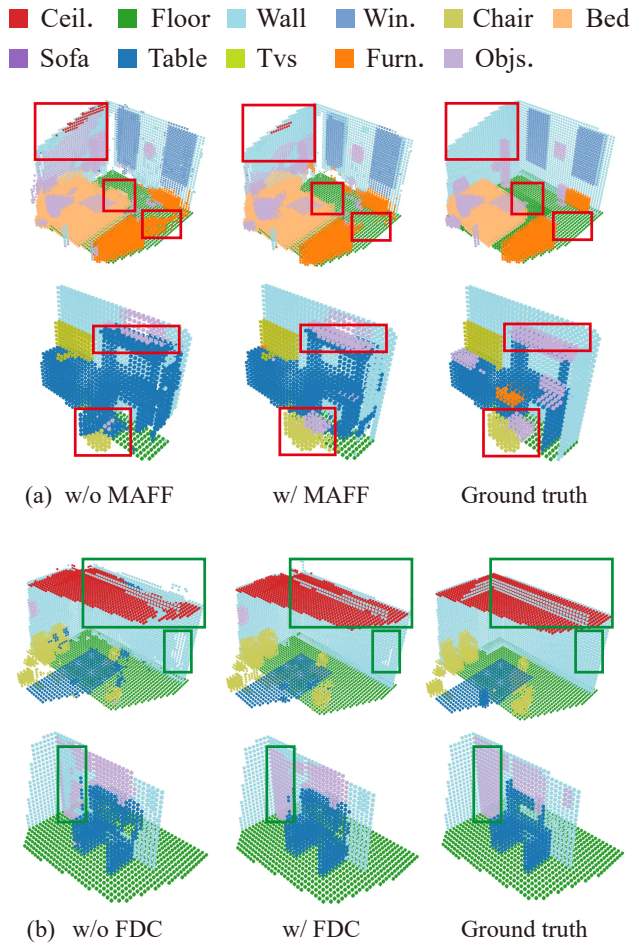


Figure 3. The additional ablation results of MAFF (a) and FDC (b) on NYUv2 test set.

References

[1] Liang-Chieh Chen. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1

[2] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8874–8883, 2023. 1

[3] Yu-Xiao Guo and Xin Tong. View-volume network for semantic scene completion from a single depth image. *arXiv preprint arXiv:1806.05361*, 2018. 1

[4] Xianzhu Liu, Haozhe Xie, Shengping Zhang, Hongxun Yao, Rongrong Ji, Liqiang Nie, and Dacheng Tao. 2d semantic-guided semantic scene completion. *International Journal of Computer Vision*, 133(3):1306–1325, 2025. 1

[5] 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

[6] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1746–1754, 2017. 1

[7] Fengyun Wang, Qianru Sun, Dong Zhang, and Jinhui Tang. Unleashing network potentials for semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10314–10323, 2024. 1

[8] Xuzhi Wang, Di Lin, and Liang Wan. Ffnet: Frequency fusion network for semantic scene completion. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2550–2557, 2022. 1

Figure 4. More completion results of different methods on NYUv2 test set.