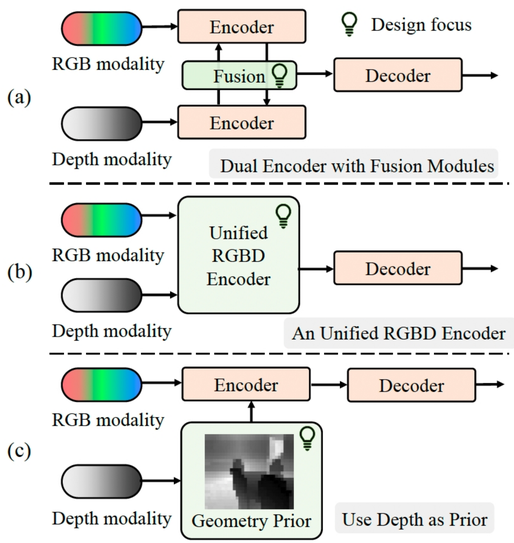
DFormerv2: Geometry Self-Attention for RGBD Semantic Segmentation

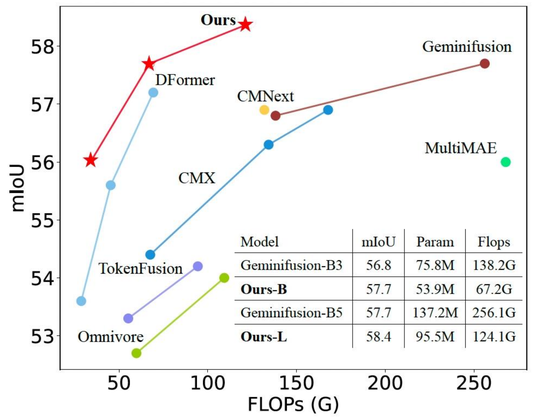
Bo-Wen Yin, Jiao-Long Cao, Ming-Ming Cheng, Qibin Hou#

#Q. Hou is the corresponding author.

Nankai University

Problem & Solution



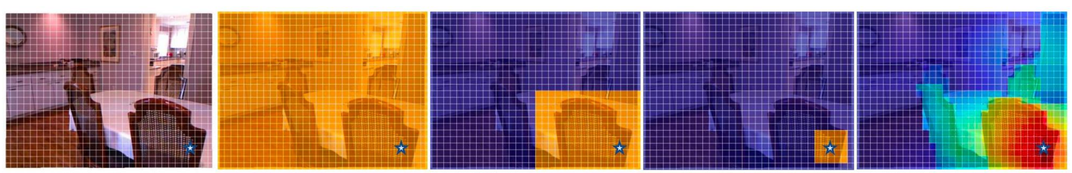


Model	mIoU	Param	Flops
Geminifusion-B3	56.8	75.8M	138.2G
Ours-B	57.7	53.9M	67.2G
Geminifusion-B5	57.7	137.2M	256.1G
Ours-L	58.4	95.5M	124.1G

Compared to current SOTA methods, our DFormerv2 achieves better performance with less than less Parms&Flops.

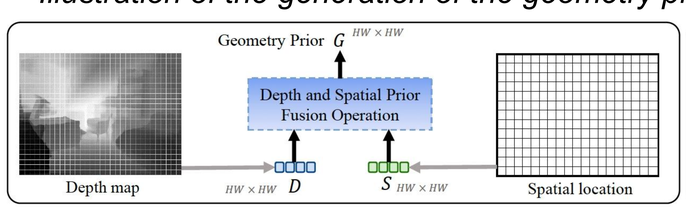
Is it possible to specifically design a depth encoding manner?

Comparison between geometry self-attention (GSA) and other attention mechanisms

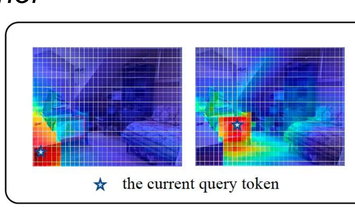


Input Image | Vanilla Attention | Window Attention | Local Attention | Geometry Attention

Illustration of the generation of the geometry prior

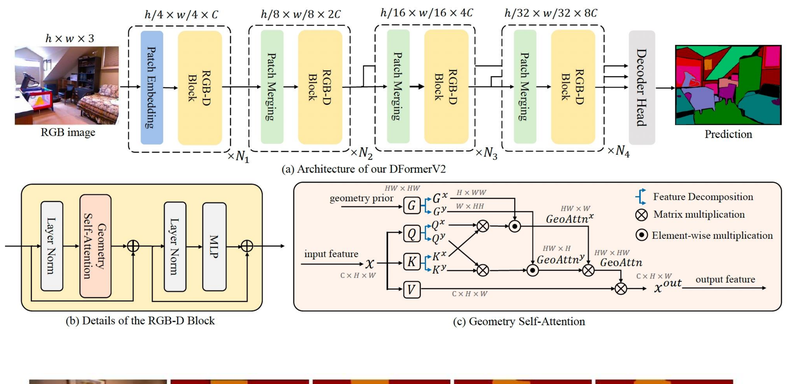


(a) Generation of geometry prior

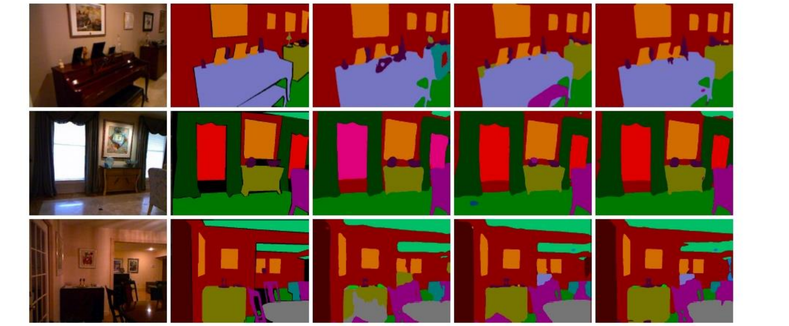


(b) Visualization of geometry prior

Model Structure



(a) Architecture of our DFormerV2



(b) Details of the RGB-D Block (c) Geometry Self-Attention

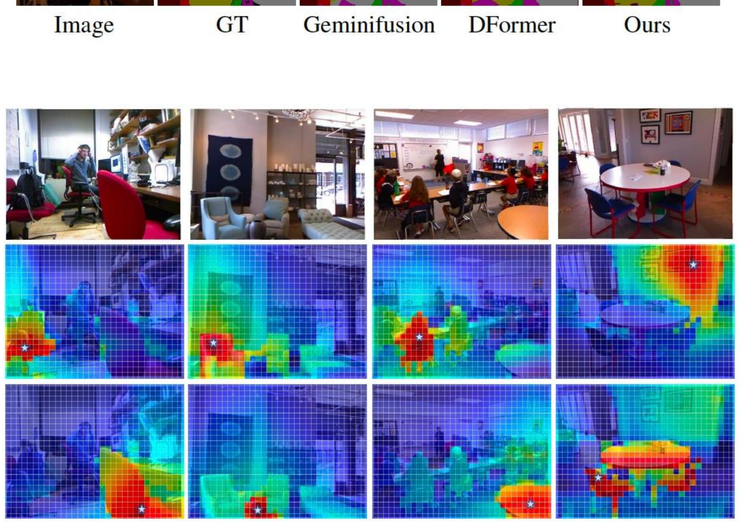


Image | GT | Geminifusion | DFormer | Ours

Feature Distribution

Visualization of the focused geometry attention

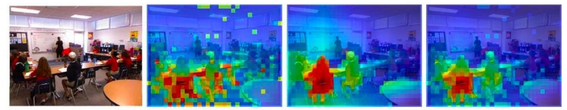


Image | Attention map | Geometry prior | Focused attention

Feature comparison with and without GSA

Detailed Results

Model	Backbone	Params	NYUDepthv2 Input size	NYUDepthv2 Flops	NYUDepthv2 mIoU	SUN-RGBD Input size	SUN-RGBD Flops	SUN-RGBD mIoU
TokenFusion$_{22}$ [54]	MiT-B2	26.0M	480 × 640	55.2G	53.3	530 × 730	71.1G	50.3
Omnivore$_{22}$ [16]	Swin-Tiny	29.1M	480 × 640	32.7G	49.7	530 × 730	—	—
DFormer$_{24}$ [59]	DFormer-Tiny	6.0M	480 × 640	11.7G	51.8	530 × 730	15.0G	48.8
DFormer$_{24}$ [59]	DFormer-Small	18.7M	480 × 640	25.6G	53.6	530 × 730	33.0G	50.0
DFormer$_{24}$ [59]	DFormer-Base	29.5M	480 × 640	41.9G	55.6	530 × 730	54.0G	51.2
AsymFormer$_{24}$ [11]	MiT-B0+ConvNeXt-Tiny	33.0M	480 × 640	39.4G	55.3	530 × 730	52.6G	49.1
★ DFormerv2-S	DFormerv2-Small	26.7M	480 × 640	33.9G	**56.0**	530 × 730	43.7G	**51.5**
SGNet$_{20}$ [4]	ResNet-101	64.7M	480 × 640	108.5G	51.1	530 × 730	151.5G	48.6
ShapeConv$_{21}$ [3]	ResNext-101	86.8M	480 × 640	124.6G	51.3	530 × 730	161.8G	48.6
FRNet$_{20}$ [68]	ResNet-34	85.5M	480 × 640	115.6G	53.6	530 × 730	150.0G	51.8
EMSANet$_{24}$ [40]	ResNet-34	46.9M	480 × 640	45.4G	51.0	530 × 730	58.6G	48.4
TokenFusion$_{22}$ [54]	MiT-B3	45.9M	480 × 640	94.4G	54.2	530 × 730	122.1G	51.4
Omnivore$_{22}$ [16]	Swin-Small	51.3M	480 × 640	59.8G	52.7	530 × 730	—	—
CMX$_{22}$ [61]	MiT-B2	66.6M	480 × 640	67.6G	54.4	530 × 730	86.3G	49.7
DFormer$_{24}$ [59]	DFormer-Large	39.0M	480 × 640	65.7G	57.2	530 × 730	84.5G	52.5
GeminiFusion$_{24}$ [28]	MiT-B3	75.8M	480 × 640	138.2G	56.8	530 × 730	179.0G	52.7
★ DFormerv2-B	DFormerv2-Base	53.9M	480 × 640	67.2G	**57.7**	530 × 730	86.9G	**52.8**
SA-Gate$_{20}$ [5]	ResNet-101	110.9M	480 × 640	193.7G	52.4	530 × 730	250.1G	49.4
CEN$_{20}$ [53]	ResNet-101	118.2M	480 × 640	618.7G	51.7	530 × 730	790.3G	50.2
CEN$_{20}$ [53]	ResNet-152	133.9M	480 × 640	664.4G	52.5	530 × 730	849.7G	51.1
PGDENet$_{22}$ [67]	ResNet-34	100.7M	480 × 640	178.8G	53.7	530 × 730	229.1G	51.0
MultiMAE$_{22}$ [1]	ViT-Base	95.2M	640 × 640	267.9G	56.0	640 × 640	267.9G	51.1†
Omnivore$_{22}$ [16]	Swin-Base	95.7M	480 × 640	109.3G	54.0	530 × 730	—	—
CMX$_{22}$ [61]	MiT-B4	139.9M	480 × 640	134.3G	56.3	530 × 730	173.8G	52.1
CMX$_{22}$ [61]	MiT-B5	181.1M	480 × 640	167.8G	56.9	530 × 730	217.6G	52.4
CMNext$_{23}$ [62]	MiT-B4	119.6M	480 × 640	131.9G	56.9	530 × 730	170.3G	51.9†
GeminiFusion$_{24}$ [28]	MiT-B5	137.2M	480 × 640	256.1G	57.7	530 × 730	332.4G	**53.3**
★ DFormerv2-L	DFormerv2-Large	95.5M	480 × 640	124.1G	**58.4**	530 × 730	160.5G	**53.3**

bowenyin@mail.nankai.edu.cn

Feel free to contact us!

欢迎与我们随时交流！

GitHub | WeChat