DFormerv2: 基于几何自注意力机制的RGBD语义分割

尹博文，曹骄龙，程明明，侯淇彬*
VCIP, CS, 南开大学
bowenyin@mail.nankai.edu.cn, houqb@nankai.edu.cn

Abstract

场景理解的最新进展得益于深度图，因为它包含*3D*几何信息，尤其是在复杂条件下（例如弱光和过度曝光）。现有方法将深度图与*RGB* 图像一起编码，并在它们之间进行特征融合，以实现更稳健的预测。考虑到深度可以被视为*RGB* 图像的几何补充，一个简单的问题出现了：我们真的需要像对*RGB* 图像那样，用神经网络明确地编码深度信息吗？基于这一见解，在本文中，我们研究了一种学习*RGBD* 特征表示的新方法，并提出了一个强大的*RGBD* 编码器*DFormerv2*，它明确地使用深度图作为几何先验，而不是用神经网络编码深度信息。我们的目标是从所有图像块标记之间的深度和空间距离中提取几何线索，然后将其用作几何先验，以在自注意力机制中分配注意力权重。大量实验表明，*DFormerv2* 在各种*RGBD* 语义分割基准测试中表现出色。代码已开源：*https://github.com/VCIP-RGBD/DFormer*。

1. 引言

语义分割旨在将图像中的每个像素分配到特定的预定义类别标签，由于其广泛的应用（例如智能交通系统和自动驾驶）而成为计算机视觉领域的一个重要研究领域。 [30]。然而，仅基于RGB 数据的方法在复杂场景下（例如杂乱的室内环境或弱光条件）通常会显著降低性能。近年来，3D 模块化传感器的进步使得深度数据更容易获取。集成RGB-D 数据可以使场景理解更加稳健和准确，因此成为推进高级视觉任务的关键。此外，RGB-D 数据已展现出显著的潜力，在包括自动驾驶 [26]、SLAM [50] 和机器人技术 [36] 在内的各种下游任务中超越了基于RGB 的范式。

Fig. 1(a) 展示了当前主流RGB-D 模型的架构。如图所示，该模型采用双编码器架构 [61, 62]，其中一个编码器从RGB 模态中提取特征，另一个编码器处理深度信息。同时，在编码过程中采用融合策略实现两种模态信息的交互。尽管取得了成功，但大多数现有的RGBD 分割方法都采用相同的主干架构，从RGB 和深度数据中提取特征进行融合，忽略了RGB 和深度之间的内在差异。

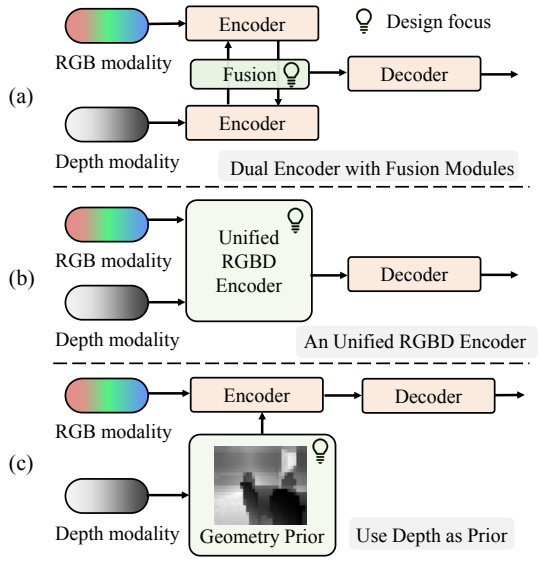


Figure 1. 主要的RGBD 分割流程与我们方法之间的比较。（a）使用双编码器分别对RGB 和深度进行编码，并设计融合模块来融合它们 [28, 61]；（b）采用统一的RGBD 编码器来提取和融合RGBD 特征 [1, 59]；（c）我们的DFormerv2 使用深度来形成场景的几何先验，然后增强视觉特征。

下列一系列研究致力于寻找处理深度图并将其与RGB 数据集成的最佳方法。Asyformer [11] 采用双流非对称主干网络，即使用更高效的深度数据编码器来减少特征提取过程中的冗余参数。PrimKD [20] 提出了一种基于知识蒸馏(KD) 的方法来指导RGB-D 语义分割中的多模态融合，重点在于充分利用RGB 的主要模态。此外，如图Fig. 1(b) 所示，DFormer [59] 提出了一种高效的RGB-D 模型，该模型通过表示学习方法 [23, 47] 在统一的编码器中对RGB 和深度数据进行编码，但分配了更多的计算资源来处理RGB 数据。这些方法承认RGB 和深度数据承载着不同的信息，它们对语义分割的贡献也不同。然而，它们未能完全解释深度模态的独特特性。总之，如何有效且高效地利用深度信息仍然是一个悬而未决的问题，值得进一步探索。

本文考虑到深度图反映给定场景几何信息的物理意

*Qibin Hou is the corresponding author.

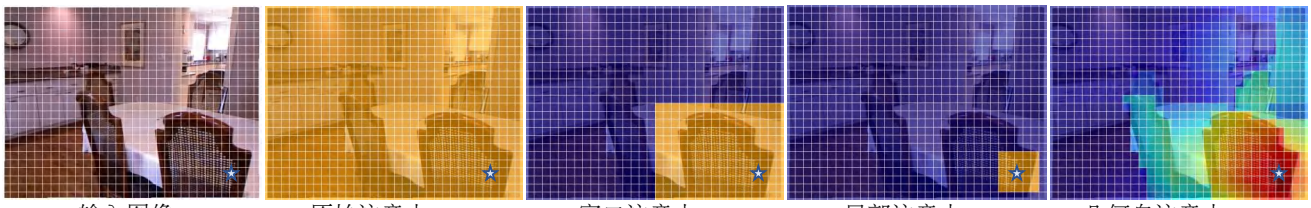


Figure 2. 几何自注意力(GSA) 与其他注意力机制（例如，原始注意力 [10]、窗口注意力 [9, 35] 和局部注意力 [52, 57]）的比较。“星号”表示当前查询的位置。在GSA 中，接近红色的颜色表示较小的衰减率，而远离红色的颜色表示较大的衰减率。在其他注意力机制中，亮色表示感受野。

义，从新的角度思考了深度图的利用方式。与之前使用神经网络同时编码RGB图像和深度图（如图Fig. 1 所示）的研究不同，我们提出直接将深度图作为几何先验，并利用它们来指导自注意力机制中的权重分布，从而产生一种新的注意力机制，称为几何自注意力机制(GSA)。关于所提出的GSA的工作原理及其与其他自注意力机制变体的区别，请参见 Fig. 2。在每个构建块中，我们基于GSA对所有patch token之间的几何和空间关系进行建模，这是一种更高效的RGB和深度信息融合方法。我们的方法不需要额外的层来处理深度图，因此与其他类型的RGBD 分割方法相比，需要的可学习参数和计算量更少。此外，为了减轻vanilla 自注意力的计算负担，我们还采用了轴分解操作，将自注意力沿特征的两个空间轴进行分解。

基于几何自注意力机制，我们构建了一个强大的RGB-D 视觉主干网络，称为DFormerv2。我们在流行的RGB-D 语义分割基准数据集上证明了DFormerv2 的有效性，例如NYU DepthV2 [42]、SUNRGBD [43] 和Deliver [62]。通过在DFormerv2 之上添加一个小型解码器头，我们的方法与以往方法相比，以更低的计算成本创造了新的SOTA 纪录。值得注意的是，我们的基础模型DFormerv2-B 在NYU DepthV2 数据集上实现了与排名第二的Gemnifuision (MiT-B5) [28] 相当的性能，即57.7% 的mIoU，而计算成本却不到其一半。同时，我们最大的模型DFormerv2-L 在NYU DepthV2 数据集上，以95.5M 的参数实现了58.4% 的mIoU。与其他方法相比，我们的DFormerv2 在分割性能和计算成本之间实现了最佳平衡。

我们的主要贡献可以概括如下：

- 据我们所知，我们的工作标志着将深度信息与空间信息相结合作为几何先验并将其应用于神经网络的首次成功尝试。
- 我们提出了几何自注意力，它在自注意力之前引入几何，以构建一个高效的RGB-D 编码器，称为DFormerv2。
- 我们的方法在三个流行的RGB-D 语义分割数据集上实现了新的最先进的性能，并且计算成本不到当前最佳方法的一半。

2. 相关工作

2.1. RGB-D语义分割

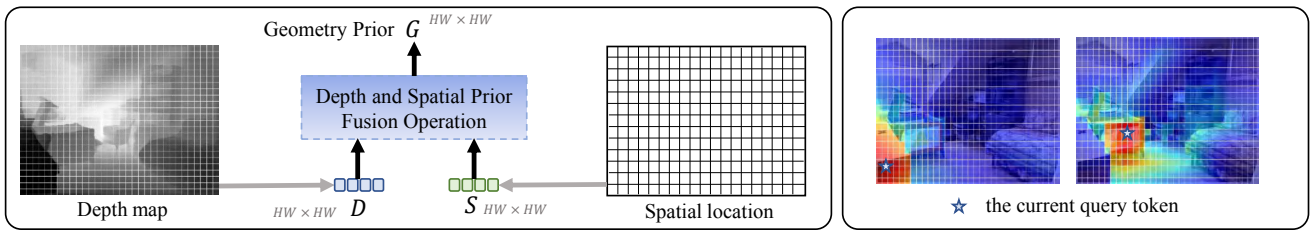
语义分割 [64]是计算机视觉领域的核心研究课题之一，旨在将图像中的每个像素归类到特定的类别中。近年来，深度学习技术 [6, 7, 18, 44, 45]在该领域取得了重大进展。然而，仅使用RGB图像理解某些现实世界场景 [13, 27, 32, 33, 56]仍然具有挑战性，因为RGB图像无法提供足够的纹理，尤其是在低照度和快速移动的场景下。为了解决这个问题，研究人员 [63, 65]提出利用包含场景三维几何信息的深度来增强RGB语义分割，即RGB-D语义分割。此后，一系列研究成果被提出，旨在实现RGB-D 数据的融合，并利用这些附加信息来捕捉更多细节。本文，我们深入研究了RGB-D 融合方案并分析了其特点。

目前主流的方法 [25, 39] 投入了大量精力来设计交互模块，以融合由两个并行预训练主干网络编码的RGB 和深度特征。例如，CMX [61]、TokenFusion [54]、GeminiFusion [28] 等方法动态地融合来自RGB 和深度编码器的RGB-D 表示，并在解码器中聚合它们。这些方法显著突破了RGB-D 语义分割应用的性能界限。然而，它们仍然面临两个共同的问题：(1) 与基于RGB 数据的方法相比，使用两个并行的主干网络平等处理RGB 和深度图会带来显著更高的计算成本；(2) 所使用的主干网络使用RGB 图像进行预训练，但在微调过程中以图像-深度对作为输入。输入之间的不一致会导致表示分布发生巨大的偏移。

最近，DFormer [59] 提出了一个RGB-D 表示学习框架，并利用一个在RGB-D 对上预训练的统一主干网络来克服这两个问题。该框架注意到了这两种模态之间信息密度的差异，并观察到深度信息仅需要一小部分通道进行编码。虽然它实现了高效的准确预测，但它忽略了深度模态的内在特性，仅仅以较低的计算成本分配深度。与此不同，在本文中，我们提出从数据特征的角度，通过深度生成几何先验。据我们所知，这是首次尝试明确使用深度的几何信息，而无需任何额外的编码层。

2.2. Vision Transformer 与先验知识

Vision Transformer (ViT) [10] 首次将Transformer 架构引入视觉任务，将图像分割成小的、不重叠的块序列。



(a) Generation of geometry prior (b) Visualization of geometry prior

Figure 3. 几何先验的图示。（a）几何先验的生成过程。（b）几何先验的一些可视化效果，其中“蓝色星号”表示当前查询。

与CNN [22, 24, 34] 最大的区别在于，Transformer [8, 9, 35, 49] 使用注意力机制替代卷积层，以实现全局上下文建模。然而，普通的自注意力机制会产生繁重的计算负担，因为它需要计算所有块之间的成对特征亲和度。为了减轻自注意力机制巨大的计算成本，各种稀疏注意力机制 [19, 35, 55, 60, 69] 已被提出。与此同时，研究人员也提出了许多研究 [17, 37, 46, 48, 51]，将先验知识融入到Transformer 模型中，以增强其表征能力。原始的Transformer [10] 利用位置编码为每个token提供位置信息。对于视觉任务，swin-transformer [35] 建议使用相对位置编码，而非原始的绝对位置编码。相比之下，我们建议将深度转化为几何先验知识，并将其引入自注意力机制，即几何自注意力机制。与位置先验相比，我们的几何先验可以对整个图像的3D 域中的关系进行建模。

3. 方法

3.1. 几何先验生成

在Vision Transformer 中，大小为$h \times w$ 的二维输入图像被均匀分割成HW 个小块，其中H 和W 分别表示每行和每列的块数。每个块标记为P_{ij}，在空间域中具有唯一的二维坐标，其中i 和j 分别索引行和列。当给定相关的深度图时，深度图中相应位置的块反映了其与相机平面的距离。基于这两种先验，我们建模所有块之间的几何关系，并将其嵌入到自注意力机制中，形成我们的几何自注意力机制。

具体而言，对于深度先验，我们对深度块中位置(i, j) 处的所有像素执行平均池化操作来表示其深度位置z_{ij}，并计算每对深度块之间的距离，其定义为：

$$D_{ij,i'j'} = |z_{ij} - z_{i'j'}|, \tag{1}$$

其中$D_{ij,i'j'}$ 表示位置(i, j) 和(i', j') 处面之间的深度距离。D形成一个形状为$HW \times HW$的深度关系矩阵。

深度关系矩阵D 不包含空间距离信息，而这对于形成几何线索也是至关重要的。因此，我们需要将深度先验与空间先验联系起来，作为几何先验，以建立图像块之间综合关系的模型。与深度先验的处理类似，我们用曼哈顿距离计算所有图像块之间的空间距离。这可以定义为：

$$S_{ij,i'j'} = |i - i'| + |j - j'|, \tag{2}$$

其中$S_{ij,i'j'}$ 表示位置(i, j) 和(i', j') 处图块之间的空间曼哈顿距离。与深度关系矩阵类似，我们也可以生成形状为$HW \times HW$ 的空间关系矩阵S。

给定深度和空间距离矩阵D 和S，我们执行融合操作来构建它们之间的桥梁，如图Fig. 3 所示。我们通过经验发现，仅仅使用两个可学习的记忆对深度和空间先验进行加权求和就已经效果很好了。值得一提的是，还可以使用更先进的技术，通过融合深度和空间先验来生成几何先验。我们整合这两种先验，生成形状为$HW \times HW$ 的几何先验G，它存储了所有图像块更全面的三维几何关系。更多G 的可视化效果见Fig. 7。

3.2. 几何自注意力机制

给定一个特征图$x \in \mathbb{R}^{HW \times C}$，在每个头中，自注意力机制可以简单地表述如下：

$$\text{SelfAtt}(Q, K, V) = \text{Softmax}(QK^T)V, \tag{3}$$

其中Q, K, V 分别是查询、键和值矩阵，可以通过线性投影获得。受[12, 41, 48] 的启发，该算法通过执行位置编码为每个token 提供空间信息，我们的几何自注意力机制可以通过以衰减的方式将几何先验G 引入自注意力机制来实现。这个过程可以写成：

$$\text{GeoAttn}(Q, K, V, G) = (\text{Softmax}(QK^T) \odot \beta^G)V, \tag{4}$$

其中$\odot$ 表示元素乘法，$\beta \in (0, 1)$ 表示衰减率，β^G 表示将G 中的每个元素取β 的幂，得到一个新的矩阵。由于$\beta \in (0, 1)$ 且G 中的元素均为非负数，因此得到的$\beta^G = [\beta^{g_{ij}}]_{ij} \in (0, 1]^{HW \times HW}$ 是一个对角线为1 的矩阵。元素值越小，几何距离越长。β^G 通过乘法将显式几何先验嵌入到注意力图中，GSA 则获得近邻区域的焦点，如图Fig. 8 所示。具体而言，对于查询，根据几何关系抑制不相关的键值对的权重，增强相关的键值对，这有助于注意力机制建模对象内和对象间的关系。在实际应用中，Eqn. (4) 也可以扩展为多头版本，同时，我们为不同的自注意力头设置不同的衰减率来增强几何指导。

正如先前针对密集预测任务的研究 [35]所示，金字塔结构通常用于编码精细特征。然而，直接使用自注意力机制来编码高分辨率特征，将带来高昂的计算量和内存成本。我们的几何自注意力机制也面临这个问题。因此，受现有稀疏注意力机制 [9, 12,

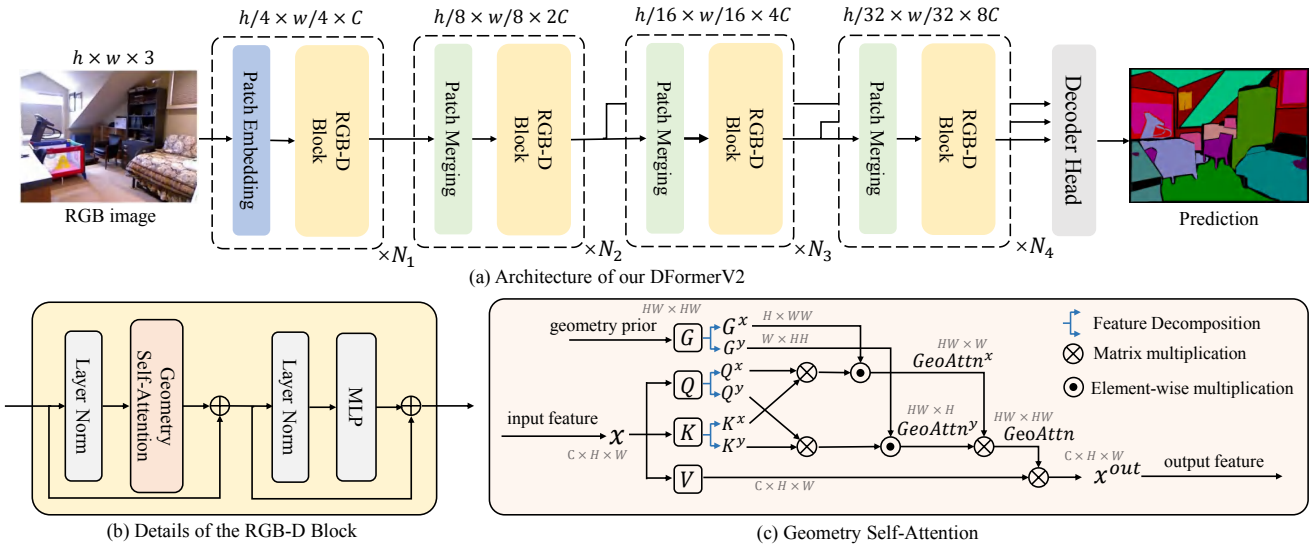


Figure 4. 我们的DFormerv2 的说明。（a）我们的DFormerv2 的整体架构，其中包含一个具有金字塔结构的编码器和一个接收来自最后三个阶段特征的输入的解码器头。（b）基本构建块的详细结构。（c）所提出的几何自注意力机制的详细说明。

Model	Backbone	Params	NYUDepthv2			SUN-RGBD		
			Input size	Flops	mIoU	Input size	Flops	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$_{22}$ [68]	ResNet-34	85.5M	480 × 640	115.6G	53.6	530 × 730	150.0G	51.8
EMSANet$_{22}$ [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**

Table 1. Results on NYU Depth V2 [42] and SUN-RGBD [43]. Some methods do not report the results or settings on the SUN-RGBD datasets, so we reproduce them with the same training configs. † indicates that we follow the results from [59]. All the backbones are pretrained on ImageNet-1K. We split the models to three sets, *i.e.,* small scale, base scale, and large scale. We can see that our method receives the best results on both datasets.

21, 57]的启发，我们使用一种简单的分解方式分别沿水平和垂直方向进行注意力机制，如图Fig. 4(c) 所示。为此，我们还需要生成水平和垂直方向的几何先验。因此，我们将几何先验G 分解为G^x 和G^y，它们分别反映所有token 在行和列上的几何关系。具体而言，$G^y = [G^y_{ij}]_{i=0,1,..,H-1,j=0,1,...,W-1}$ 是一个形状

为(HW, H) 的矩阵，其中G^y_{ij} 表示在(i, j) 处的patch 与第j 列所有patch 之间的几何关系。类似地，我们可以得到形状为(HW, W) 的G^x。然后，几何自注意力的计算公式如下：

$$\text{GeoAttn}^y = (\text{Softmax}(Q^y(K^y)^T) \odot \beta^{G^y}), \quad (5)$$

$$\text{GeoAttn}^x = (\text{Softmax}(Q^x(K^x)^T) \odot \beta^{G^x}), \quad (6)$$

$$\text{GeoAttn} = \text{GeoAttn}^y(\text{GeoAttn}^x V)^T, \quad (7)$$

其中$Q^y(K^y)^T$ 和$Q^x(K^x)^T$ 表示沿垂直轴和水平轴进行注意力计算。

3.3. DFormerv2 结构

Fig. 4 展示了DFormerv2 的整体架构，它遵循广泛使用的编码器-解码器框架。编码器由四个阶段组成，用于生成多尺度特征。每个阶段包含一堆几何自注意力模块。前三个阶段对几何自注意力模块进行分解，最后一个阶段不进行分解。使用轻量级解码器头将这些视觉特征转换为RGB-D 语义分割结果。

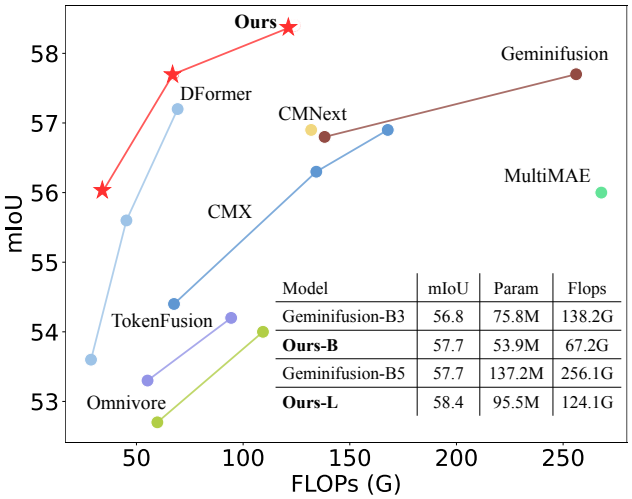
给定一张RGB 图像，首先由一个主干层(stem layer)处理，该层由两个卷积组成，核大小为3×3，步长为2。然后，将RGB 特征输入到分层编码器中，以原始图像分辨率的$\{1/4, 1/8, 1/16, 1/32\}$ 编码多尺度特征。与现有方法不同，我们的DFormerv2 无需显式编码深度图。我们只需在深度图上对编码器中几何自注意力模块对应的四个尺度执行具有不同池化核和步长的平均池化操作，然后利用它们为每个模块生成几何先验即可。基于每个阶段几何自注意力模块的配置，我们设计了一系列编码器变体，分别称为DFormerv2-S、DFormerv2-B 和DFormerv2-L，它们具有相同的架构，但模型大小不同。这些变体的详细配置可在补充材料中找到。

4. 实验

4.1. 实现细节

预训练设置 继DFormer [59] 和MultiMAE [1] 之后，我们在ImageNet-1K 上对我们的DFormerv2 进行RGB-D 预训练，使编码器能够实现RGB 和深度模态之间的交互，并生成具有丰富语义和空间信息的可迁移表征。ImageNet 的深度图由深度估计方法 [59] 生成。我们采用标准交叉熵损失作为优化目标，训练轮数设置为300，与大多数预训练模型 [34] 相同。参照前人研究 [59, 61]，我们采用了AdamW [31]，学习率为1e-3，权重衰减为5e-2，并将批次大小设置为1024。DFormerv2 各个变体的更详细设置请参见补充材料。

用于微调的数据集和一些设置 遵循RGB-D 语义分割研究文献 [20, 28, 59] 中常用的实验设置，我们在两个常用数据集（即NYU DepthV2 [42] 和SUNRGBD [43]）上评估了我们的DFormerv2。此外，我们还在Deliver 数据集 [62] 上进行了实验，就像在 [28] 中所做的那样。与DFormer [59] 一致，我们使用轻量级的head [15] 作为



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

Figure 5. 我们的DFormerv2 与其他SOTA 方法在NYU DepthV2 上的性能计算比较 [42].

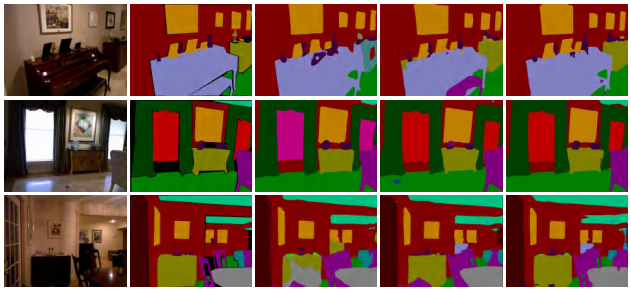


Figure 6. 与GeminiFusion-B5 [28] 和DFormer-L [59]的定性比较。'GT' 表示真值

解码器来构建我们的RGB-D 语义分割模型。我们在微调模型时仅采用两种简单的数据增强方法，即随机水平翻转和随机缩放（从0.5 到1.75）。我们使用交叉熵损失作为优化目标，并使用AdamW [31] 作为优化器，初始学习率为6e-5，采用多项式衰减方案。对于NYU DepthV2 和SUNRGBD 数据集，我们分别将图像裁剪并调整为480×640 和480×480 进行训练。在评估过程中，我们采用所有语义类别的平均交并比(mIoU) 作为衡量分割准确率的主要评估指标。借鉴近期研究成果 [59, 61, 62]，我们在尺度$\{0.5, 0.75, 1, 1.25, 1.5\}$ 上采用多尺度(MS) 翻转推理策略。我们在DeLiVER 上采用与CMNeXt [62] 相同的训练和测试策略，其中图像尺寸调整为1024×1024。更多详情请参阅补充材料。

4.2. 和其它方法的对比

我们将我们的DFormerv2 与17 种最新的RGB-D 语义分割方法在NYU DepthV2 [42]、SUNRGBD [43] 和Deliver [62] 数据集上进行了比较。在Tab. 1 中，我们根据模型规模将所有方法的变体分为三类，即小规模、基础规模和大规模，以便进行更直观、更公平的比较。可以看出，DFormerv2 在两个基准测试的所有模型规模设置上都取得了新的SOTA 性能。我们

Model	Backbone	Params	**Flops**	**mIoU**
HRFuser [2]	HRFormer-T	30.5M	223.0G	51.9
TokenFusion [54]	MiT-B2	26.0M	55.0G	60.3
★ DFormerv2-S	DFormerv2-S	26.7M	28.9G	**63.7**
CMX [61]	MiT-B2	66.6M	65.7G	62.7
CMNext [62]	MiT-B2	58.7M	62.9G	63.6
★ DFormerv2-B	DFormerv2-B	53.9M	60.8G	**65.2**
CMNext [62]	MiT-B4	116.6M	112.0G	66.3
GeminiFusion$_{24}$ [28]	MiT-B5	137.2M	218.4G	66.9
TokenFusion [54]	MiT-B5	83.3M	144.7G	63.5
★ DFormerv2-L	DFormerv2-L	95.5M	114.5G	**67.1**

Table 2. Deliver [62] 数据集上的结果。以下 [62] 数据集上的Flops 计算基于形状为512 × 512 的图像。

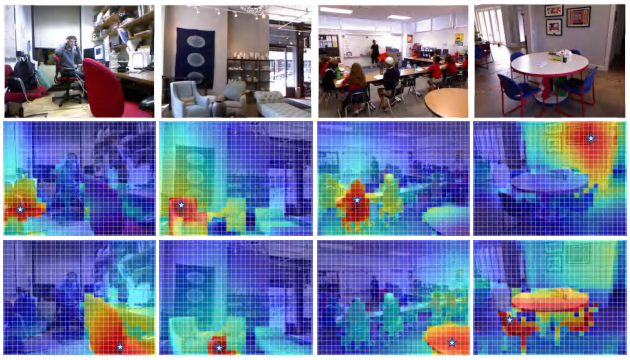


Figure 7. 一些几何先验的可视化示例。蓝色“星号”表示当前查询标记。需要注意的是，可视化先验仅从深度图中获得。

还在Fig. 5 中绘制了不同方法的性能计算成本曲线。与其他方法相比，DFormerv2 实现了更佳的性能和计算量平衡。具体而言，我们最大的模型DFormerv2-L 在95.5M 参数和124.1G Flops 的条件下实现了58.4% mIoU，比排名第二的方法Gemnifusion [28] 快0.7同样，在基础和小规模上，我们的DFormerv2 也始终以更高的效率优于其他SOTA 方法。在SUNRGBD和Deliver (Tab. 2) 数据集上，我们的DFormerv2 也带来了显著的改进。此外，Fig. 6 展示了我们的DFormerv2与Gemnifusion [28] 的语义分割结果之间的视觉对比。这些改进表明，我们的DFormerv2 可以更有效地利用深度图中的几何先验，而无需显式编码，从而以更低的计算成本获得更准确的预测。

4.3. 模型分析

几何自注意力机制 所提出的几何自注意力机制由深度先验、空间先验和先验融合组成，并将它们集成为统一的几何先验。为了评估每个组件的有效性，我们在Tab. 3 中展示了从原始自注意力机制到几何自注意力机制的路线图。首先，我们分别将深度先验和空间先验引入原始自注意力机制（步骤1-2），以观察其对性能的影响。与基线相比，这些替换分别使NYUDepthV2 数据集的准确率提高了2.6% 和1.8%，以及SUNRGBD 数据集的准确率提高了1.7% 和1.3%，这凸显了将这些先验引入自注意力机制的重要性。然而，同样明显的是，简单地将深度先验和空间先验相

Step	Attention	Params	Flops	NYUDepthV2	SUNRGBD
0	Vanilla Attn	26.5M	51.4G	51.7	47.8
1	+Only Depth Prior	26.5M	51.4G	54.3 (**+2.6**)	49.9 (**+1.7**)
2	+Only Spatial Prior	26.5M	51.4G	53.5 (**+1.8**)	49.1 (**+1.3**)
3	+Both Priors	26.7M	51.7G	56.2 (**+4.5**)	51.7 (**+3.9**)
4	+decomposition	26.7M	33.9G	56.0 (**+4.3**)	51.5 (**+3.7**)

Table 3. 消融实验展示了从原始自注意力机制到几何自注意力机制的完整路线图，该路线图在DFormerv2 的小规模上得以实现。在步骤0 和2 中，我们仅输入RGB 图像，而在其他所有步骤中均使用RGB-D 图像。

Figure 8. focused attention的可视化

Figure 9. 具有和不具有几何先验的特征可视化。它们是从第一阶段的输出中随机挑选出来的。

加，相比仅使用深度先验，效果仅略有提升，这表明这种集成方法可能并不有效。在步骤3中，当我们引入融合操作来桥接两个先验并形成几何先验时，我们观察到NYUDepthV2和SUNRGBD的性能得到了进一步提升，而计算成本的增加却微乎其微。这些结果（步骤1-3）表明，几何先验显著提升了性能，而复杂度几乎没有增加。此外，在步骤4中，我们对注意力机制进行了分解，这进一步减轻了计算负担，同时保持了几乎相同的性能水平。总体而言，与自注意力机制相比，集成几何先验能够以最小的计算开销和略微增加的参数实现更好的RGB-D分割。

关于几何先验的更多理解 形状为$HW \times HW$ 的几何先验G 源自深度图，表示每对标记之间的几何关系。为了更深入地了解此先验，我们随机选择了几个标记，并在Fig. 7 中可视化它们与其他标记的几何关系。对于每个查询标记，几何先验可以准确识别其所属的对象，并捕捉该对象与其附近对应对象之间的几何关系。对物体之间几何关系的感知可以帮助我们的模型更好地区分不同的语义物体，例如，椅子通常位于桌子下方，而人通常坐在椅子上。GSA 中的集中注意力机制在Fig. 8 中进行了可视化。在自注意力机制之前引入几何关系，使模型能够更好地理解物体的几何结构以及复杂场景中物体之间的位置关系，从而获得更

Method	Params	Flops	NYUDepthV2	SUNRGBD
Conv	26.9M	34.3G	55.8	51.3
Addition	26.5M	33.5G	54.6	50.4
Hadamard	26.5M	33.6G	54.9	50.9
Memory	26.7M	33.9G	56.2	51.7

Table 4. 在我们的小规模模型上，使用不同的操作来连接深度先验和空间先验。

Settings	NYUDepthV2	SUNRGBD
fixed to 0.25 for all heads	55.7	51.1
fixed to 0.5 for all heads	55.5	51.0
fixed to 0.75 for all heads	55.7	51.2
linearly sampled in [0.5, 1.0)	55.9	51.5
linearly sampled in [0.75, 1.0)	56.0	51.5

Table 5. 几何自注意力中不同衰减策略对DFormerv2-S 的影响。

Model	Params	FLOPs	Latency↓	NYU DepthV2
Omnivore [16]	29.1M	32.7G	40.1ms	49.7
DFormer-B [59]	29.5M	41.9G	42.8ms	55.6
DFormerv2-S	26.7M	33.9G	43.9ms	56.0
CMX-B2 [61]	66.6M	65.6G	71.5ms	54.4
DFormer-L [59]	39.0M	69.3G	44.5ms	57.2
GeminiFusion-B3 [28]	75.8M	138.2G	68.2ms	56.8
DFormerv2-B	53.9M	67.2G	50.7ms	57.7
CMX-B5 [61]	181.1M	167.8G	114.9ms	56.9
CMNext-B4 [62]	119.6M	131.9G	98.5ms	56.9
MultiMAE [1]	95.2M	267.9G	76.9ms	56.0
GeminiFusion-B5 [28]	137.2M	256.1G	108.7ms	57.7
DFormerv2-L	95.5M	124.1G	79.9ms	58.4

Table 6. 我们的方法与最近的SOTA 模型之间的推理延迟比较。'↓'：越低越好。

准确的分割结果。此外，我们还对使用和不使用几何先验的特征图进行了可视化，如图Fig. 9 所示。可以看出，引入几何先验可以帮助我们的模型更好地捕捉物体的细节，并提升分割性能。

融合操作 为了在深度先验和空间先验之间建立桥梁并形成几何先验，我们利用记忆权重，根据所有图像块标记之间的深度和空间距离形成几何线索。为了验证融合操作的有效性，我们还使用了一些其他操作来替代它，包括加法、Hadamard 积和卷积层。如Tab. 4 所示，我们可以看到记忆权重比其他操作能带来更好的结果。

衰减率 当将几何先验引入到如公式 (4) 所示的注意机制中时，我们使用衰减率β 来控制先验对特征的影响程度。本文，我们研究了当采用不同的衰减率策略时，模型性能会如何变化。结果可在Tab. 5 中找到。结果表明，在我们的几何自注意力机制中，为不同的头部分配不同的衰减率β 可以实现多尺度增强和多样性，从而进一步提升性能。因此，我们默认在$[0.75, 1.0]$ 范围内采样β 的值。

Modality	**Params**	Classification	Segmentation	
		Top-1 Acc↑	wF↑	MAE↓
RGB	26.5M	83.1	0.818	0.054
Depth	26.5M	43.8	0.715	0.061
RGB+Depth	26.7M	83.4	0.868	0.048

Table 7. 不同输入模态对捕捉语义类别和物体形状的影响。加权F值（wF）和平均绝对误差（MAE）是前景分割任务的两个常用指标 [29, 58, 66]。

推理延迟 实时推理速度对于RGB-D 模型在各种下游应用中的实际部署至关重要 [5]。为了确保公平比较，所有测试均在相同的硬件配置下进行，使用单个3090 RTX GPU，图像分辨率均为480×640。如Tab. 6 所示，DFormerv2 在速度和准确率之间取得了良好的平衡。

关于**RGB**和深度的影响的讨论 语义分割为每个像素分配一个类别标签，可以看作是物体分类和分割的结合。本文，我们探讨了这两种模式如何有助于捕捉语义类别和物体形状，从而更深入地理解我们提出的几何自注意力机制的设计。为此，我们在LUSS [14] 数据集上进行了实验，该数据集为来自ImageNet [38] 的5万张图像提供了分割标注。我们将数据分为训练集、验证集和测试集，并针对分类和前景分割任务训练模型。如Tab. 7 所示，我们可以看到深度内的3D 几何信息主要帮助模型分割对象，并略微帮助捕获语义。

5. 总结

我们提出了DFormerv2，这是一种融合了显式几何先验的RGBD 视觉主干模型。DFormerv2 利用深度来建模图像块之间的几何关系，然后利用此先验在自注意力机制（称为几何自注意力）中分配注意力权重。得益于这种定制的注意力机制，我们的方法能够更有效地利用深度模态。实验表明，DFormerv2 在RGB-D 语义分割方面比近期方法取得了更好的结果，并且计算成本更低。

致谢 本研究由国家自然科学基金（编号：62225604、62176130）、天津市科技支撑计划（编号：23JCZDJC01050）和深圳市科技计划（编号：JCYJ20240813114237048）资助。南开大学超级计算中心提供了部分计算支持。

References

[1] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. MultiMAE: Multi-modal multi-task masked autoencoders. In *ECCV*, 2022. 1, 4, 5, 7

[2] Tim Broedermann, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Hrfuser: A multi-resolution sensor fusion architecture for 2d object detection. In *IEEE ITSC*, 2023. 6

[3] Jinming Cao, Hanchao Leng, Dani Lischinski, Danny Cohen-Or, Changhe Tu, and Yangyan Li. ShapeConv: Shape-aware convolutional layer for indoor RGB-D semantic segmentation. In *ICCV*, 2021. 4

[4] Lin-Zhuo Chen, Zheng Lin, Ziqin Wang, Yong-Liang Yang, and Ming-Ming Cheng. Spatial information guided convolution for real-time RGBD semantic segmentation. *TIP*, 30: 2313–2324, 2021. 4

[5] Xiaokang Chen et al. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In *ECCV*, 2020. 4, 7

[6] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 2

[7] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 2

[8] Senlin Cheng and Haopeng Sun. Spt: Sequence prompt transformer for interactive image segmentation. *arXiv preprint arXiv:2412.10224*, 2024. 3

[9] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *CVPR*, 2022. 2, 3

[10] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 3

[11] Siqi Du, Weixi Wang, Renzhong Guo, Ruisheng Wang, and Shengjun Tang. Asymformer: Asymmetrical cross-modal representation learning for mobile platform real-time rgb-d semantic segmentation. In *CVPRW*, 2024. 1, 4

[12] Qihang Fan, Huaibo Huang, Mingrui Chen, Hongmin Liu, and Ran He. Rmt: Retentive networks meet vision transformers. In *CVPR*, 2024. 3

[13] Huan Gao, Jichang Guo, Guoli Wang, and Qian Zhang. Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation. In *CVPR*, 2022. 2

[14] Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *IEEE TPAMI*, 45(6):7457–7476, 2022. 7

[15] Zhengyang Geng, Meng-Hao Guo, Hongxu Chen, Xia Li, Ke Wei, and Zhouchen Lin. Is attention better than matrix decomposition? *NeurIPS*, 2021. 5

[16] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *CVPR*, 2022. 4, 7

[17] Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. Cmt: Convolutional neural networks meet vision transformers. In *CVPR*, 2022. 3

[18] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *NeurIPS*, 2022. 2

[19] Dongchen Han, Tianzhu Ye, Yizeng Han, Zhuofan Xia, Siyuan Pan, Pengfei Wan, Shiji Song, and Gao Huang. Agent attention: On the integration of softmax and linear attention. In *ECCV*, 2025. 3

[20] Zhiwei Hao, Zhongyu Xiao, Yong Luo, Jianyuan Guo, Jing Wang, Li Shen, and Han Hu. Primkd: Primary modality guided multimodal fusion for rgb-d semantic segmentation. In *ACM MM*, 2024. 1, 5

[21] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *CVPR*, 2023. 4

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1

[24] Qibin Hou, Cheng-Ze Lu, Ming-Ming Cheng, and Jiashi Feng. Conv2former: A simple transformer-style convnet for visual recognition. *IEEE TPAMI*, 2024. 3

[25] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. ACNet: Attention based network to exploit complementary features for RGBD semantic segmentation. In *ICIP*, 2019. 2

[26] Keli Huang, Botian Shi, Xiang Li, Xin Li, Siyuan Huang, and Yikang Li. Multi-modal sensor fusion for auto driving perception: A survey. *arXiv preprint arXiv:2202.02703*, 2022. 1

[27] Wei Ji, Jingjing Li, Qi Bi, Tingwei Liu, Wenbo Li, and Li Cheng. Segment anything is not always perfect: An investigation of sam on different real-world applications. *MIR*, 2024. 2

[28] Ding Jia, Jianyuan Guo, Kai Han, Han Wu, Chao Zhang, Chang Xu, and Xinghao Chen. Geminifusion: Efficient pixel-wise multimodal fusion for vision transformer. In *ICML*, 2024. 1, 2, 4, 5, 6, 7

[29] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *CVPR*, 2013. 7

[30] Peng-Tao Jiang, Yuqi Yang, Yang Cao, Qibin Hou, Ming-Ming Cheng, and Chunhua Shen. Traffic scene parsing through the tsp6k dataset. In *CVPR*, 2024. 1

[31] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[32] Zheng Lin, Zhao Zhang, Zi-Yue Zhu, Deng-Ping Fan, and Xia-Lei Liu. Sequential interactive image segmentation. *CVM*, pages 753–765, 2023. 2

[33] Chang Liu, Xudong Jiang, and Henghui Ding. Primitivenet: decomposing the global constraints for referring segmentation. *Visual Intelligence*, 2024. 2

[34] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 3, 5

[35] Ze Liu et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2, 3

[36] Nicolas Marchal, Charlotte Moraldo, Hermann Blum, Roland Siegwart, Cesar Cadena, and Abel Gawel. Learning densities in feature space for reliable segmentation of indoor scenes. *RA-L*, 5(2):1032–1038, 2020. 1

[37] Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *ICLR*, 2022. 3

[38] Olga Russakovsky et al. ImageNet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 7

[39] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. Efficient rgb-d semantic segmentation for indoor scene analysis. In *ICRA*, 2021. 2

[40] Daniel Seichter, Söhnke Benedikt Fischedick, Mona Köhler, and Horst-Michael Groß. Efficient multi-task rgb-d scene analysis for indoor environments. In *IJCNN*, 2022. 4

[41] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *NAACL*, 2018. 3

[42] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012. 2, 4, 5

[43] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *CVPR*, 2015. 2, 4, 5

[44] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, 2021. 2

[45] Boyuan Sun, Yuqi Yang, Le Zhang, Ming-Ming Cheng, and Qibin Hou. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. In *CVPR*, 2024. 2

[46] Haopeng Sun. Ultra-high resolution segmentation via boundary-enhanced patch-merging transformer. *AAAI*, 2025. 3

[47] Haopeng Sun, Lumin Xu, Sheng Jin, Ping Luo, Chen Qian, and Wentao Liu. Program: Prototype graph model based pseudo-label learning for test-time adaptation. In *ICLR*, 2024. 1

[48] Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023. 3

[49] Lichun Tang, Zhaoxia Yin, Hang Su, Wanli Lyu, and Bin Luo. Wfss: weighted fusion of spectral transformer and spatial self-attention for robust hyperspectral image classification against adversarial attacks. *Visual Intelligence*, 2024. 3

[50] Hengyi Wang, Jingwen Wang, and Lourdes Agapito. Co-slam: Joint coordinate and sparse parametric encodings for neural real-time slam. In *CVPR*, 2023. 1

[51] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *CVM*, 8(3):415–424, 2022. 3

[52] W Wang, L Yao, L Chen, B Lin, D Cai, X He, and W Liu. Crossformer: A versatile vision transformer hinging on cross-scale attention. In *ICLR*, 2022. 2

[53] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. *NeurIPS*, 2020. 4

[54] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *CVPR*, 2022. 2, 4, 6

[55] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *CVPR*, 2022. 3

[56] Qi Xu, Yinan Ma, Jing Wu, Chengnian Long, and Xiaolin Huang. Cdada: A curriculum domain adaptation for nighttime semantic segmentation. In *ICCV*, 2021. 2

[57] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021. 2, 4

[58] Bowen Yin, Xuying Zhang, Deng-Ping Fan, Shaohui Jiao, Ming-Ming Cheng, Luc Van Gool, and Qibin Hou. Camoformer: Masked separable attention for camouflaged object detection. *IEEE TPAMI*, 2024. 7

[59] Bowen Yin, Xuying Zhang, Zhongyu Li, Li Liu, Ming-Ming Cheng, and Qibin Hou. Dformer: Rethinking rgbd representation learning for semantic segmentation. *ICLR*, 2024. 1, 2, 4, 5, 7

[60] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *CVPR*, 2022. 3

[61] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *T-ITS*, 2023. 1, 2, 4, 5, 6, 7

[62] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *CVPR*, 2023. 1, 2, 4, 5, 6, 7

[63] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *CVPR*, 2019. 2

[64] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 2

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

[66] Tao Zhou, Deng-Ping Fan, Geng Chen, Yi Zhou, and Huazhu Fu. Specificity-preserving rgb-d saliency detection. *CVM*, pages 297–317, 2023. 7

[67] 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 TMM*, 2022. 4

[68] Wujie Zhou, Enquan Yang, Jingsheng Lei, and Lu Yu. Frnet: Feature reconstruction network for rgb-d indoor scene parsing. *JSTSP*, 16(4):677–687, 2022. 4

[69] Lei Zhu, Xinjiang Wang, Zhanghan Ke, Wayne Zhang, and Rynson WH Lau. Biformer: Vision transformer with bi-level routing attention. In *CVPR*, 2023. 3