基于卷积神经网络的实时 RGBD 语义分割

Lin-Zhuo Chen, Zheng Lin, Ziqin Wang, Yong-Liang Yang, and Ming-Ming Cheng

摘要—众所周知，3D 空间信息对于语义分割任务是十分有利的。现有的方法大多数都将 3D 空间数据作为额外的输入，使用双流的分割网络来分别处理 RGB 和 3D 信息。这样的解决方法大大的增加了推理时间，严重的限制了其实时推断的能力。为了解决这个问题，我们提出了一种基于空间信息的卷积方法 (S-Conv)，它可以有效地集成 RGB 特征和 3D 空间信息。S-Conv 能够在输入图片空间信息的指导下推断卷积核的权重与采样偏移量，帮助卷积层自适应调整感受野并适应物体的几何变换。由于空间信息的直接输入，S-Conv 可以直接分析出物体的尺度和空间变换，并生成对应的权值与卷积核分布，从而更好地感知场景中物体的空间关系与几何形状。其可以在增加少量计算量和参数量的情况下，充分地利用空间信息并显著地提升语义分割网络的性能。本文基于 S-Conv 进一步设计了一个实时 RGBD 语义分割网络，名为空间信息引导卷积网络 (SGNet)。SGNet 在通用数据集上，例如 NYUDv2 数据集与 SUNRGBD 数据集，达到了实时推理速度，并与其他方法相比有着最优的性能。

Index Terms—**空间信息**, **感受野**, RGBD **语义分割**.

I. 引言

随着 3D 传感器的应用普及，图片的空间信息变得易于获取。因此，用于高级场景理解的 RGBD 语义分割变得非常重要，有利于诸如自动驾驶 [1]、SLAM [2] 和机器人等广泛的应用。由于卷积神经网络 (CNN) 的有效性和额外的空间信息，最近的工作在室内场景分割任务上有显著的提升 [3]–[5]。然而，对于实时推理的网络结构现存的一个重大挑战是需要同时考虑环境的复杂性和利用空间数据的方式。

一种常见的方法是将 3D 空间信息作为额外的输入，然后结合 RGB 图像的特征来融合多模态信息 [6]–[10] (参照 Fig. 1(a))。该方法在显著增加参数数量和计算时间的代价下获得了较好的结果，因而不适合实时任务。同时，有的工作 [3], [6], [9], [11], [12] 将原始空间信息编码为三个通道 (HHA)，由水平视差、地面高度和标准角度组成。然而，从原始数据到 HHA 的转换也很耗时 [9]。

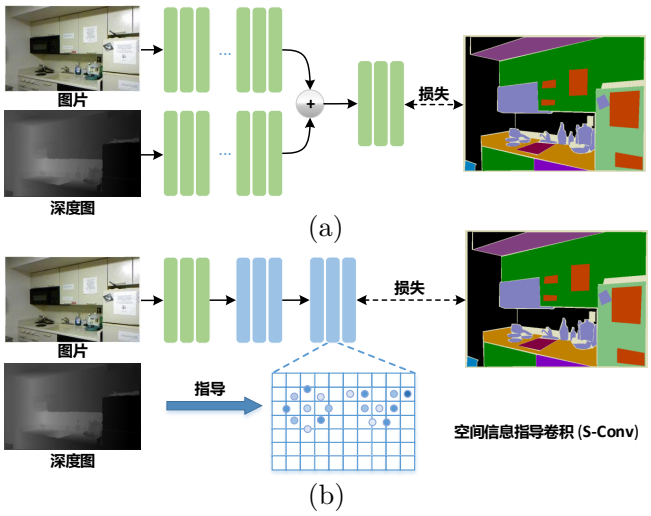


图 1. **不同多模态融合方法的网络结构**. (a) 传统的双流结构 [6]–[10]. (b) 本文提出的 SGNet. 可以看出，(a) 中的方法由于处理空间信息，大大增加了参数数量和推断时间，不太适合实时应用。我们用 (b) 中的 S-Conv 代替卷积，其中卷积核的分布和权值是自适应的空间信息。S-Conv 在减少额外参数和计算的情况下，大大提高了网络的空间感知能力，从而可以有效地利用空间信息。在彩色图中可以更好的体现。

值得注意的是，室内场景的空间关系比室外场景更为复杂。这就要求网络具有较强的适应几何变换的能力。但由于卷积核的固定结构，上述方法中的二维卷积不能很好地适应空间变换，和自适应地调整感受野，会限制语义分割的准确性。虽然可以通过修改池化操作和先验数据增强 [13], [14] 来缓解上述问题，但卷积仍需要更好的空间自适应采样机制。

此外，室内场景中物体的颜色和纹理并不总是具有代表性 [15]。相反，几何结构在语义分割中往往起着至关重要的作用。例如，要识别冰箱和墙壁，由于纹理相似，几何结构是首要线索。但对 RGB 数据进行二维卷积会忽略这些空间信息。深度感知卷积 (depth-aware convolution) [16] 的提出就是为了解决这个问题。其迫使与核中心深度相似的像素拥有比其他像素更高的权值。然而，这个先验条件是人工添加的，可能导致次优

LZ Chen(linzhuochen@mail.nankai.edu.cn), Z Lin, MM Cheng (通讯作者, cmm@nankai.edu.cn) 在中国南开大学计算机科学学院 TKLNDST 工作。

Ziqin Wang 目前在悉尼大学工作。

YL Yang 目前在巴斯大学工作。

结果。

可以得知，二维卷积的固定结构与变化的空间变换之间存在矛盾，同时还包括 RGB 和空间数据分别处理时的效率瓶颈。为了克服上述局限性，我们提出了一种新的卷积运算方式，称为空间信息引导卷积 (S-Conv)，其卷积核分布随空间信息自适应变化 (参照 Fig. 1(b))。具体来讲，此操作可以生成自适应空间信息的含有不同采样分布的卷积核。进而增强网络的空间变换适应能力和感受野调节能力。S-Conv 在卷积权值与底层像素空间关系之间建立联系，将几何信息融入到卷积核的权值中，以便更好地捕捉场景的空间结构。由于会输入空间信息到 S-Conv 中，S-Conv 可以直接分析对象的尺度和空间变换，生成空间自适应偏移量和权值。

我们提出的 S-Conv 是轻而灵活的，仅仅使用了少量的参数量和计算量即可显著的提升语义分割的精度，使其可以胜任实时语义分割的任务。此方法可以看作是一种新的多模态融合方法。具体来说，与其他双流方法相比，我们利用空间信息引导卷积过程，以达到多模态融合的目的。它的性能优于其他依靠双流网络的方法，与双流方法相比，大大减少了参数量和计算量，实现了实时应用。本文通过实验来说明 S-Conv 设计的的有效性以及高效性。我们首先设计了消融实验研究，并将 S-Conv 与双流方法、可变形卷积 (deformable convolution [13], [14]) 和深度感知卷积 (depth-aware convolution [16]) 进行了比较，以展示 S-Conv 的优点。通过测试 S-Conv 对深度、HHA 和三维坐标等不同类型空间数据的影响，验证了 S-Conv 在空间变换中的适用性。我们证明空间信息比可变形卷积使用的 RGB 特征更适合生成偏移量。通过实验，基于 S-Conv 的空间信息引导卷积网络 (Spatial information Guided convolutional Network, SGNet) 不仅可以对 NYUDv2 [17] 和 SUNRGBD [18], [19] 数据集进行实时推理，还获得了高质量的结果。

我们的贡献如下：

- 我们提出了一种新的 S-Conv 算子，该算子能够在有效适应空间变换的同时自适应调整感受野，并能以较低的代价感知复杂的几何形状。
- 基于 S-Conv，我们提出了一个新的 SGNet 网络，在 NYUDv2 [17] 和 SUNRGBD [18], [19] 数据集上实现了具有竞争力的实时 RGBD 分割性能。

II. 相关工作

A. 语义分割

近年来，卷积神经网络（CNN）的发展为语义分割的研究提供了新的思路 [20], [21]。FCN [3] 是将 CNN 应用在语义分割上的先驱工作，在各个语义分割数据集上取得了令人信服的结果。并成为成为各大像素级分类任务的基本框架。目前的方法可以依据网络结构分为两类，包括基于空洞卷积的方法 [4], [22]–[24] 和编码器-解码器的架构 [25]–[30].

空洞卷积: 标准的方法依赖于步长大于 1 的卷积或池化来减少 CNN 主干网络的的输出尺寸，并尽可能使用大的感受野。然而，生成的特征图的分辨率降低 [4]，许多细节丢失。一种方法是利用空洞卷积增强感受野，同时保持特征图的分辨率，以缓解感受野与特征图分辨率之间的矛盾。[4], [22], [26], [31] 我们在提出的 SGNet 中使用基于空洞卷积的主干网络。

编码器-解码器架构: 其他方法是用编码器解码器结构 [25]–[30], [32]，其学习解码器以逐渐恢复细节。DeconvNet [28] 采用一系列反卷积层 (deconvolutional layers) 来产生高分辨率预测。SegNet [27] 通过在编码器中使用池化索引来指导解码器中的恢复过程，以便获得更好的结果。RefineNet [25] 将编码器中的低级特征与解码器融合以细化预测。 [29], [30] 提出了门控和 (gated sum) 的方案，可以控制编码器-解码器架构中不同尺度的信息流。虽然这种方法可以获得更精确的结果，但它需要更长的推理时间。

B. RGBD 语义分割

如何有效利用额外的几何信息（深度、3D 坐标）是 RGBD 语义分割的关键。一些工作更多关注如何从几何中提取更多信息 [7]–[10], [33]。[6], [8]–[10], [12] 使用双流网络分别对 RGB 图像和几何信息进行处理，最后一层将两者的结果合并。这些方法以参数和计算成本加倍为代价取得了有效的结果。3D CNN 或 3D KNN 图网络也用于考虑几何信息 [34]–[36]。此外，还有的方法 [37]–[42] 探索了 3D 点云上的各种深度学习方法。然而，这些方法消耗大量内存并且计算成本高。另一类方法将几何信息合并到特征提取操作中。[43] 提出基于深度引导卷积执行 3D 目标检测，其权重是位置变量和深度自适应 Cheng et al. [44] 使用几何信息来构建一个特征亲和矩阵，用于平均池化和上采样池化 (up-pooling)。Lin 等人 [45] 根据几何信息将图像分成不同的分支。Wang

等人 [16] 提出了深度感知 CNN(Depth-aware CNN)，它在卷积权重之前增加了深度。虽然它通过卷积改进了特征提取，但先验是人工提出的，而不是从数据中学习的。其他方法，例如多任务学习 [7], [46]–[50] 或时空分析 [51]，进一步用于提高分割精度。所提出的 S-Conv 旨在有效地利用空间信息来提高特征提取能力。由于只使用少量的参数，它可以显着提高性能和效率。

C. 卷积神经网络的动态结构

还有方法探索了使用动态结构来处理卷积神经网络的不同输入。在 [4], [22] 中使用了扩张卷积 (Dilation Convolution) 来增加感受野大小而不降低特征图分辨率。空间变换器网络 (Spatial transformer network) [52] 通过令特征图变形来适应空间变换。动态过滤器 (Dynamic filter) [53] 根据输入自适应地改变其权重。此外，基于自注意力的方法 [54]–[57] 从中间特征图生成注意图，以调整每个位置的响应或自适应地捕获远程上下文信息。专注于上下文语义的理解，Shape-variant convolution [57] 通过基于语义相关区域的位置变量卷积来限制其上下文区域。还有从 2D 图像到 3D 点云的卷积的一些泛化之后的工作。PointCNN [42] 是一项开创性的工作，可以在一组无序 3D 点上启用 CNN。还有其他改进 [39]–[41] 关于利用神经网络从 3D 点集中有效提取深度特征。可变形卷积 [13], [14] 可以生成具有自适应权重的不同分布。然而，他们的输入是中间特征图而不是空间信息。我们的工作通过实验验证了基于空间信息可以获得更好的结果，详见 Sec. IV。

III. S-Conv 与 SGNet

在本节中，我们首先详细说明空间信息引导卷积 (S-Conv) 的细节，它是传统卷积在 RGBD 输入情况下，充分利用空间信息的泛化。然后，我们讨论我们的 S-Conv 和其他方法之间的关系。最后，我们描述了空间信息引导卷积网络 (SGNet) 的网络架构，它以 S-Conv 为基础，实现 RGBD 语义分割任务。

A. 空间信息引导卷积网络 (S-Conv)

我们首先回顾传统的卷积操作。我们使用 $\mathrm{A}_i(\mathrm{j}), \mathrm{A} \in \mathbb{R}^{c\times h\times w}$ 来表示一个张量，其中 i 是第一维对应的索引，$\mathrm{j} \in \mathbb{R}^2$ 表示第二维和第三维的两个索引。为方便起见，非标量值以粗体突出显示。

对于一个输入的特征图 $\mathrm{F} \in \mathbb{R}^{c\times h\times w}$. 为简单起见，我们用 2D 来描述它，因此我们将 X 作为输入特征图。$\mathrm{X} \in \mathbb{R}^{1\times h\times w}$。其扩展到 3D 情况很简单。

输入 X 输出 Y 的常规卷积可以表述为如下：

$$\mathrm{Y}(\mathrm{p}) = \sum_{i=1}^{K} \mathrm{W}_i \cdot \mathrm{X}(\mathrm{p} + \mathrm{d}_i), \tag{1}$$

其中 $\mathrm{W} \in \mathbb{R}^K$ 表示卷积核的权重，核大小为 $k_h \times k_w$，$K = k_h \times k_w$。$\mathrm{p} \in \mathbb{R}^2$ 是二维卷积中心，$\mathrm{d} \in \mathbb{R}^{K\times 2}$ 表示周围的核分布 p. 对于 3×3 卷积，d 定义为 Equ. (2)：

$$\mathrm{d} = \{[-1,-1],[-1,0],...,[0,1],[1,1]\}. \tag{2}$$

从上面的等式可以看出，卷积核在 X 上是常数。换句话说，W 和 d 是固定的，这意味着卷积核在特征图的任何位置结构固定，并不知道特征图的空间信息。

在 RGBD 上下文中，我们希望通过使用自适应卷积核有效地融合 3D 空间信息。我们首先根据空间信息生成偏移量，然后使用给定偏移量对应的空间信息生成新的空间自适应权重。我们的 S-Conv 需要两个输入。一种是与常规卷积相同的特征图 X。另一个则是空间信息 $\mathrm{S} \in \mathbb{R}^{c'\times h\times w}$。在实践中，S 可以是 HHA（$c'=3$）、3D 坐标（$c'=3$）或深度（$c'=1$）。深度编码成 3D 坐标和 HHA 的方法同 [36]。请注意，输入空间信息不包含在特征图中。

作为 S-Conv 的第一步，我们将输入的空间信息投影到一个高维特征空间中，可以表示为：

$$\mathrm{S}' = \phi(\mathrm{S}), \tag{3}$$

其中 ϕ 是空间变换函数，而 $\mathrm{S}' \in \mathbb{R}^{64\times h\times w}$，比 S 具有更高的维度。

然后，我们考虑变换后的空间信息 S′，感知其几何结构，并在不同的 p 上生成不同卷积核的分布（$x-$ 和 $y-$ 轴上的像素坐标偏移）。这个过程可以表示为：

$$\mathrm{d} = \eta(\mathrm{S}'), \tag{4}$$

其中 $\mathrm{d} \in \mathbb{R}^{K\times h'\times w'\times 2}$，为简单起见，我们并没有在 Equ. (4) 中展示 d 的 reshape 过程。reshape 之前 $\mathrm{d} \in \mathbb{R}^{2K\times h'\times w'}$。$h', w'$ 表示在卷积之后的特征图尺寸。$K = k_h \times k_w$，其中 k_h 和 k_h 是卷积核大小。对于 3×3 卷积，$\mathrm{d} \in \mathbb{R}^{9\times h'\times w'\times 2}$. η 是一个非线性函数，这可以通过一系列卷积来实现。

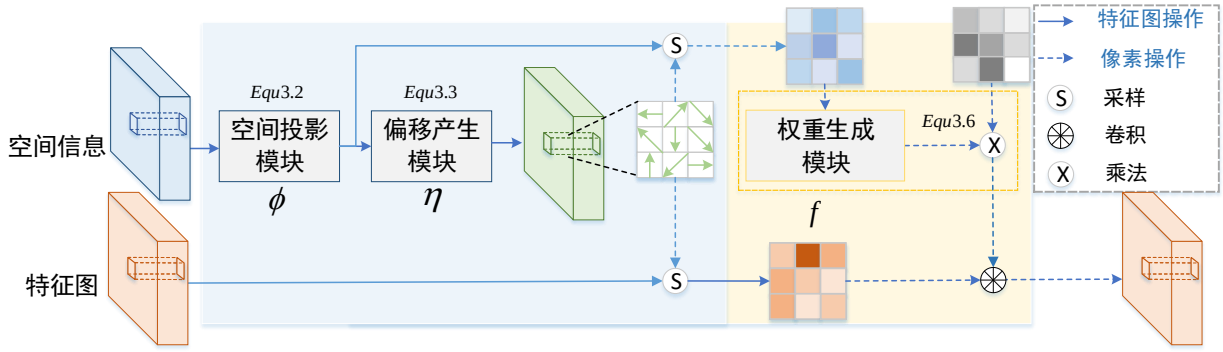


图 2. **空间信息引导卷积（S-Conv）示意图**。首先，输入的 3D 空间信息由 **空间投影模块**投影以匹配输入的特征图。其次，自适应卷积核分布由**偏移产生模块**生成。最后，根据核分布对投影的空间信息进行采样，并将其输入**权重生成模块**以生成自适应卷积权重。

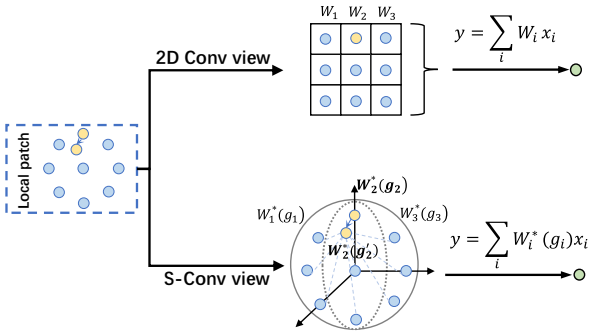


图 3. **有关 2D 卷积中的权重 W 和 S-Conv 中的 W^* 的说明**。黄点表示空间位置沿箭头变化的点。2D 卷积的插图在顶部，S-Conv 在底部。传统的 2D 卷积操作将局部点有序地放置在具有固定权重的规则网格中，同时忽略了空间信息。我们可以看到，黄点的空间位置变化并不能反映在权重上。我们的 S-Conv 可以被视为将数据的邻域放入权重空间中，权重空间是由邻域的空间引导生成的。因此，每个点的权重与其空间位置建立了联系，有效地捕捉了邻域的空间变化。黄点与其他点之间的空间关系可以通过自适应权重来体现。

在使用 $\mathrm{d}(\mathrm{p})$ 为每个可能的 p 生成卷积核分布后，我们通过建立几何结构和卷积权重的关联来强化其特征提取能力。由于 Equ. (4) 中卷积核的移位，卷积核对应的深度信息也发生了变化。我们要收集移位后卷积核对应的深度信息，以生成空间自适应权重。更具体地说，我们对平移后卷积核对应的像素的几何信息进行采样：

$$\mathrm{S}^*(\mathrm{p}) = \{\mathrm{S}'(\mathrm{p} + \mathrm{d}_i + \Delta\mathrm{d}_i(\mathrm{p}))|_{i=1,2,\ldots,K}\}, \tag{5}$$

其中 $\mathrm{d}(\mathrm{p})$ 是 p 处卷积核的空间分布。$\mathrm{S}^*(\mathrm{p}) \in \mathbb{R}^{64K}$ 是以 p 为中心，并在变换后的卷积核的特征图对应的空间信息。

最后，我们根据最终的空间信息生成卷积权重如下：

$$\mathrm{W}^*(\mathrm{p}) = \sigma(f(\mathrm{S}^*(\mathrm{p}))) \cdot \mathrm{W}, \tag{6}$$

其中 f 是一个非线性函数，具有实现为一系列包含非线性激活函数的全连接层的功能，σ 是 sigmoid 函数，$\cdot$ 是元素乘积，$\mathrm{W} \in \mathbb{R}^K$ 表示卷积权重，可以通过梯度下降算法更新。$\mathrm{W}^*(\mathrm{p}) \in \mathbb{R}^K$ 表示以 p 为中心平移后卷积的空间自适应权重。

总的来说，我们广义的 S-Conv 公式如下：

$$\mathrm{Y}(\mathrm{p}) = \sum_{i=1}^{K} \mathrm{W}_i^*(\mathrm{p}) \cdot \mathrm{X}(\mathrm{p} + \mathrm{d}_i + \Delta\mathrm{d}_i(\mathrm{p})). \tag{7}$$

我们可以看到 $\mathrm{W}_i^*(\mathrm{p})$ 建立了空间信息和卷积权重之间的相关性。此外，卷积核分布也因 $\Delta\mathrm{d}$ 而与空间信息相关。注意 $\mathrm{W}_i^*(\mathrm{p})$ 和 $\Delta\mathrm{d}_i(\mathrm{p})$ 不是常数，意思是广义卷积适应不同的 p。此外，由于 $\Delta\mathrm{d}$ 通常是小数，我们使用双线性插值来计算 $\mathrm{X}(\mathrm{p} + \mathrm{d}_i + \Delta\mathrm{d}_i(\mathrm{p}))$ 正如同在 [13], [52] 之中。上面讨论的主要公式在 Fig. 2 中有所介绍。

B. 与其他方法的关联

2D 卷积是不包含几何信息的 S-Conv 的特例。具体来说，在没有几何信息的情况下，如果我们删除 $\mathrm{W}_i^*(\mathrm{p})$ 和 $\Delta\mathrm{d}_i(\mathrm{p})$，其由 Equ. (7) 中的几何信息生成，这个过程退化为 2D 卷积。而对于 RGBD 情况，我们的 S-Conv 可以通过引入空间自适应权重来在点级别提取特征并且不限于离散网格，如 Fig. 3所示。可变形卷积 [13], [14] 也通过生成不同的分布权重来缓解这个问题。然而，它们的分布是从 2D 特征图所推断出来的，而不是像我们是从 3D 空间信息推断出来的。我们将通过实验验证我们的方法比可变形卷积获得更好的结果 [13], [14]。与形变（SV）卷积 [57] 相比，SV 卷积通过基于语义相关区域的位置变体卷积来限制其上下文区域。它实现了一个位置变量卷积算子，其权重是位置变量，由特征图生

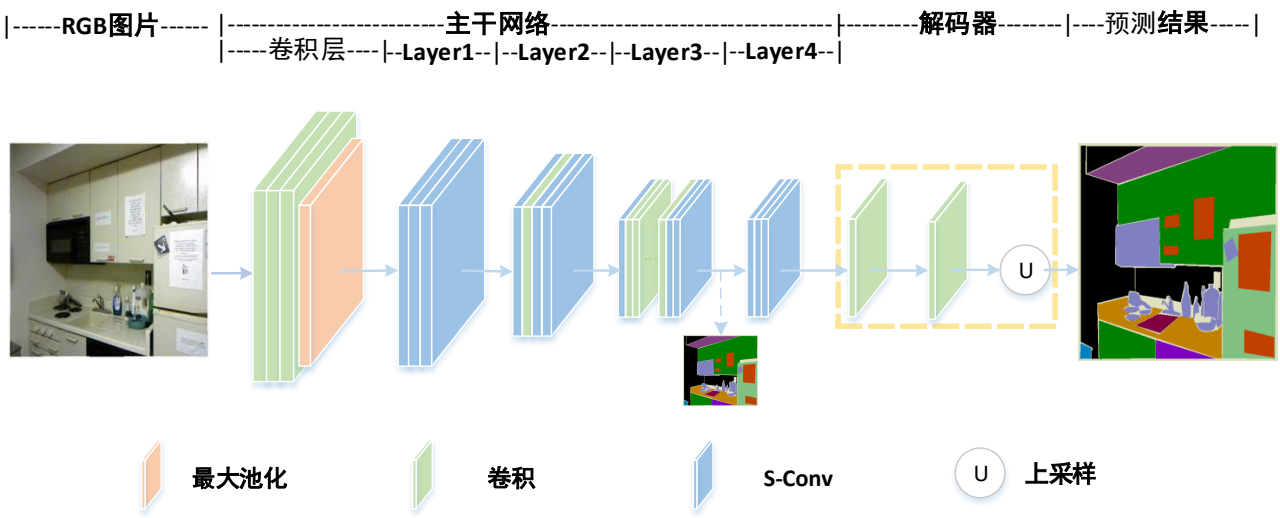


图 4. **配备 S-Conv 用于 RGBD 语义分割的 SGNet 网络架构**。SGNet 由主干网络和解码器组成。在第 3 层和第 4 层之间添加了深度监督（deep supervision）以改进网络优化。

成，侧重于上下文语义的理解。我们的 S-Conv 使用深度图而不是特征图来生成空间自适应偏移和权重。并且 S-Conv 的权重和偏移量从空间信息（深度图）得到的。这有助于卷积层根据空间信息调整感受野并适应几何变换。与基于 3D KNN 图的方法相比，我们的 S-Conv 自适应地选择相邻像素，而不是使用不灵活且计算成本高的 KNN 图。

C. SGNet 结构

我们的语义分割网络称为 SGNet，配备 S-Conv，由主干网络和解码器组成。SGNet 的结构如图 Fig. 4 所示。我们使用 ResNet101 [58] 作为我们的主干网络，并用我们的 S-Conv 替换每层的第一个和最后两个常规卷积（3×3 filter）。我们添加了一系列卷积来进一步提取特征，然后使用双线性上采样来生成最终的分割概率图，它对应于 SGNet 的解码器部分。Equ. (3) 中的 ϕ 对应三个 3×3 卷积层，i.e. Conv(3, 64) - Conv(64, 64) - Conv(64, 64) 与非线性激活函数。Equ. (4) 中的 η 和 Equ. (6) 中的 f 分别对应单个卷积层和两个全连接层。S-Conv 实现是从可变形卷积修改而来的 [13], [14]。我们在第 3 层和第 4 层之间添加深度监督以提高网络优化能力，这与 PSPNet [59] 相同。

IV. 实验

在本节中，我们首先通过分析其在不同层的使用情况来验证 S-Conv 的性能；进行消融实验/与其替代品进行比较；评估使用不同输入信息生成偏移量的结果；并测试推理速度。然后我们在 NYUDV2 和 SUNRGBD 数据集上比较了配备 S-Conv 的 SGNet 与其他 state-of-the-art 语义分割方法。最后，我们将每一层的深度自适应感受野和分割结果可视化，证明所提出的 S-Conv 可以很好地利用空间信息。

数据集和指标: 我们在公共数据集上评估 S-Conv 算子和 SGNet 分割方法的性能：

- NYUDv2 [17]：此数据集包括 1,449 张 RGB 图片和对应的深度图与分割标注图。与之前的方法 [60] 保持一致，本文使用 795 张图片用于训练，654 张图片用于测试。本文使用 40 类别的设置来进行实验。
- SUNRGBD [18], [19]：此数据集包括 10,335 张 RGB 图片和对应的深度图与分割标注图，有 37 个类别，其中 5,285 张图片用于训练，5,050 张图片用于测试。
- Cityscapes [61]: 此数据集为室外场景数据集。本文将这个数据集分为训练，测试和验证集，分别有 2,975, 500 和 1,525 张图片。

本文使用 3 个常见的评测指标来验证，包括精度 (Acc)，平均精度 (mAcc)，和平均交并比 (mIoU)。这三个指标定义如下：

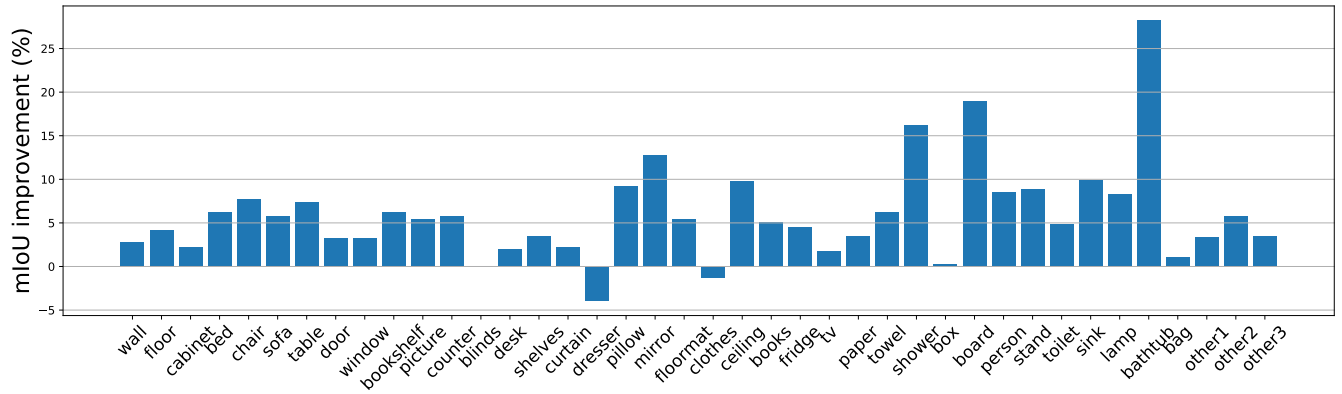


图 5. **引入 S-Conv 后，基础网络在 NYUDv2 测试集上每个类别的提升。**

表 I
在 NYUDv2 测试集上将卷积替换为 S-Conv 的结果。“layerx_y” 代表替换第 x 层的第 y 个残差网络的 3 × 3 卷积。

layer3_0	layer3_1	layer3_2	layer3_20	layer3_21	layer3_22	其他	平均交并比 (%)	参数量 (M)	FPS
							43.0	56.8	34
✓							47.0	56.9	34
✓	✓	✓					46.6	57.2	33
			✓	✓	✓		46.5	57.2	33
✓				✓	✓		47.8	57.2	33
✓				✓	✓	✓	49.0	58.3	26

表 II
S-Conv **替换网络不同层在 NYUDv2 测试集上的表现。结果为百分数。**
Acc: 准确率；mAcc: 平均准确率；mIoU: 平均交并比。

Layer1	Layer2	Layer3	Layer4	Acc	mAcc	mIoU
				72.1	54.6	43.0
✓				74.3	58.1	46.3
✓	✓			74.4	58.4	46.7
✓	✓	✓		75.2	60.3	48.5
✓	✓	✓	✓	75.5	60.9	49.0

$$
\begin{aligned}
Acc &= \sum_i \frac{p_{ii}}{g}, \\
mAcc &= \frac{1}{p_c} \sum_i \frac{p_{ii}}{g_i}, \\
mIoU &= \frac{1}{p_c} \sum_i \frac{p_{ii}}{g_i + \sum_j p_{ji} - p_{ii}},
\end{aligned}
\tag{8}
$$

其中 p_{ij} 是被预测为类 j 的像素数量，真实数据（ground truth）为 i，p_c 是类别数，g_i 是真实类别为 i 的像素数。$g = \sum_i g_i$ 是像素数。除非另有说明，否则深度图用作空间信息的默认格式。

实现细节：与 [4] 一致，本文使用在 ImageNet [20] 上预训练后的 ResNet101 [58] 作为本文特征提取的主干网络。默认的输出步长为 16。本文使用 Pytorch 实现整个系统。采用 SGD 优化器进行训练，其学习率策略（“poly” 策略）与 [4], [26] 相同，其中网络的初始学习率在消融实验上为 5e-3，在 NYUDv2 [17] 数据集上为 8e-3，在 SUNRGBD [19] 数据集上为 1e-3。权重衰减设置为 5e-4。网络默认使用 ReLU 激活函数，批大小默认为 8，同 [6]，网络使用常用的数据增广策略，包括随机尺度变换，随机剪裁，随机翻转。剪裁的大小为 480 × 640。在测试阶段，网络将图片下采样到训练裁剪的大小，然后预测的结果上采样到输入大小。网络使用交叉熵损失，由于 SUNRGBD 严重类别不平衡，本文依据类别的分布，对不同的类别添加不同的权重。本文使用两张 NVIDIA 1080Ti 显卡进行训练，在 NYUDv2 数据集上训练 500 轮，在 SUNRGBD 上训练 200 轮。

A. S-Conv 的分析

我们首先在 NYUDv2 [17] 数据集上进行消融实验。使用带有简单解码器和深度监督的 ResNet101 作为基础网络。

S-Conv **替代卷积实验**: 我们通过替换不同层中的传统卷积（3×3 过滤器）来评估 S-Conv 的有效性。我们首先在第 3 层替换卷积，然后将探索的规则扩展到其他层。FPS（每秒帧数）为在 NVIDIA 1080Ti 上进行测试的结果，输入图像大小为与 [16] 相同的 425×560 。最终结果参照 Tab. I.

我们可以从 Tab. I 中的结果得出以下两个结论。1) 基础网络的推理速度很快，但性能较差。用 S-Conv 代替卷积可以通过更多的参数和计算时间来改善基础网络的结果。2) 除了替换第 3 层第一个步长为 2 的卷积之外，替换后面的卷积的效果更好。主要原因是空间信息可以更好地指导第一次卷积中的下采样操作。因此我们选择用 S-Conv 替换每层的第一个卷积和最后两个卷积。我们将第 3 层中发现的规则推广到其他层并获得更好的结果。上面的实验表明，我们的 S-Conv 只需很少的参数就可以显着提高网络性能。值得注意的是，我们的网络没有空间信息流。空间信息只影响卷积核的分布和权重。我们还探索了嵌入不同层的 S-Conv 的性能。结果显示在 Tab. II 中。我们可以观察到性能随着配备 S-Conv 的层数而提高。

我们还在 Fig. 5 中展示了 S-Conv 在大多数类别上的 IoU 改进。很明显，我们的 S-Conv 在大多数类别中都提高了 IoU，尤其是对于缺乏代表性纹理信息的物体，例如镜子、木板和浴缸。对于具有丰富空间变换的对象，例如椅子和桌子，也有明显的改进。这表明我们的 S-Conv 在推理过程中可以很好地利用空间信息。

S-Conv **结构消融实验**: 为了评估我们提出的 S-Conv 中每个组件的有效性，我们设计了消融实验。结果显示在 Tab. III 中。默认情况下，我们根据 Tab. I替换每层的第一个卷积和最后两个卷积。我们可以看到，S-Conv 的偏移生成器、空间投影模块和权重生成器都有助于结果的改进。

与其他备选方案的比较: 大多数方法 [6], [9], [33], [62] 使用双流网络从两种不同的模态中提取特征，然后将它们组合起来。我们的 S-Conv 专注于利用空间信息提升网络的特征提取过程。在这里，我们将我们的 S-Conv 与双流网络、可变形卷积 [13], [14] 和深度感知卷积 [16] 进行比较。我们使用一个简单的基础网络，它由一个具有深度监督的 ResNet101 网络和一个简单的解码器组成。我们添加了一个额外的 ResNet101 网络，称为 HHANet，以提取 HHA 特征并将其与我们在双流网络最后一层的基础网络特征融合。为了与深度感知卷积

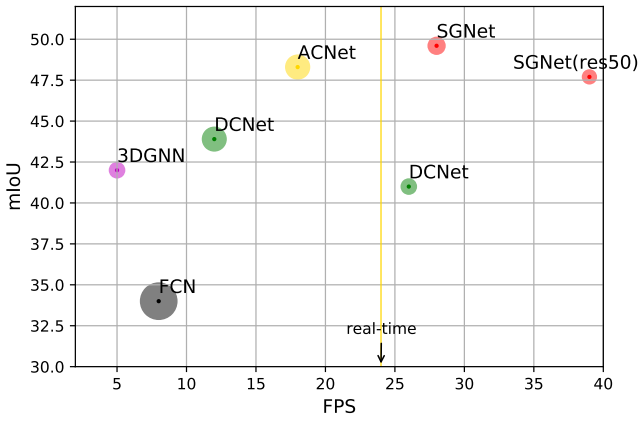


图 6. FPS、mIoU，**以及 NYUDv2 上不同方法的参数个数**。所有单尺度速度比较的输入图像大小为与 [16] 相同的 425×560 。圆的半径对应模型的参数个数。DCNet [16] 和 3DGNN [36] 的结果来自 [16]。我们的 SGNet 可以实现最快的推理时间和最先进的性能。

和可变形卷积进行比较，类似于 SGNet，我们替换了每层的第一个卷积和最后两个卷积。对于 “Baseline + DAC + DCV2”，我们在前两层用深度感知卷积替换卷积 [16] (DAC)，在最后两层用可变形卷积替换卷积 [13] (DCV2) ，因为 DCV2 不适用于较低层 [13]。结果显示在 Tab. IV 中。我们发现我们的 S-Conv 比双流网络、可变形卷积 [13]、深度感知卷积 [16] 以及它们的组合取得了更好的结果。这表明我们的 S-Conv 可以有效地利用空间信息。配备权重生成器的基础网络也可以取得比深度感知卷积更好的结果，说明从空间信息中学习权重是必要的。

空间信息对比: 我们还评估了不同格式的空间信息对 S-Conv 的影响。结果显示在 Tab. V 中。我们可以看到深度信息导致与 HHA 和 3D 坐标相当的结果，并且比可变形卷积 [13], [14] 使用的中间 RGB 特征更好。这显示了使用空间信息进行偏移和权重生成相对于 RGB 特征的优势。但是，将深度转换为 HHA 是很耗时的 [9]。因此 3D 坐标和深度图更适合使用 SGNet 进行实时分割。可以看出，即使没有空间信息输入（有 RGB 特征），我们的 S-Conv 也比基础网络有 3.4% 以上的提升。

推理速度测试: 为了证明 S-Conv 的轻量级，我们在这部分测试了 SGNet 的推理速度。我们还将我们的 S-Conv 与双流方法进行了比较。图像的输入大小为 480×640。结果显示在 Tab. VI 中。我们可以观察到，与双流方法相比，S-Conv 只需要少量的额外计算。我们的 SGNet 还可以使用 ResNet101 和 ResNet50 [58] 主干网络实现实时推理速度。

表 III
SGNet 在 NYUDv2 测试集上的消融实验。 OG: S-Conv 的偏移产生模块；WG: S-Conv 的权重产生模块；SP: S-Conv 的空间投影变换模块。

SP	OG	WG	准确率 (%)	平均准确率 (%)	平均交并比 (%)
			72.1	54.6	43.0
	✓		73.9	58.2	46.3
✓	✓		75.2	60.0	48.4
✓		✓	74.5	58.4	46.8
✓	✓	✓	75.5	60.9	49.0

表 IV
NYUDv2 测试集上的比较结果。 DCV2: 可变形卷积 [14]；DAC: 深度感知卷积 [16]；SP: S-Conv 中的空间投影模块；WG: S-Conv 中的权重产生模块。表中的结果为百分数。

模型	准确率	平均准确率	平均交并比
基础网络	72.1	54.6	43.0
基础网络 +DCV2	73.0	56.1	44.5
基础网络 +HHANet	73.5	56.8	45.4
基础网络 +DAC	73.8	57.1	45.4
基础网络 +HHANet+DCV2	74.3	58.4	47.0
基础网络 +DAC+DCV2	74.5	58.3	46.5
基础网络 +SP+WG	74.5	58.4	46.8
基础网络 +S-Conv(SGNet)	75.5	60.9	49.0

表 V
使用不同的空间信息在 NYUDv2 测试集上的比较结果。 Acc: 准确率；mAcc: 平均准确率；mIoU: 平均交并比。

空间信息	Acc(%)	mAcc(%)	mIoU(%)
深度图	75.5	60.9	49.0
RGB 特征	73.9	58.5	46.4
HHA	75.7	60.8	48.9
3D 空间坐标	75.3	61.2	48.5

表 VI
480 × 640 输入图片分辨率下的推理速度测试. OG: S-Conv 的偏移量产生模块, †：SGNet 中不应用产生的偏移量与权重, HHANet: 额外的主干网络 (ResNet101) 来处理空间信息。骨干网络 (ResNet101) 负责使用空间信息。

模型	时间 (秒)	帧率	参数量 (M)
基础网络	0.029	34	56.8
基础网络 +OG	0.033	30	57.7
基础网络 +HHANet	0.053	18	99.4
SGNet(ResNet50)	0.028	36	39.3
SGNet†	0.032	31	58.3
SGNet	0.037	26	58.3

B. 与其他主流方法的对比

我们在 NYUDv2 [17] 和 SUNRGBD [18], [19] 数据集上将我们的 SGNet 与其他 state-of-the-art 方法进行比较。SGNet 的架构显示在 Fig. 4 中。

NYUDv2 **数据集**: 比较结果可以在 Tab. VII 和 Fig. 6 中找到。我们将学习率从 5e-3 更改为 8e-3。我们将输入图像下采样到 480×640 并对其预测图进行上采样以获得测试期间的最终结果。为了将推理速度与其他方法进行比较，Tab. VII 中所有单尺度速度比较的输入图像大小为 425×560 以下 [16]。DCNet [16] 和 3DGNN [36] 的推理速度结果来自 [16]。我们使用 NVIDIA 1080Ti 在相同条件下测试了其他方法的单尺度速度。此外，输入大小为 480×640 的 SGNet 的推理速度测试显示在 Tab. VI 中。请注意，Tab. VII 中的某些方法不报告参数数量或开源。所以我们只是列出了这些方法的 mIoU。我们可以从 Tab. VI 和 Tab. VII 得出以下结论。我们的 SGNet (ResNet50) 不是使用额外的网络来提取空间特征，而是可以以最少的参数实现有竞争力的性能和

表 VII
NYUDv2 **测试集上的比较结果**。 MS: 多尺度测试; SI: 空间信息，Acc: 准确率 (%)，mIoU：平均交并比 (%)，param：参数量 (M)。输入测试图片的大小为 425 × 560，在 NVIDIA 1080Ti 环境下，速度均为单尺度速度。本文在 SGNet 的最后一层添加了 ASPP 模块 [4] ，命名为 “SGNet*”。

Network	Backbone	MS	SI	Acc	mAcc	mIoU	FPS	param (M)
FCN [3]	2×VGG16		HHA	65.4	46.1	34.0	8	272.2
LSD-GF [44]	2×VGG16		HHA	71.9	60.7	45.9	-	-
3DGNN [36]	VGG16		HHA	-	55.2	42.0	5	47.2
D-CNN [16]	VGG16		Depth	-	53.6	41.0	26	47.0
D-CNN [16]	2×ResNet152		Depth	-	61.1	48.4	-	-
ACNet [33]	2×ResNet50		Depth	-	-	48.3	18	116.6
RefineNet [25]	ResNet152	✓	-	73.6	58.9	46.5	16	129.5
RDFNet [6]	2×ResNet152	✓	HHA	76.0	62.8	50.1	9	200.1
RDFNet [6]	2×ResNet101	✓	HHA	75.6	62.2	49.1	11	169.1
CFNet [45]	2×ResNet152	✓	HHA	-	-	47.7	-	-
SGNet	ResNet50		Depth	75.0	59.6	47.7	39	39.3
SGNet	ResNet101		Depth	75.6	61.9	49.6	28	58.3
SGNet*	ResNet101		Depth	76.1	62.7	50.2	26	64.7
SGNet*	ResNet101	✓	Depth	76.8	63.3	51.1	26	64.7

最快的推理。我们的 SGNet (ResNet101) 可以实现更具竞争力的性能和实时推理。这得益于 S-Conv，它可以有效地利用空间信息，只需要少量的额外参数和计算成本。此外，我们的 S-Conv 可以在不使用 HHA 信息的情况下获得良好的结果，使其适用于实时任务。这验证了我们的 S-Conv 在利用空间信息方面的效率。通过在 SGNet 标记为 SGNet* 之后添加 ASPP 模块 [4] 以增加一点推理时间为代价，所提出的 SGNet 可以获得比其他方法和使用多尺度测试、HHA 信息和两个 ResNet152 主干网络的 RDFNet 更好的结果。使用其他方法使用的多尺度测试后，SGNet 的性能可以进一步提高。

SUNRGBD **数据集**: SUNRGBD 数据集上的比较结果显示在 Tab. VIII 中。值得注意的是，Tab. VII 中的一些方法没有报告关于 SUNRGBD 数据集的结果。Tab. VIII中模型的推理时间和参数数量与 Tab. VII中的相同。与没有实时性能的模型相比，我们的 SGNet 可以实时实现最先进的结果。

Cityscapes **数据集**: 我们在 SGNet 之后添加 ASPP [4] 模块并设置 *output stride* = 8. 我们在训练集上用 2975 张图像进行训练以进行验证。我们还在 Cityscapes 服务器上提供我们的测试结果。Cityscapes 数据集的比较结果显示在 Tab. IX 中。值得注意的是，由于 Cityscapes 中深度图的严重噪声，以前大多数基于 RGB-D 的方法的性能都比基于 RGB 的方法差。我们可以观察到，我们的网络受益于 S-Conv 可以取得比基础网络更好的结果，并在 Cityscapes 上取得有竞争力的结果。

C. 定量表现

S-Conv **中感受野的可视化**: 合适的感受野对于场景识别非常重要。我们在 S-Conv 生成的不同层中可视化 SGNet 的输入自适应感受野。具体来说，我们通过在 S-Conv 操作期间总结每个像素的偏移量的范数来获得每个像素的感受野，然后我们将每个值归一化为 [0, 255] 并使用灰度图像将结果可视化。结果显示在 Fig. 7 中。像素越亮，相对感受野越大。我们还使用圆的半径来表示相对感受野的大小。我们观察到不同卷积的感受野随着输入图像的深度自适应变化。例如，在 layer1_1 中，感受野与深度成反比。在每一层学习的自适应感受野的组合可以帮助网络更好地解析具有复杂空间关系的室内场景。

定性比较结果: 我们在 Fig. 8 中展示了 NYUDv2 测试数据集的定性比较结果。对于 Fig. 8(a) 中的视觉结果，浴缸和墙壁的纹理不足，无法通过基础方法轻松区分。一些物体可能有反射，例如 Fig. 8(b) 中的桌子，这对基础网络也具有挑战性。然而，SGNet 可以通过在 S-Conv

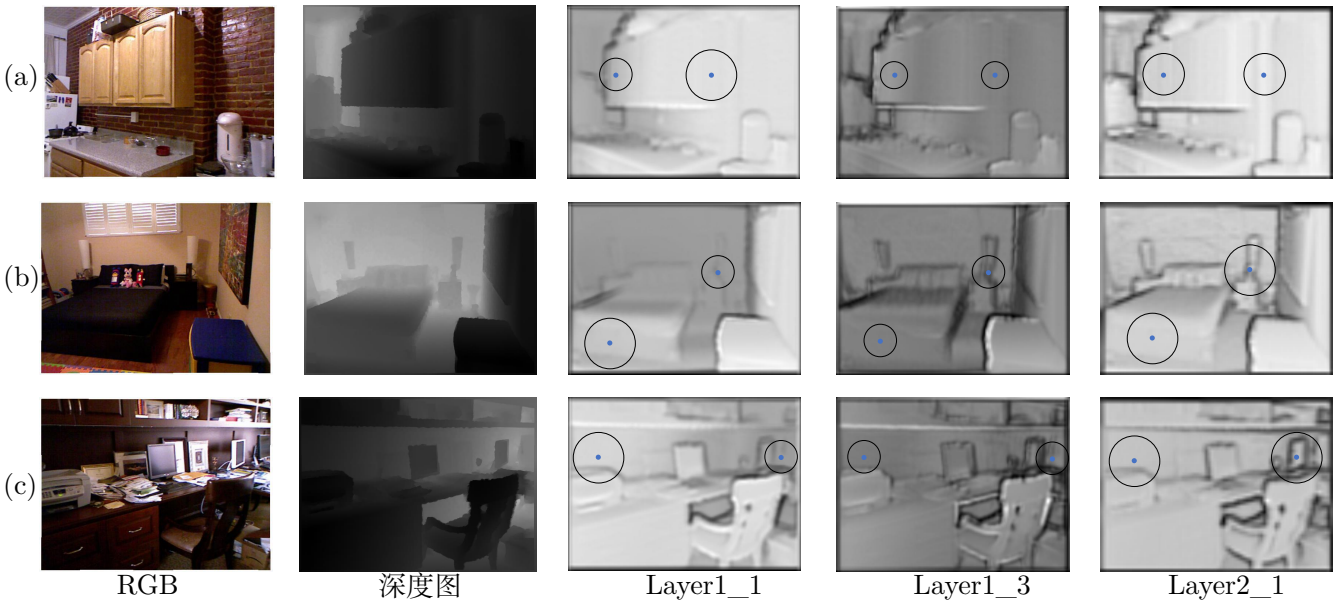


图 7. S-Conv **中相对感受野的可视化**。

表 VIII
SUNRGBD **测试数据集上的比较结果**。 MS: 多尺度测试, SI: 空间信息我们在 SGNet 的最后一层之后添加了 ASPP 模块 [4]，记为 “SGNet*”。

Network	Backbone	MS	SI	Acc	mAcc	mIoU	param (M)
LSD-GF [44]	2×VGG16		HHA	-	58.0	-	-
RefineNet [25]	ResNet152	✓	-	80.6	58.5	45.9	129.5
CGBNet [30]	ResNet101		-	82.3	61.3	48.2	-
3DGNN [36]	VGG16	✓	HHA	-	57.0	45.9	47.2
D-CNN [16]	2×VGG16		HHA	-	53.5	42.0	92.0
ACNet [33]	2×ResNet50		HHA	-	-	48.1	272.2
RDFNet [6]	2×ResNet152	✓	HHA	81.5	60.1	47.7	200.1
CFNet [45]	2×ResNet152	✓	HHA	-	-	48.1	-
SGNet	ResNet101		Depth	81.0	59.6	47.1	58.3
SGNet*	ResNet101		Depth	81.0	59.8	47.5	64.7
SGNet*	ResNet101	✓	Depth	82.0	60.7	48.6	64.7

表 IX
Cityscapes **验证数据集的比较结果**。‡: **测试数据集的结果**。 iterations: 迭代次数、MS: 多尺度、mIoU: 平均交并比

Network	Backbone	iterations	MS	mIoU
Baseline	ResNet101	40k		78.2
SGNet	ResNet101	40k		79.2
SGNet	ResNet101	65k	✓	80.6
SGNet	ResNet101	65k	✓	81.2‡

的帮助下合并空间信息来很好地识别它。Fig. 8(c, d) 中的椅子由于对比度低和纹理混乱而难以被 RGB 数据识别，而受益于配备的 S-Conv，SGNet 可以轻松辨别它们。同时，SGNet 可以很好地恢复对象的几何形状，正如 Fig. 8(e) 的椅子所示。我们还在 Fig. 9 中展示了 SUNRGBD 测试数据集的定性结果。可以看出，我们的 SGNet 也可以在 SUNRGBD 上实现精确分割。

V. 总结

在本文中，我们提出了一种新颖的 空间信息引导卷积 (S-Conv) 算子。与传统的 2D 卷积相比，它可以根据

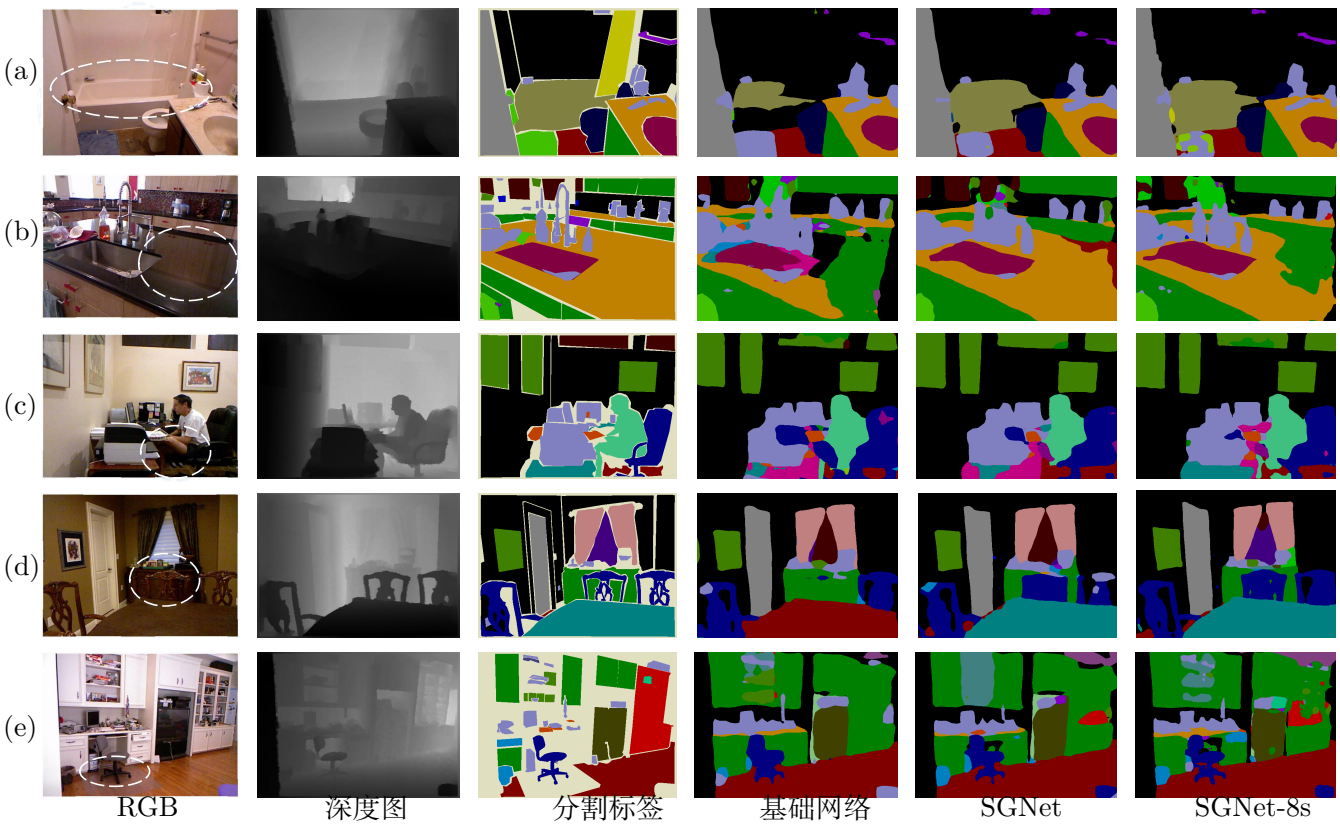


图 8. NYUDv2 **测试数据集上的定性语义分割比较结果**。 SGNet-8s: 输出步长为 8。

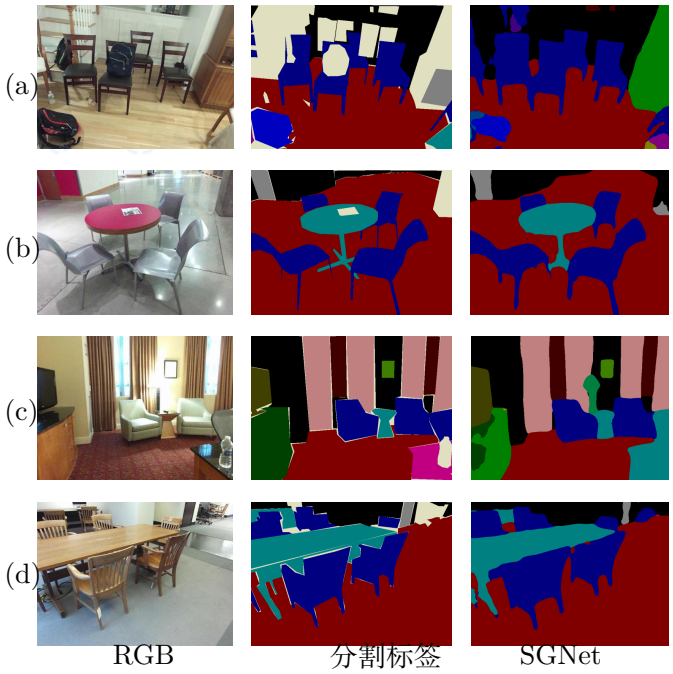


图 9. SUNRGBD **测试数据集上的定性语义分割比较结果**。

输入的空间信息自适应地调整卷积权重和分布，从而只需要很少的附加参数和计算成本就可以更好地感知几何结构。我们还提出了空间信息引导卷积网络 (SGNet) 配备了 S-Conv，具有实时推理速度并在 NYUDv2 和 SUNRGBD 数据集上获得具有竞争力的 RGBD 语义分割结果。我们还比较了使用不同输入来生成偏移的性能，展示了使用空间信息优于 RGB 特征的优势。此外，我们将每层中的深度自适应感受野可视化以显示有效性。未来，我们将同时研究不同模态信息的融合和 S-Conv 结构的自适应变化，使这两种方法相互受益。我们还将探索 S-Conv 在不同领域的应用，例如姿势估计和 3D 对象检测。

致谢

该研究得到了国家自然科学基金（61620106008）、天津市自然科学基金（17JCJQJC43700）和教育部科技创新项目的资助，新一代人工智能重大专项: 2018AAA0100400。

参考文献

[1] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," in Eur. Conf. Comput. Vis., 2018, pp. 405–420.

[2] B. Bescos, J. M. Fácil, J. Civera, and J. Neira, "Dynaslam: Tracking, mapping, and inpainting in dynamic scenes," IEEE Robotics and Automation Letters, vol. 3, no. 4, pp. 4076–4083, 2018.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2015, pp. 3431–3440.

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834–848, 2017.

[5] Q. Hou, L. Han, and M.-M. Cheng, "Autonomous learning of semantic segmentation from internet images (in chinese)," Sci Sin Inform, 2021.

[6] S.-J. Park, K.-S. Hong, and S. Lee, "Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation," in Int. Conf. Comput. Vis., 2017, pp. 4980–4989.

[7] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in Int. Conf. Comput. Vis., 2015, pp. 2650–2658.

[8] L. Ma, J. Stückler, C. Kerl, and D. Cremers, "Multi-view deep learning for consistent semantic mapping with rgb-d cameras," in IROS, 2017, pp. 598–605.

[9] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in Asian Conf. Comput. Vis., 2016, pp. 213–228.

[10] J. Wang, Z. Wang, D. Tao, S. See, and G. Wang, "Learning common and specific features for rgb-d semantic segmentation with deconvolutional networks," in Eur. Conf. Comput. Vis. Springer, 2016, pp. 664–679.

[11] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in Eur. Conf. Comput. Vis. Springer, 2014, pp. 345–360.

[12] Z. Li, Y. Gan, X. Liang, Y. Yu, H. Cheng, and L. Lin, "Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling," in Eur. Conf. Comput. Vis. Springer, 2016, pp. 541–557.

[13] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in Int. Conf. Comput. Vis., 2017, pp. 764–773.

[14] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in IEEE Conf. Comput. Vis. Pattern Recog., 2019, pp. 9308–9316.

[15] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," Computational Visual Media, vol. 5, no. 2, pp. 117–150, 2019.

[16] W. Wang and U. Neumann, "Depth-aware cnn for rgb-d segmentation," in Eur. Conf. Comput. Vis., 2018, pp. 135–150.

[17] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in Eur. Conf. Comput. Vis. Springer, 2012, pp. 746–760.

[18] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in IEEE Conf. Comput. Vis. Pattern Recog., 2015, pp. 567–576.

[19] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell, "A category-level 3d object dataset: Putting the kinect to work," in Consumer depth cameras for computer vision, 2013, pp. 141–165.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Adv. Neural Inform. Process. Syst., 2012, pp. 1097–1105.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Int. Conf. Learn. Represent., 2015.

[22] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in Int. Conf. Learn. Represent., 2016.

[23] H. Ding, X. Jiang, A. Q. Liu, N. M. Thalmann, and G. Wang, "Boundary-aware feature propagation for scene segmentation," in Int. Conf. Comput. Vis., 2019, pp. 6819–6829.

[24] B. Shuai, H. Ding, T. Liu, G. Wang, and X. Jiang, "Toward achieving robust low-level and high-level scene parsing," IEEE Trans. Image Process., vol. 28, no. 3, pp. 1378–1390, 2018.

[25] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 1925–1934.

[26] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in Eur. Conf. Comput. Vis., 2018, pp. 801–818.

[27] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," vol. 39, no. 12, pp. 2481–2495, 2017.

[28] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in Int. Conf. Comput. Vis., 2015, pp. 1520–1528.

[29] H. Ding, X. Jiang, B. Shuai, A. Qun Liu, and G. Wang, "Context contrasted feature and gated multi-scale aggregation for scene segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 2393–2402.

[30] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Semantic segmentation with context encoding and multi-path decoding," IEEE Trans. Image Process., vol. 29, pp. 3520–3533, 2020.

[31] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 3684–3692.

[32] R. Fan, M.-M. Cheng, Q. Hou, T.-J. Mu, J. Wang, and S.-M. Hu, "S4net: Single stage salient-instance segmentation," Computational Visual Media, vol. 6, no. 2, pp. 191–204, June 2020.

[33] X. Hu, K. Yang, L. Fei, and K. Wang, "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation," in IEEE Int. Conf. Image Process. IEEE, 2019, pp. 1440–1444.

[34] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 1746–1754.

[35] S. Song and J. Xiao, "Deep sliding shapes for amodal 3d object detection in rgb-d images," in IEEE Conf. Comput. Vis. Pattern Recog., 2016, pp. 808–816.

[36] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, "3d graph neural networks for rgbd semantic segmentation," in Int. Conf. Comput. Vis., 2017, pp. 5199–5208.

[37] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 652–660.

[38] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in Adv. Neural Inform. Process. Syst., 2017, pp. 5099–5108.

[39] L.-Z. Chen, X.-Y. Li, D.-P. Fan, M.-M. Cheng, K. Wang, and S.-P. Lu, "Lsanet: Feature learning on point sets by local spatial aware layer," arXiv preprint arXiv:1905.05442, 2019.

[40] Y. Xu, T. Fan, M. Xu, L. Zeng, and Y. Qiao, "Spidercnn: Deep learning on point sets with parameterized convolutional filters," in Eur. Conf. Comput. Vis., 2018, pp. 87–102.

[41] C. Wang, B. Samari, and K. Siddiqi, "Local spectral graph convolution for point set feature learning," in Eur. Conf. Comput. Vis., 2018, pp. 52–66.

[42] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, "Pointcnn: Convolution on x-transformed points," in Adv. Neural Inform. Process. Syst., 2018, pp. 828–838.

[43] M. Ding, Y. Huo, H. Yi, Z. Wang, J. Shi, Z. Lu, and P. Luo, "Learning depth-guided convolutions for monocular 3d object detection," in IEEE Conf. Comput. Vis. Pattern Recog. Worksh., 2020, pp. 1000–1001.

[44] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, "Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 3029–3037.

[45] D. Lin, G. Chen, D. Cohen-Or, P.-A. Heng, and H. Huang, "Cascaded feature network for semantic segmentation of rgb-d images," in Int. Conf. Comput. Vis., Oct 2017.

[46] J. Jiao, Y. Wei, Z. Jie, H. Shi, R. W. Lau, and T. S. Huang, "Geometry-aware distillation for indoor semantic segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2019, pp. 2869–2878.

[47] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille, "Towards unified depth and semantic prediction from a single image," in IEEE Conf. Comput. Vis. Pattern Recog., 2015, pp. 2800–2809.

[48] J. Hoffman, S. Gupta, and T. Darrell, "Learning with side information through modality hallucination," in IEEE Conf. Comput. Vis. Pattern Recog., 2016, pp. 826–834.

[49] I. Kokkinos, "Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory," in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 6129–6138.

[50] Z. Zhang, Z. Cui, C. Xu, Y. Yan, N. Sebe, and J. Yang, "Pattern-affinitive propagation across depth, surface normal and semantic segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2019.

[51] Y. He, W.-C. Chiu, M. Keuper, and M. Fritz, "Std2p: Rgbd semantic segmentation using spatio-temporal data-driven pooling," in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 4837–4846.

[52] M. Jaderberg, K. Simonyan, A. Zisserman et al., "Spatial transformer networks," in Adv. Neural Inform. Process. Syst., 2015, pp. 2017–2025.

[53] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," in Adv. Neural Inform. Process. Syst., 2016, pp. 667–675.

[54] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in IEEE Conf. Comput. Vis. Pattern Recog., 2019, pp. 510–519.

[55] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 7794–7803.

[56] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in IEEE Conf. Comput. Vis. Pattern Recog., June 2018.

[57] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Semantic correlation promoted shape-variant context for segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., 2019, pp. 8885–8894.

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in IEEE Conf. Comput. Vis. Pattern Recog., 2016, pp. 770–778.

[59] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 2881–2890.

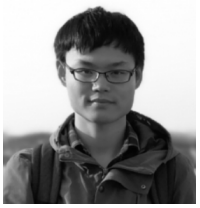
[60] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in Eur. Conf. Comput. Vis., 2012, pp. 746–760.

[61] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in IEEE Conf. Comput. Vis. Pattern Recog., 2016.

[62] J. Jiang, L. Zheng, F. Luo, and Z. Zhang, "Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation," arXiv preprint arXiv:1806.01054, 2018.

Lin-Zhuo Chen is currently a Master student with College of Computer Science, Nankai University. His research interests include computer vision and deep learning.

Zheng Lin is currently a Ph.D. candidate with College of Computer Science, Nankai University, under the supervision of Prof. Ming-Ming Cheng. His research interests include deep learning, computer graphics and computer vision.



Ziqin Wang Received the Bachelor Degree in Information Engineering (2011-2015), and received the Master Degree in Control Engineering (2015-2018) form Xian Jiaotong University. Currently he is a PhD candidate at the University of Sydney. His research interests include computer vision, deep learning, image segmentation and pattern recognition.

Yong-Liang Yang is a Senior Lecturer in Department of Computer Science, University of Bath. His research interests are broadly in visual computing and interactive techniques. His work has led to more than 40 publications in major journals and conferences, including SIGGRAPH, SIGGRAPH Asia, CHI, UIST, NeurIPS, and ICCV. He has served on program committees of multiple major conferences including Symposium on Geometry Processing, Pacific Graphics, and Solid Physical Modeling. He is a recipient of the Computer Aided Geometric Design Most Cited Paper Award in 2011 and 2012. His work has also been selected as research highlights by the Communications of the ACM.

Ming-Ming Cheng received his PhD degree from Tsinghua University in 2012. Then he did 2 years research fellow, with Prof. Philip Torr in Oxford. He is now a professor at Nankai University, leading the Media Computing Lab. His research interests includes computer graphics, computer vision, and image processing. He has published 60+ refereed research papers, with $15,000+$ Google Scholar citations. He received research awards including ACM China Rising Star Award, IBM Global SUR Award, etc. He is a senior member of IEEE and on the editor board of IEEE TIP.