Enhancing Mamba Decoder with Bidirectional Interaction in Multi-Task Dense Prediction

Supplementary Material

A. Details of DBIM

To further validate the effectiveness of our method, we present Dilated BIM (DBIM), a lightweight version of BIM, which achieves superior performance with reduced computational complexity and parameter count compared to MTMamba. In DBIM, we replace MS-Scan with a more lightweight variant, DMS-Scan, which conducts sparse scanning within each scanning branch $\mathcal{B}$. Specifically, as shown in Fig. 7, we perform dilated sampling in generating multi-scale sequences from image features instead of using all tokens. When restoring sequences to image features, we perform linear interpolation. These operations do not introduce any parameters and exhibit a reduced computational burden due to sampling a subset of tokens for modeling.

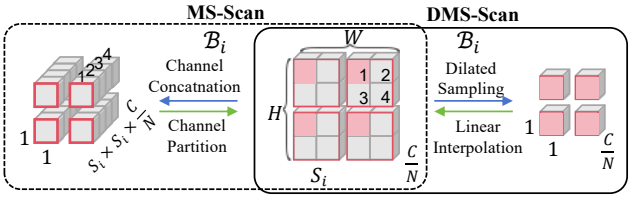

Figure 7. Comparison of MS-Scan and DMS-Scan.

B. More Ablation Studies

Effect of scan mode order in BI-Scan. We performed an ablation study to assess the impact of the scanning order in BI-Scan, containing two sequences: TF → PF (Task-First mode then Position-First mode) and PF → TF (the reverse). The results in Tab. 9 show that the TF → PF setting achieves better performance. Importantly, both combined strategies surpass the performance of using either TF or PF individually, which verifies the inherent complementarity between the two scanning modes.

Table 9. Effect of scan mode order in BI-Scan.

Setting	Semseg mIoU ↑	Depth RMSE↓	Normal mErr ↓	Boundary odsF ↑	FLOPs (G) ↓	# Params (M) ↓
TF	56.36	0.4905	18.67	78.70	510	289
PF	56.96	0.4883	18.68	78.70	510	289
TF → PF	**57.11**	0.4856	**18.66**	**78.80**	547	290
PF → TF	56.55	**0.4806**	18.71	78.70	547	290

Model efficiency with varying task quantities. To further validate the linear complexity of our method, we conducted an experiment on the NYUD dataset with a varying number of tasks. Specifically, we benchmarked our model's performance using 2, 3, and 4 tasks. The results, presented in Tab. 10, show that as the number of tasks increases, the incremental computational cost (GFLOPs) and the number of additional parameters both remain constant for each new task. This observation empirically verifies that our model's complexity scales linearly with the number of tasks.

Table 10. Model efficiency with varying task quantities.

Methods	Semseg mIoU ↑	Depth RMSE↓	Normal mErr ↓	Boundary odsF ↑	FLOPs (G) ↓	# Params (M) ↓
2 tasks	57.20	0.4715	-	-	375	304
3 tasks	55.91	0.4891	18.63	-	461(**+86**)	346(**+42**)
4 tasks	57.40	0.4733	18.55	78.72	547(**+86**)	388(**+42**)

C. More Visual Comparison Results

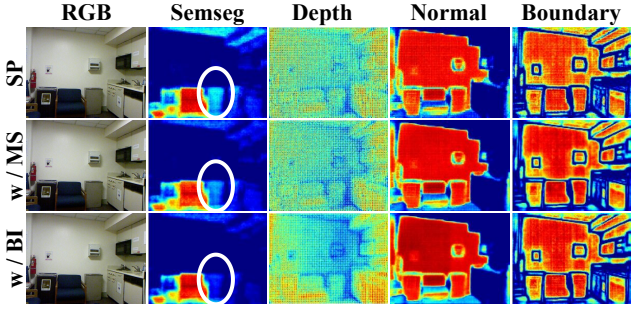

Figure 8. Effect of BI-Scan and MS-Scan on attention patterns.

Effect of BI-Scan and MS-Scan on Task Attention Patterns. To systematically validate the efficacy of our proposed scanning mechanisms in task representation enhancement, we conduct a quantitative visual analysis of attention patterns in the final MFR block, as illustrated in Figure 8. Our comparative investigation examines three distinct feature configurations: (a) Task-specific features (**SP**), (b) Features enhanced by MS-Scan in MSST block (**W/ MS**), and (c) Features refined via the BI-Scan in BCFR block (**W/ BI**). The results reveal two critical insights: First, the MS-Scan mechanism substantially enriches task-specific feature details through multi-scale contextual integration. Second, the BI-Scan induces task-aligned attention redistribution, effectively suppressing irrelevant spatial responses while amplifying task-critical regions. This synergistic effect is demon-

strated by the progressive attention focusing observed in white-circled areas of Fig. 8.

Effect of Bidirectional Scan on Task Attention Patterns. To validate the bidirectional scan efficacy in the BI-Scan mechanism, we conduct a controlled comparative study of task attention patterns in the final MFR block under two experimental configurations: (1) Unidirectional modeling (**w/o BD**) with forward scanning only, and (2) Bidirectional modeling (**w/ BD**), while maintaining identical feature channel dimensions for fair comparison. As shown by the white-circled regions in Fig. 9, in the upper panel exemplars, bidirectional scanning exhibits enhanced attention localization in task-critical regions while preserving the structural integrity of target areas. The lower panel reveals the model without bidirectional scanning, which erroneously groups architecturally distinct elements (doors/walls) due to a flawed understanding of the scene structure. whereas bidirectional implementation achieves precise separation through detailed and comprehensive cross-task interactions.

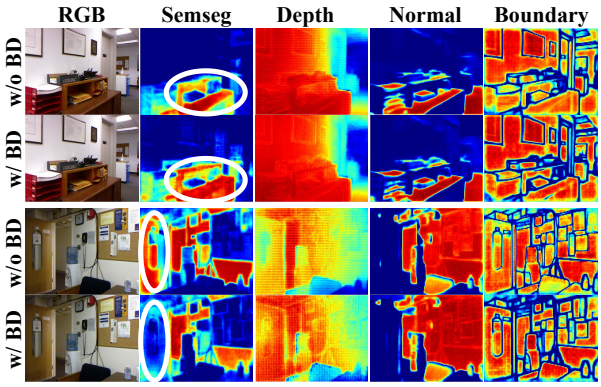

Figure 9. Effect of bidirectional scan on attention patterns.

Qualitative Comparison with state-of-the-art method. We present more qualitative results compared with the SOTA methods. In Figs. 10 to 12, the results indicate that our method generates more detailed multi-task predictions, as highlighted in the circled regions.

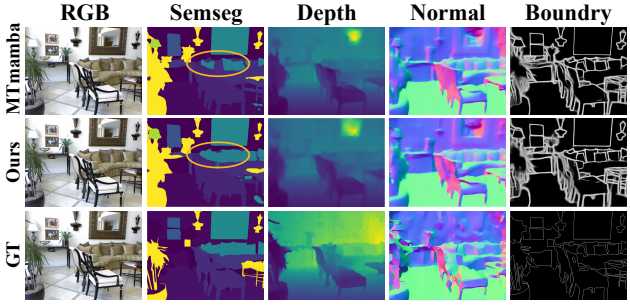

Figure 10. More qualitative comparison on NYUD-v2.

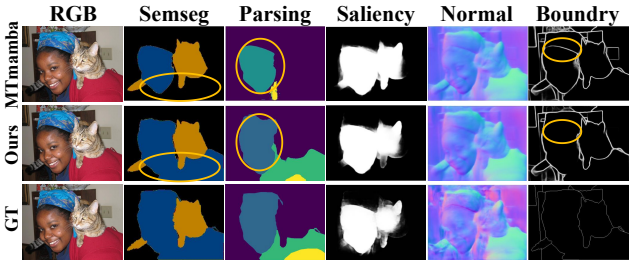

Figure 11. More qualitative comparison on Pascal-Context.

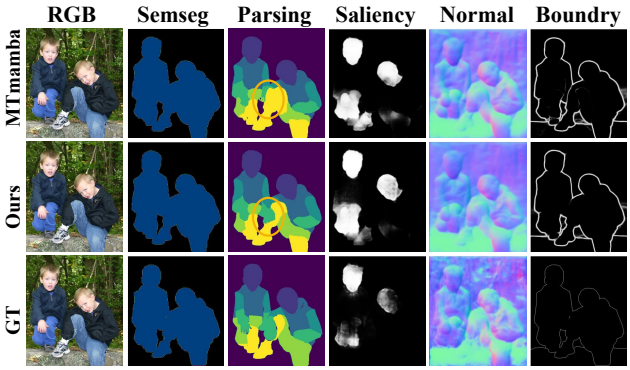

Figure 12. More qualitative comparison on Pascal-Context.