
MANGO: Multimodal Attention-based Normalizing Flow Approach to Fusion Learning

Anonymous Author(s)

Affiliation

Address

email

1 A Additional Ablation Studies

2 **Effectiveness of Number of Cross-Attention Blocks.** We conducted an ablation study with 16
3 cross-attention blocks. As shown in Table 1, although using more cross-attention blocks will increase
4 the computation, it helps to enhance the model performance.

Table 1: Effectiveness of Number of Cross-Attention Blocks.

# Blocks	NYUv2			SUN RGBD		
	Acc.	mAcc.	mIoU	Acc.	mAcc.	mIoU
6	77.5	65.8	52.3	79.6	60.5	48.1
8	78.1	65.3	54.1	84.4	60.0	51.4
12	81.5	71.6	59.2	83.9	67.2	54.1
16	83.1	75.1	61.7	85.4	68.7	55.6

5 **Computational Cost.** As shown in Table 2, the parameters, GFLOPs, and inference time of our
6 method are competitive with prior methods. Meanwhile, we achieved state-of-the-art performance on
7 two segmentation benchmarks.

Table 2: The Comparison of Computational Cost.

Method	NYUd2 mIoU	SUN RGB-D mIoU	PARAMS	GFLOPS	Inference Time
TokenFusion [2]	54.2	53.0	45.9M	108	126 ms
GeminiFusion [1]	57.7	53.3	75.8M	174	153 ms
MANGO	59.2	54.1	72.9M	152	144 ms

8 **Attention Visualization.** As shown in Figure 1, our Invertible Cross-Attention layer can capture the
9 attention interaction from the region in the depth image (red box) to the RGB image. This result
10 has illustrated the effectiveness of our proposed attention layer in capturing the correlation across
11 modalities.

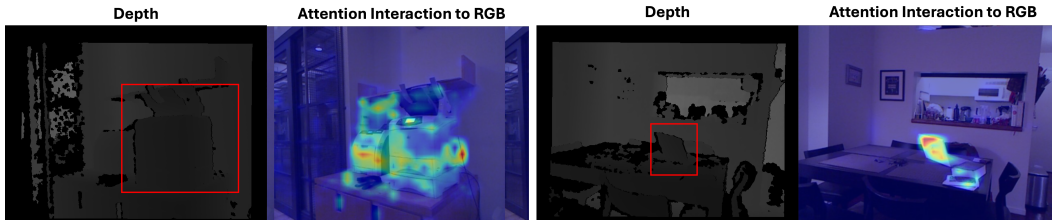


Figure 1: The Attention Visualization of ICA Layer.

12 **B Dicussion of Limitations**

13 Our experiments have chosen a set of learning hyper-parameters and benchmarks to support our
14 hypothesis. However, our work could contain several limitations. Our work studied the effectiveness
15 of our proposed invertible cross-attention layers in multimodal learning. Thus, the investigation of
16 balance weights among learning objectives has not been fully exploited, and we leave this experiment
17 as our future work. Due to computation limitations, our experiments are limited to the standard
18 scale of the benchmarks. However, we hypothesize that the proposed approaches can generalize
19 to larger-scale data and benchmark settings according to the fundamental theories presented in our
20 paper.

21 **References**

- 22 [1] D. Jia, J. Guo, K. Han, H. Wu, C. Zhang, C. Xu, and X. Chen. Geminifusion: Efficient pixel-wise
23 multimodal fusion for vision transformer. *arXiv preprint arXiv:2406.01210*, 2024.
- 24 [2] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang. Multimodal token fusion for vision
25 transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern*
26 *recognition*, pages 12186–12195, 2022.