AsymFormer: Asymmetrical Cross-Modal Representation Learning for Mobile Platform Real-Time RGB-D Semantic Segmentation

Siqi Du[1,†], Weixi Wang[1,†], Renzhong Guo[1], Ruisheng Wang[1,2], Shengjun Tang[1,*]

[1]Shenzhen University, China
[2]University of Calgary, Canada

dusiqi2021@email.szu.edu.cn

{wangwx, guorz, shengjuntang}@szu.edu.cn

ruiswang@ucalgary.ca

Abstract

Understanding indoor scenes is crucial for urban studies. Considering the dynamic nature of indoor environments, effective semantic segmentation requires both real-time operation and high accuracy.To address this, we propose AsymFormer, a novel network that improves real-time semantic segmentation accuracy using RGB-D multi-modal information without substantially increasing network complexity. AsymFormer uses an asymmetrical backbone for multimodal feature extraction, reducing redundant parameters by optimizing computational resource distribution. To fuse asymmetric multimodal features, a Local Attention-Guided Feature Selection (LAFS) module is used to selectively fuse features from different modalities by leveraging their dependencies. Subsequently, a Cross-Modal Attention-Guided Feature Correlation Embedding (CMA) module is introduced to further extract cross-modal representations. The AsymFormer demonstrates competitive results with 54.1% mIoU on NYUv2 and 49.1% mIoU on SUNRGBD. Notably, AsymFormer achieves an inference speed of 65 FPS (79 FPS after implementing mixed precision quantization) on RTX3090, demonstrating that AsymFormer can strike a balance between high accuracy and efficiency. Code: https://github.com/Fourier7754/AsymFormer

[1] [2]

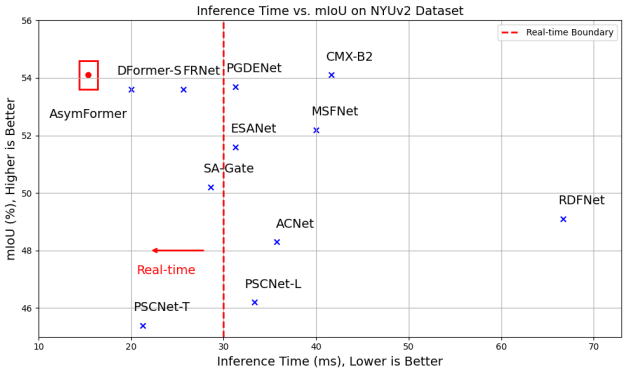

Figure 1. The AsymFormer has 33.0 million parameters and 36.0 GFLOPs computational cost, and it can achieve 65 FPS inference speed on RTX 3090, 54.1% mIoU on NYUv2.

1. Introduction

Indoor scenes are essential to urban environments. Current urban studies necessitate understanding indoor scene semantic information for tasks like emergency evacuation [26], robotic navigation [24], and virtual reality [29]. The dynamic nature of indoor scenes demands perception of environmental changes in real time for tasks such as emergency evacuation [26], stressing the importance of algorithms' real-time capabilities. Existing real-time semantic segmentation methods often falter when applied to the complex semantic information of indoor scenes, usually requiring a sacrifice in accuracy [27]. This often requires a trade-off between inference speed and increased network complexity to achieve adequate segmentation accuracy indoors. Consequently, a critical research question is how to improve the accuracy of semantic segmentation in indoor environments without substantially increasing complexity, while ensuring real-time performance.

Aside from increasing the complexity of the network, introducing additional information, such as RGB-D data, is also an effective way to improve the accuracy of semantic segmentation networks. RGB-D cameras are widely used devices for indoor information acquisition. RGB-D information consists of RGB (color, texture and shape) and

[1] † indicates equal contribution.
[2] * corresponding author.

Depth (boundaries and relative location) features, which are somewhat complementary [28]. Several studies have explored how to improve indoor scene semantic segmentation performance by integrating RGB-D information [15, 28, 33].

Existing research has explored the implementation of attention mechanisms to extract valuable information from RGB-D features without significantly increasing computational complexity. However, due to the additional feature extraction branch for depth features and the lack of discussion on how to allocate computational resources based on feature importance, these methods often introduce a substantial amount of redundant parameters, significantly reducing their computational efficiency [4].

To address this issue, this paper introduces AsymFormer, a high-performance real-time network for RGB-D semantic segmentation that employs an asymmetric backbone design. This includes a larger parameter backbone for important RGB features and a smaller backbone for the Depth branch. Regarding framework selection, at the same computational complexity, Transformer often achieves higher accuracy but has a slower inference speed compared to CNN [10]. In order to speed up the main branch, this paper employ a hardware-friendly CNN [12] for the RGB branch and a light-weight Transformer [23] for the Depth branch to further compress the parameters. Considering the differences between different modal representation, to effectively select and fuse asymmetric features, this paper proposes a learnable method for feature information compression and constructs a Local Attention Guided Feature Selection (LAFS) module. Additionally, a Cross-Modal Attention (CMA) module is introduced to embed cross-modal information into pixel-wise fused features. Finally, we employ a lightweight MLP-Decoder[23] to decode semantic information from shallow features.

This paper evaluates AsymFormer on two classic indoor scene semantic segmentation datasets: NYUv2 and SUN-RGBD. Meanwhile, the inference speed test is also preformed on Nvidia RTX 3090 platform. The AsymFormer achieves 54.1% mIoU on NYUv2 and 49.1 mIoU on SUN-RGBD, with 65 FPS inference speed (79 FPS with mixed precision quantization using the TensorRT). Our experiments highlight AsymFormer's ability to acquire high accuracy and efficiency at the same time. The main contributions are summarized as follows:

- We employed an asymmetric backbone that compressed the parameters of the Depth feature extraction branch, thus reducing redundancy.
- We introduced the LAFS module for feature selection, utilizing learnable feature weights to calculate spatial attention weights.
- We introduce a novel efficient cross-modal attention (CMA) for modeling of self-similarity in multi-modal features, validating its capability to enhance network accuracy with minimal additional model parameters.

2. Related Works

2.1. Indoor Scene Understanding

Current research in indoor scene understanding leverages diverse data sources, including RGB images [2], RGB-D images [28], point clouds [34], and mesh data [30]. Each of these sources provides unique insights and benefits for analyzing and interpreting indoor environments. For instance, RGB images are accessible and straightforward for visual representations, whereas point clouds and meshes offer intricate 3D spatial data. However, when considering the simultaneous use of both 2D and 3D information to optimize efficiency and effectiveness in scene understanding, RGB-D images emerge as the optimal choice.

2.2. RGB-D Representation Learning

One of the earliest works on RGB-D semantic segmentation, FCN [13], treated RGB-D information as a single input and processed it with a single backbone. However, subsequent works have recognized the need to extract features from RGB and Depth information separately, as they have different properties. Therefore, most of them have adopted two symmetric backbones for RGB and Depth feature extraction [3, 15, 16, 28]. Primarily, asymmetric backbones doubles the overall computational complexity [4]. However, it is generally observed that for semantic segmentation, RGB information typically plays a more prominent role than Depth information, as indicated by [4]. Using a asymmetric backbone for feature extraction will obviously lead to redundant parameters on the less important side, reducing the efficiency of the network.

2.3. RGB-D feature fusion

The performance and efficiency of different frameworks depends largely on how they fuse RGB and Depth features. Some early works, such as RedNet [8], fused RGB and Depth feature maps pixel-wise in the backbone. Later, ESANet series [15, 16] proposed channel attention to select features from different channels, as RGB and Depth feature maps may not align well on the corresponding channels. PSCNet [4] further extended channel attention to both spatial and channel directions and achieved better performance. Recently, more complex models have been proposed to exploit cross-modal information and select features for RGB-D fusion. For example, SAGate [3] proposed a gated attention mechanism that can leverage cross-modal information for feature selection. CANet [31] extended non-local attention [20] to cross-modal semantic information and achieved significant improvement. CMX [28] extended SA-Gate to spatial and channel directions and proposed a novel cross-

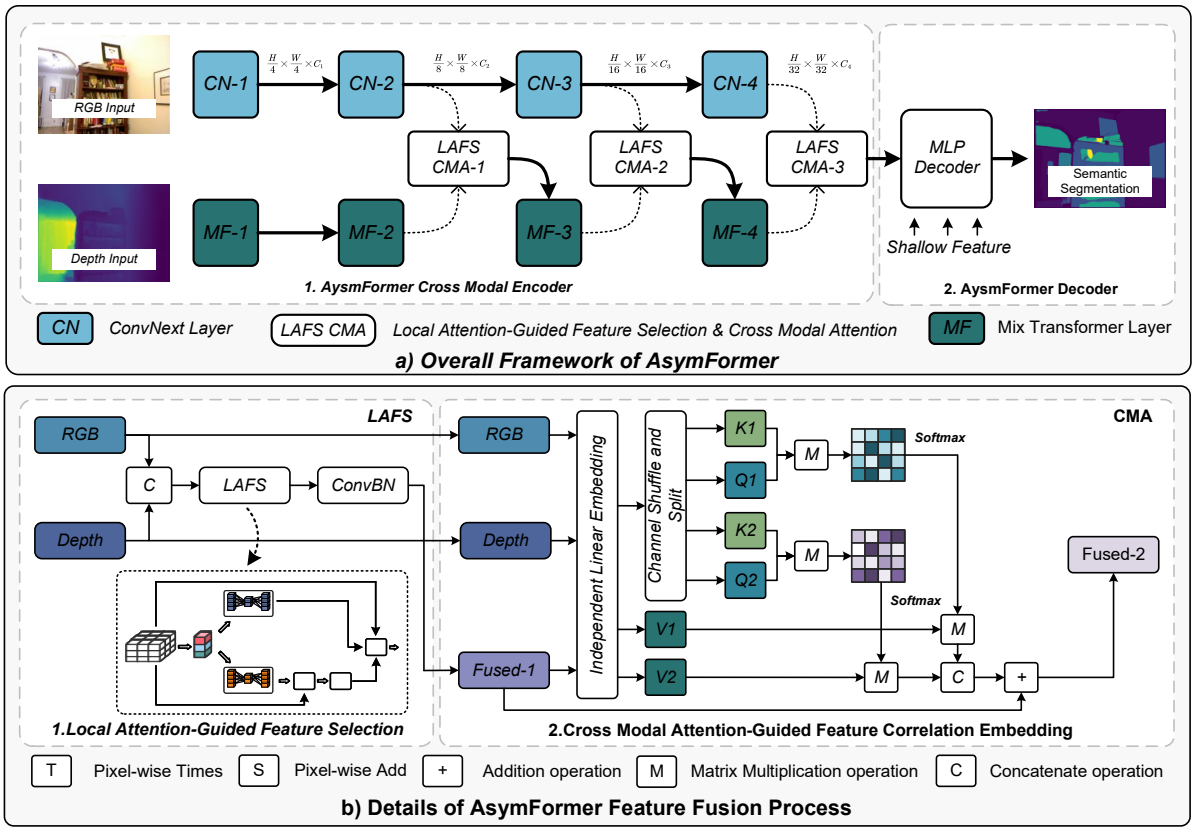

Figure 2. Overview of AsymFormer.

modal attention with global receptive field. However, integrating cross-modal information and learning cross-modal similarity is still an open question in vision tasks.

3. Method

3.1. Framework Overview

This paper introduces AsymFormer, a high-precision multimodal real-time semantic segmentation method. AsymFormer employs a dual-stream asymmetric backbone to reduce redundant parameters during the feature extraction stage. A hardware friendly convolution network ConvNext [12] is used for RGB feature extraction and a light-weight Mix-Transformer [23]is used for processing Depth feature. To effectively fuse RGB-D features, the study introduces a Local Attention Guided Feature Selection (LAFS) module, which uses learnable strategy to extract global information and selects multimodal features in both spatial and channel dimensions. Moreover, the study embeds the information contained in multimodal features through a novel Cross-Modal Attention (CMA) module. The overall framework of AsymFormer is shown in Fig.2.

3.2. Local Attention-Guided Feature Selection

Recent studies have demonstrated the effectiveness of attention mechanisms in selecting complementary features from multimodal representations [4, 11]. However, existing attention mechanisms often employ fixed, non-learnable feature information compression strategies [22], which may result in neglecting the disparities between different modal features, leading to sub-optimal information utilization.

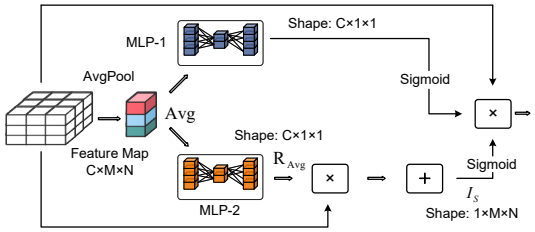

Figure 3. LAFS.

In this paper, we propose a Local Attention-Guided Feature Selection (LAFS) module. The LAFS abandons the traditional fixed strategy when extracting global information and adopts a feedforward neural network to learn a set of dynamic spatial information compression rules. Figure 3 illustrates the detailed architecture of the LAFS module. The input features of the LAFS module are the ten-

sor concatenation of RGB and $Depth$ features. LAFS first extracts the global information vector Avg through adaptive average pooling. When calculating channel attention, LAFS employs the same strategy as SE [6]. When calculating spatial attention, LAFS first processes the global information vector Avg through another feedforward neural network with a compression-expansion structure, outputting a vector R_{Avg} representing the spatial similarity description of pixels. Subsequently, LAFS obtains the global spatial information I_S by calculating the inner product of R_{Avg} and the input feature map (this operation is equivalent to calculating the weighted sum of pixels in the channel direction with dynamic weights R_{Avg}). Then, LAFS calculates the spatial attention weights W_S through sigmoid normalization.

$$W_S = \text{Sigmoid}(\frac{\text{Dot}(Input.\text{Reshape}(C, H \times W)^T, R_{Avg})}{C^2}) \tag{1}$$

Here, all results are divided by C^2 to avoid sigmoid function overflow. Finally, LAFS multiplies the calculated channel attention weights and spatial attention weights with the input features to select features in both spatial and channel directions.

3.3. Cross-Attention Guided Feature Embedding

The existing multi-head self-attention (MHSA) in Transformer [19] is limited to learning self-similarity within a single modality. When the network input is extended to multiple modalities, jointly utilizing multimodal information to mine features becomes a new goal in representation learning. To address this issue, we define a new cross-modal self-similarity measure and construct a Cross-Modal Attention (CMA) module. By embedding cross-modal self-similarity information into the fused features, CMA can complement the insufficiency of MHSA in utilizing multimodal information.

3.3.1 Definition of Cross-Modal Self-Similarity

Assuming that RGB and $Depth$ features are embedded into Key and $Query$, for a pixel (i_0, j_0), its cross-modal self-similarity with other pixels (i, j) can be defined as:

$$W(i,j) = \sum_{n=1}^{N}(Kr_{n,i,j} \cdot Qr_{n,i_0,j_0}) + \sum_{n=1}^{N}(Kd_{n,i,j} \cdot Qd_{n,i_0,j_0}) \tag{2}$$

where $Kr_{n,i,j}$ represents the n-th Key_{RGB} feature of the pixel (i, j), and $Qr_{n,i,j}$ represents the n-th $Query_{RGB}$ feature of the pixel (i, j). Similarly, $Kd_{1,i,j}$ and $Qd_{1,i,j}$ represent the n-th Key_{Depth} and $Query_{Depth}$ feature of the pixels (i, j).

3.3.2 Feature Embedding

CMA has three input features: RGB features, $Depth$ features, and the fused features $Fused$ selected by LAFS. In the calculation process of CMA, the first step is to embed the input features RGB, $Depth$, and $Fused$ into the vector space, generating Key and $Query$ corresponding to different modal features, as well as $Value$ representing the fused features. We employ several independent linear layers to embed different features into the vector space. In addition, to learn the features of two subspaces, the embedding vector Value of Fused is divided into two independent vectors V_1 and V_2 in the channel direction.

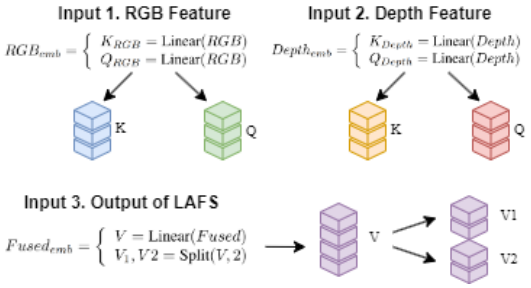

Figure 4. Feature Embedding.

3.3.3 Splitting and Mixing of Multimodal Information

After performing feature embedding calculations, the embedded feature vectors are K_{RGB}, K_{Depth}, Q_{RGB}, and Q_{Depth}. Subsequently, CMA concatenates Key and $Query$ in the channel direction to calculate cross-modal self-similarity:

$$Key, Query = \begin{cases} Key = \text{Cat}[K_{RGB}, K_{Depth}] \\ Query = \text{Cat}[Q_{RGB}, Q_{Depth}] \end{cases} \tag{3}$$

To calculate the self-similarity of features in two different subspaces, Key and $Query$ need to be split, but equally dividing Key and $Query$ cannot simultaneously include RGB and $Depth$ features. Therefore, we introduce a Shuffle mechanism to ensure that the split results of Key and $Query$, $K1$, $K2$, $Q1$, and $Q2$, all contain information from both modalities. As shown in Figure 5.

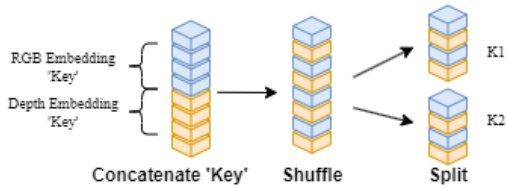

Figure 5. Splitting and Mixing of Multimodal Information.

3.3.4 Representation Learning in Multiple Subspaces

Finally, CMA calculates the cross-modal similarity based on the computed K, Q, V and embeds the calculation results into the fused features $Fused$. Firstly, CMA calculates the dot product of K_1, Q_1 and K_2, Q_2 to obtain the representations W_1 and W_2 of the two subspaces:

$$W_1 = \text{Softmax}(\frac{Q_1 \cdot K_1^T}{\sqrt{C_1/4}})\ W_2 = \text{Softmax}(\frac{Q_2 \cdot K_2^T}{\sqrt{C_1/4}}) \tag{4}$$

Subsequently, W_1 and W_2 embed information into V_1 and V_2 through dot product. The fused features $Fused_2$ are the concatenation of the above calculation results in the channel dimension:

$$Fused_2 = \text{Cat}[W_1 \cdot V_1, W_2 \cdot V_2] \tag{5}$$

Finally, CMA converts $Fused_2$ to the same number of channels as $Fused$ through a linear layer and adds it to the residual connection $Fused$ pixel by pixel to obtain the output features of CMA.

4. EXPERIMENT RESULTS

4.1. Implementation Details

To evaluate our Real-Time semantic segmentation network design, we conduct a series of experiments on two widely-used datasets NYUv2[17] (795 training and 654 testing RGB-D images) and SUNRGBD[18] (5825 training and 5050 testing RGB-D images). We conduct the model training and testing on different platforms. For the training, we use Nvidia A100-40G GPU. For the evaluation and inference speed testing, we use Nvidia RTX 3090 GPU, Ubuntu 20.04, CUDA 12.0 and Pytorch 2.0.1. We apply data augmentation to all datasets by randomly flipping (p=0.5), random scales between 1.0 and 2.0, random crop 480×640 and random HSV. The MLP-decoder in AsymFormer has the same structure as Segformer and an embedding dimension of 256. We choose AdamW optimizer with a weight decay of 0.01. The initial learning rate is $5e^{-5}$ and we use a poly learning rate schedule $(1 - \frac{iter}{max_iter})^{0.9}$ with a warm-up of 10 epochs. We train with a batch size of 8 for NYUv2 (500 epochs) and SUNRGBD (200 epochs). We employ cross-entropy as the loss function and do not use any auxiliary loss during training process. The evaluation metric is mean Intersection over Union (mIoU).

4.2. Ablation Experiment

We conduct a series of ablation experiments on NYUv2 dataset to evaluate the effectiveness of the LAFS and CMA module. We set two common feature fusion methods as our comparative baseline: **1. Cat:** This method directly concatenates two features and then uses convolution layers to adjust the channel numbers. Essentially, it is a pixel-wise fusion without feature selection. **2. SE+MHSA:** This method combines the popular SE attention [6] and MHSA attention [19] for feature fusion. Here, SE is used for feature selection in the channel direction, while MHSA is employed for further feature extraction on the fused features.

In our experiments, the Cat fusion method, used as a baseline, achieved a segmentation accuracy of 47.0 mIoU and an inference speed of 77.5 FPS. When using LAFS alone, we achieved a performance improvement of 2.1% while sacrificing only 1.8 FPS of inference speed. This demonstrates that LAFS provides performance gains without significantly impacting inference speed. In comparison to the other baseline, using Cat+MHSA, which resulted in a reduction of inference speed by 11.8 FPS, an improvement of only 2.9% in mIoU was achieved. This further highlights the efficiency of LAFS. Furthermore, when using CMA alone, we observed a 2.6% improvement in segmentation accuracy but encountered a significant decrease in inference speed of 10.1 FPS. Compared to LAFS, CMA showed a more noticeable reduction in inference speed.

Finally, we combined LAFS with CMA (LAFS+CMA). Since LAFS had minimal impact on inference speed and served a different purpose than CMA, the network's inference speed decreased by only 2 FPS. This change achieved a significant improvement of 7.1% in segmentation accuracy compared to the baseline Cat. At this point, the inference speed of LAFS+CMA was similar to SE+MHSA, but with a 4.2% performance improvement. This validates our experimental hypothesis: by re-modeling feature selection and mining cross-modal self-similarity, we can enhance the segmentation performance of the network without sacrificing inference speed compared to existing models. This demonstrates that we have indeed improved the efficiency of the network.

4.3. Comparison With State-of-The-Arts

4.3.1 NYUv2 Comparison Results

According to Table 2, despite the lack of ImageNet-1k pretraining—a common practice among competing methods—our AsymFormer still achieves leading scores in Real-Time semantic segmentation. Th AsymFormer achieves 54.1 % mIoU, demonstrating competitive accuracy compared to those high-performance heavy designs. AsymFormer also has faster inference speed than other methods. For instance, AsymFormer outperforms PSCNet-T[4] by 8.7% mIoU and 18 FPS inference speed improvement. Similarly, AsymFormer is two times faster than ESANet[15] and three times faster than CMX-B2 [28] with the same performance. Finally, by using multi-scale inference strategy, the AsymFormer achieves 55.3 % mIoU on NYUv2. In terms of semantic segmentation accuracy, AsymFormer does not show a significant disadvantage compared to those

Table 1. Ablation experiment results for different multi-modal feature fusion method.

Model	Feature Fusion Method				Metric		
	Cat	SE+MHSA	LAFS	CMA	Params/M	mIoU(%)	Inf.Speed(FPS)
Baseline	✓				31.9	47.0	77.5
		✓			32.6(+0.7M)	49.9 (+2.9%)	65.7
Ours			✓		32.4 (+0.5M)	49.1 (+2.1%)	75.7
				✓	32.5 (+0.6M)	49.6 (+2.6%)	67.4
			✓	✓	33.0 (+1.1M)	54.1 (+7.1%)	65.5

Table 2. Comparison Results on NYUv2. The inference speed is tested on RTX 3090 platform, (480 × 640) inputs. MS denotes Multi-Scale inference strategy.

Method	Year	Backbone	Params/M	mIoU (%)	Real-Time	Speed/FPS	Speed (%)
CMX-B2 [28]	2022	Segformer-B2	67	54.1	×	24	36.9%
Token-Fusion [21]	2022	Token-Fusion(S)	-	54.2	×	-	-
CMX-B2 (MS) [28]	2022	Segformer-B2	67	54.4	×	-	-
Multi-MAE [1]	2022	Vit-B	-	56.0	×	-	-
CMX-B4 (MS) [28]	2022	Segformer-B4	140	56.3	×	-	-
Omnivore [5]	2022	Swin-L	-	56.8	×	-	-
CMX-B5 (MS) [28]	2022	Segformer-B5	181	56.9	×	-	-
SA-Gate [3]	2021	Res50	65	50.2	✓	35	53.8%
ESANet [15]	2022	Res34-Nbt1D	34	51.6	✓	32	49.2%
PSCNet-L [4]	2022	Res50	52	46.2	✓	30	46.2%
PSCNet-T [4]	2022	Res50	40	45.4	✓	47	72.3%
PGDENet [32]	2022	Res34	101	53.7	✓	32	49.2%
FRNet [33]	2022	Res34	86	53.6	✓	39	60.0%
DFormer-S [25]	2023	DFormer-S	18.7	53.6	✓	50	76.9%
AsymFormer	2024	B0+T	33	54.1	✓	65	100.0%
AsymFormer (FP16)	2024	B0+T	33	54.1	✓	79	121.5%
AsymFormer (MS)	2024	B0+T	33	55.3	×	-	-

high-performance methods, such as Omnivore, included in the comparison. This validates the effectiveness of our various efforts in reducing network redundancy parameters.

4.3.2 SUNRGBD Comparison Results

Table 3 reports the performance of AsymFormer on the SUNRGBD dataset. AsymFormer achieves competitive accuracy with 49.1 % mIoU. The advantage of AsymFormer is not as significant as in NYUv2 experiment. For example, AsymFormer improves 3.9 mIoU over SA-Gate[3] in NYUv2 dataset (54.1% vs 50.2% mIoU), but decreases 0.3 mIoU in SUNRGBD dataset (49.1% vs 49.4% mIoU). A similar performance degradation can be observed in CMX-B2 result which also uses Transformer based backbone. We conjecture that this phenomenon may be caused by low quality depth images in SUNRGBD dataset. The aim of our research is not to construct a state-of-the-art method that has a marginal mIoU improvement over other methods, but to construct a method that has a better performance-speed balance and is more suitable for robot platform. Given that AsymFormer still has faster inference speed than other methods, we consider this performance acceptable for AsymFormer.

Table 3. Comparison Results on SUNRGBD. MS denotes Multi-Scale inference strategy.

Method	Pixel Acc. (%)	mIoU (%)
RDFNet [14]	81.5	47.7
ESANet [15]	-	48.0
ACNet [7]	-	48.1
SA-Gate [3]	82.5	49.4
CMX-B2 (MS) [28]	82.8	49.7
DFormer-S [25]	-	50.0
MSFNet [9]	-	50.3
FRNet [33]	87.4	51.8
PGDENet [32]	87.7	51.0
CMX-B4 (MS) [28]	83.5	52.1
CMX-B5 (MS) [28]	83.8	52.4
AsymFormer	81.9	49.1

4.4. Visualization

4.4.1 LAFS Attention Map

As shown in Figure 6, to demonstrate that LAFS performs better than CBAM [22] in selecting features in the spatial dimension, we visualized the spatial attention weights of both methods. It can be observed that LAFS provides better coverage of informative regions in the image while maintaining consistency within objects and preserving the integrity of edges.

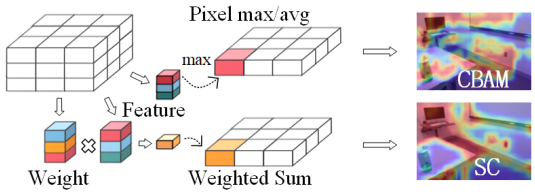

Figure 6. Difference between CBAM and LAFS.

4.4.2 Semantic Segmentation Results

The Figure 7 demonstrates the segmentation results of AsymFormer on the NYUv2 dataset. As observed, while maintaining a significantly faster inference speed compared to other methods, AsymFormer achieves comparable semantic segmentation accuracy to mainstream approaches.

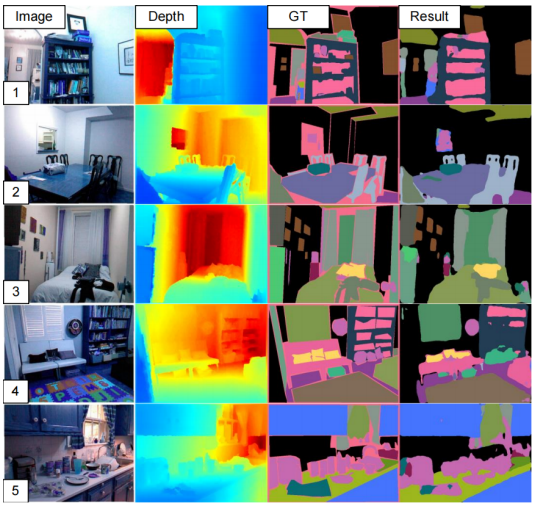

Figure 7. Visualization of AsymFormer Semantic Segmentation Results.

5. CONCLUSIONS

In this work, we proposed AsymFormer, which aims to construct a less-redundant real-time indoor scene understanding network. To enhance efficiency and reduce redundant parameters, we implemented the following improvement: 1. the asymmetric backbone that compressed the parameters of the Depth feature extraction branch, thus reducing redundancy. 2. the LAFS module for feature selection, utilizing learnable strategy for global information compressing and improving spatial attention calculations 3. the self-similarity in multi-modal features, validating its capability to enhance network accuracy with minimal additional model parameters. The experiments demonstrated that the AsymFormer achieves a balance between accuracy and speed. Moving forward, we will continue to optimize the modules and address issues such as self-supervised pre-training of the model, aiming for further improvements.

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. 6

[2] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 2

[3] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*, pages 561–577. Springer, 2020. 2, 6

[4] SQ Du, SJ Tang, WX Wang, XM Li, YH Lu, and RZ Guo. Pscnet: Efficient rgb-d semantic segmentation parallel network based on spatial and channel attention. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, (1), 2022. 2, 3, 5, 6

[5] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens Van Der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112, 2022. 6

[6] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4, 5

[7] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1440–1444. IEEE, 2019. 6

[8] Jindong Jiang, Lunan Zheng, Fei Luo, and Zhijun Zhang. Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation. *arXiv preprint arXiv:1806.01054*, 2018. 2

[9] Shiyi Jiang, Yang Xu, Danyang Li, and Runze Fan. Multi-scale fusion for rgb-d indoor semantic segmentation. *Scientific Reports*, 12(1):20305, 2022. 6

[10] Jiashi Li, Xin Xia, Wei Li, Huixia Li, Xing Wang, Xuefeng Xiao, Rui Wang, Min Zheng, and Xin Pan. Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios. *arXiv preprint arXiv:2207.05501*, 2022. 2

[11] 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. 3

[12] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 2, 3

[13] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[14] Seong-Jin Park, Ki-Sang Hong, and Seungyong Lee. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 4980–4989, 2017. 6

[15] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. Efficient rgb-d semantic segmentation for indoor scene analysis. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13525–13531. IEEE, 2021. 2, 5, 6

[16] Daniel Seichter, Söhnke Benedikt Fischedick, Mona Köhler, and Horst-Michael Groß. Efficient multi-task rgb-d scene analysis for indoor environments. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE, 2022. 2

[17] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. *ECCV (5)*, 7576:746–760, 2012. 5

[18] 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. 5

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4, 5

[20] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 2

[21] 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. 6

[22] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 3, 7

[23] 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, 3

[24] Yao Yeboah, Cai Yanguang, Wei Wu, and Zeyad Farisi. Semantic scene segmentation for indoor robot navigation via deep learning. In *Proceedings of the 3rd International Conference on Robotics, Control and Automation*, pages 112–118, 2018. 1

[25] Bowen Yin, Xuying Zhang, Zhongyu Li, Li Liu, Ming-Ming Cheng, and Qibin Hou. Dformer: Rethinking rgbd representation learning for semantic segmentation. *arXiv preprint arXiv:2309.09668*, 2023. 6

[26] Sang-Jo Yoo and Seung-Hee Choi. Indoor ar navigation and emergency evacuation system based on machine learning and iot technologies. *IEEE Internet of Things Journal*, 9 (21):20853–20868, 2022. 1

[27] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, 129: 3051–3068, 2021. 1

[28] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 2, 5, 6

[29] Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, Joon-Young Lee, Hailin Jin, and Thomas Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5287–5295, 2017. 1

[30] Bo Zheng, Yibiao Zhao, Joey C. Yu, Katsushi Ikeuchi, and Song-Chun Zhu. Beyond point clouds: Scene understanding by reasoning geometry and physics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3127–3134, 2013. 2

[31] Hao Zhou, Lu Qi, Hai Huang, Xu Yang, Zhaoliang Wan, and Xianglong Wen. Canet: Co-attention network for rgb-d semantic segmentation. *Pattern Recognition*, 124:108468, 2022. 2

[32] Wujie Zhou, Enquan Yang, Jingsheng Lei, Jian Wan, and Lu Yu. Pgdenet: Progressive guided fusion and depth enhancement network for rgb-d indoor scene parsing. *IEEE Transactions on Multimedia*, 2022. 6

[33] Wujie Zhou, Enquan Yang, Jingsheng Lei, and Lu Yu. Frnet: Feature reconstruction network for rgb-d indoor scene parsing. *IEEE Journal of Selected Topics in Signal Processing*, 16(4):677–687, 2022. 2, 6

[34] Wei Zhou, Yao Yue, Meng Fang, Xujun Qian, Rong Yang, and Li Yu. Bcinet: Bilateral cross-modal interaction network for indoor scene understanding in rgb-d images. *Information Fusion*, 78:84–94, 2023. 2