PointGroup: Dual-Set Point Grouping for 3D Instance Segmentation

摘要

PointGroup 是一种新的端到端自下而上架构，特别关注通过探索对象之间的空白空间来更好地对点进行分组。PointGroup 设计了两个分支的网络来提取点的特征来预测语义分割和偏移量，从而将每个点移向其实例质心，使用聚类组件利用原始点坐标集和偏移后的点坐标集，充分利用它们的互补优势。接下来，他们使用 ScoreNet 评估候选实例，并使用非最大值抑制（NMS）去除重复预测，最终实现了较为准确的实例分割任务。在这篇研究报告中实现了 PointGroup 方法的完整复现，并且在原方法的基础上实现了数据增强技术。通过将多个场景进行融合来形成训练集对模型进行训练，提高了模型的泛化能力。通过增强的数据集对模型进行训练，最终结果表示混合后的数据可以增强模型在较小物体上的实例分割性能。

关键词： 3D 实例分割；数据增强；点云；ScanNet；深度学习；卷积神经网络；点聚类；

1 引言

复现的论文是关于 3D 实例分割的研究，它提出了一种名为 PointGroup 的底层自下而上的架构，旨在通过探索对象之间的空白空间来更好地对点进行分组。论文指出，相比于已经发展成熟的 2D 实例分割，针对点云的 3D 实例分割仍有很大的改进空间。因此，作者设计了一种双分支网络，用于提取点特征并预测语义标签和偏移量，以将每个点移动到其相应的实例质心。论文还介绍了一种聚类组件，利用原始坐标和偏移后的坐标集。此外，作者提出了 ScoreNet 来评估候选实例，并使用非极大值抑制（NMS）来消除重复。论文在两个具有挑战性的数据集 ScanNet v2 和 S3DIS 上进行了广泛实验，结果显示该方法在 mAP（交并比阈值为 0.5）方面取得了最高性能，分别为 63.6% 和 64.0%，而之前最好的解决方案仅为 54.9% 和 54.4%。

根据论文中的介绍，实例分割是一项重要且具有挑战性的任务，它要求对场景中的每个对象进行语义标签和实例 ID 的预测。随着自动驾驶、机器人导航等领域的发展，实例分割在室外和室内环境中都具有潜在的应用前景。然而，由于 3D 点云的无序和非结构特性，传统的 2D 方法无法直接扩展到 3D 点云，因此 3D 实例分割仍然是一个具有挑战性的问题。因此，本文提出了 PointGroup 方法，旨在通过探索对象之间的空白空间以及语义信息，更好地实现对个体对象的分割。

这篇论文的研究对于改进 3D 实例分割具有重要意义。通过提出 PointGroup 架构和双分支网络，论文在 3D 点云实例分割任务中取得了显著的性能提升。该方法能够更好地将点分

组成对象，并区分相邻对象的空间关系。这对于实现精确的实例分割和场景理解具有重要作用，并为自动驾驶、机器人导航等领域的应用提供了有力支持。此外，通过在具有挑战性的数据集上进行广泛实验，并与现有最佳解决方案进行比较，论文证明了 PointGroup 方法的有效性和通用性，为进一步推动 3D 实例分割的研究和应用提供了有益的参考。

对物体进行 3D 实例分割在实际的应用中具有重要的应用，在与物联网有关的应用中同时具有重要的意义。在我们的日常生活中存在着大量的物联网设备如 Wifi、蓝牙、zigbee 和 LoRa 等，这些设备在实际的部署过程中需要考虑各种因素如设备旁的物体遮挡情况，设备之间的距离等，这些因素都会对设备终端在实际的覆盖情况具有重要的影响。3D 实例分割的应用可以帮助我们对设备部署中的实例环境进行分割任务的处理，帮助我们进行设备部署后其覆盖范围的评估。具体而言，我们可以通过首先应用 3D 实例分割来实现对建筑物中的每个设备进行分割，为不同的设备分配各自的实例 ID。然后我们利用 3D 实例分割后的结果来评估设备的大致覆盖范围，我们科技将设备旁的环境信息分成两大类，一类是对设备的信号传输产生较小影响的环境类型即直视路径，一类是对设备的信息传输产生较大影响的物体即非直视路径。

2 相关工作

这部分将对论文中已经存在的相关工作的研究进行概括与描述，讨论目前已有的方法，并且讨论 PointGroup 方法的优势。

2.1 2D 实例分割

2D 实例分割旨在在场景中找到前景对象，并为每个对象实例标记一个唯一的标签。总体而言，有两个主要方向。第一个是基于检测或自上而下的方法，直接检测对象实例。早期的工作 [7] 使用 MCG [17] 的提议进行特征提取。 [5] 的方法采用池化特征以加快处理速度。Mask R-CNN [8] 被广泛认为是一种在检测框架中具有额外分割头的有效方法，类似于 Faster R-CNN [18]。进一步的工作 [11] 增强了实例分割的特征学习能力。

另一种方法是基于分割或自下而上的方法，首先进行像素级语义分割，然后对像素进行分组以找到对象实例。张等人 [24] 利用 MRF 进行局部块合并。Arnab 和 Torr [1] 使用 CRF。Bai 和 Urtasun [2] 结合经典的分水岭变换和深度学习，产生能量图以区分个体实例。刘等人 [15] 使用一系列神经网络从像素构建对象。

2.2 3D 场景中的深度学习

2D 图像像素位于规则的网格中，因此可以通过卷积神经网络 [9] 进行自然处理。相比之下，3D 点云是无序的，散布在 3D 空间中，导致在点云场景理解方面存在额外的困难 [19]。有几种方法处理数据的不规则性。多层感知器（MLP）风格的网络，例如 PointNet [3]，直接应用 MLP 和最大池化来捕获 3D 中的局部和全局结构。然后使用学习到的特征进行点云分类和分割。

除了直接处理不规则输入外，还有几种方法将无序点集转换为有序点集以应用卷积操作。PointCNN [12] 学习点的顺序变换以进行点的重新加权和排列。其他一些方法 [4] 对齐和体素

化点云，以生成用于 3D 卷积的规则 3D 有序张量。多视角策略 [20] 也被广泛探索，其中 3D 点云被投影到 2D 视图中进行视域处理。

2.3 3D 实例分割

3D 实例分割也可以被分为两种情况，其中第一种情况是基于候选区域的 3D 实例分割方法，这种方法通过在 3D 场景中首先选出具有代表性的实例区域以对接下来的实例分割方法提出建议，方便进行实例分割。第二种方法是非基于候选区域的 3D 实例分割方法，这种方法不进行候选区域的预选，而是直接对 3D 点云数据进行处理进行实例分割，这种方法可以免除进行候选时引入的误差。

基于检测的方法提取 3D 边界框，在每个框内利用掩码学习分支预测对象掩码。Li 等人 [23] 提出了 GSPN，采用分析合成策略为实例分割生成提议。Hou 等人 [10] 将多视图 RGB 输入与 3D 几何结合起来，以端到端的方式联合推断对象边界框和相应的实例掩码。3D-BotNet [22] 直接回归用于目标候选的 3D 边界框，采用多标准损失。3D-MPA [6] 从偏移的质心中采样提议，并使用图神经网络 [Wang 等，2019b] 增强特征。GICN [14] 将实例的质心视为高斯分布。

基于语义分割的方法。Wang 等人 [21] 通过基于语义分割预测（如 PointNet++）对点进行聚类来设计 SGPN。Liu 和 Furukawa [13] 预测不同尺度下相邻体素之间的语义标签和关联性，以分组实例。Pham 等人 [16] 开发了一个多任务学习框架，使用多值 CRF 模型来共同推理语义和实例标签。

3 本文方法

3.1 本文方法概述

本文方法的主要概述如图 2所示，PointGroup 方法使用 U-Net 作为骨干网络，如图所示，网络的输入是 3D 点云数据，其中每个点都有点的 3 个 RGB 特征，代表着点的颜色，同时也具有点的空间坐标特征信息。其中 $f_i = \{r_i, g_i, b_i\}$ 代表了点的颜色，$p_i = \{x_i, y_i, z_i\}$ 代表了点的位置。然后使用骨干网络对点云信息进行特征的提取，将 U-Net 作为骨干网络，由于 U-Net 网络的特殊对称的结构，是的其在图像处理中具有非常好的效果。骨干网络也可以被其他网络替代，骨干网络提取的特征可以表示为 $F = F_i\ R^{N \times K}$ 其中 k 是特征的通道数。

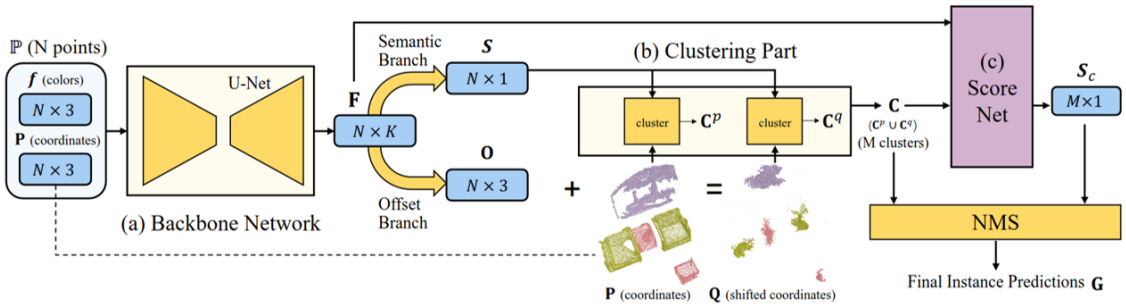

图 1. 方法概述

接着会进行两个分支的步骤，一个是用于语义分割的语义分支，另一个是用于对点进行

偏移的偏移分支。其中语义分支的主要任务是对提取得到的特征值进行语义分割任务，其结果是已经进行分割任务的点云数据，目的是为实例分割任务提供有利的特征数据。偏移分支的任务是通过探索物体与物体之间的空白信息，其认为物体之间存在的空白间隙可以更好的帮助算法实现实例分割任务，对于相同的实例的点云数据，其会尝试将点云移动到实例质心上面，以此完成点的聚类任务。偏移分支不仅将点的偏移结果作为下一阶段的输入，而且还输入了原始点数据。并且将这些点数据与语义分支结果组成簇，将两种点坐标的簇取并集作为输入。同时论文还使用使用 NMS 对其进行最终预测，并且提出了一个 ScoreNet 评分机制为最终的结果进行打分，以此来实现最好的实例分割结果。

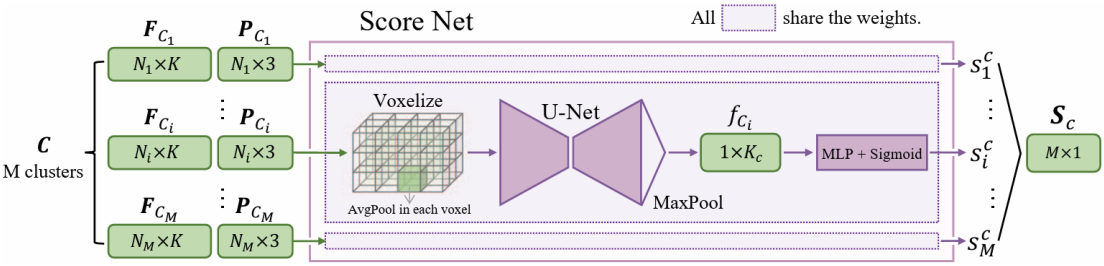

图 2. ScoreNet

ScoreNet 的输入是所有候选聚类 C，输出是每一个反应聚类质量的得分。首先将特征 F 和聚类 Ci 进行体素化，再利用小型 U-Net 提取特征，再经过 max pooling 得到单个聚类特征向量，最后经过 MLP+sigmoid 得到聚类得分。

3.2 特征提取模块

在特征提取模块，使用了一个 U-Net 作为骨干网络进行特征的提取如图 2所示。U-Net 网络作为特征提取的骨干网络具有很多优点。首先，它能够捕捉丰富的上下文信息，通过编码器-解码器结构有效地利用输入数据的整体信息。其次，U-Net 网络可以将低级特征和高级特征进行跨层连接和重建，以保留细节并提高特征的表达能力。此外，U-Net 网络具有大小可调性，可以根据任务需求灵活地调整架构。最后，由于其使用反卷积和 Skip 连接来进行特征重建，U-Net 网络更易于训练，并能够更快地收敛到优秀的结果。

3.3 损失函数定义

偏移分支将 **F** 编码为 N 个偏移向量 $\mathbf{O} = \{o_1, \ldots, o_N\} \in \mathbb{R}^{N\times 3}$ 用于表示这 N 个点的偏移量。对于属于同一实例的点，通过 L_1 回归损失来规范它们学习到的偏移量：

$$L_{o\_reg} = \frac{1}{\sum_i m_i} \sum_i \|o_i - (\hat{c}_i - p_i)\| \cdot m_i$$

其中 $\mathbf{m} = \{m_1, \ldots, m_N\}$ 是一个二元值，表示点 i 是否为某个实例上的点，ci 表示点 i 所属实例的中心。

$$\hat{c}_i = \frac{1}{N^I_{g(i)}} \sum_{j \in I_{g(i)}} p_j$$

考虑到不同类别的不同物体大小，网络很难精确的回归偏移，特别是对于大型物体的边界点，因为这些点离实例质心相对较远。因此制定了一个方向损失来约束预测偏移向量的方向，将

损失定义为减去余弦相似度的方式。这种损失与偏移向量的范数无关，并确保点向它们所属的实例中心点移动。

$$L_{o\_dir} = -\frac{1}{\sum_i m_i} \sum_i \frac{o_i}{\|o_i\|_2} \cdot \frac{\hat{c}_i - p_i}{\|\hat{c}_i - p_i\|_2} \cdot m_i$$

为了反映聚类质量，使用软标签来替代二进制的 0/1 标签，用于监督预测的聚类分数:

$$\hat{s}_i^c = \begin{cases} 0 & \text{iou }_i < \theta_l \\ 1 & \text{iou }_i > \theta_h \\ \frac{1}{\theta_h - \theta_l} \cdot (\text{ iou }_i - \theta_l) & \text{otherwise} \end{cases}$$

其中 ioui 是聚类 Ci 与真实实例之间的最大并集交集（IoU）。然后使用二进制交叉熵损失作为分数损失

$$L_{\text{c\_score}} = -\frac{1}{M} \sum_{i=1}^{M} (\hat{s}_i^c \log(s_i^c) + (1 - \hat{s}_i^c) \log(1 - s_i^c))$$

在训练中使用的损失函数是将上面的损失函数进行直接相加，即:

$$L = L_{\text{sem}} + L_{o\_\text{dir}} + L_{\text{o\_reg}} + L_{c\_\text{score}}$$

4 复现细节

4.1 与已有开源代码对比

在论文的复现过程中使用了数据增强技术，如图 3所示，目前的大部分 3D 分割模型都可以能够充分捕捉输入的 3D 场景的全局纹理。模型在训练的过程中使用的是全局的纹理信息，这就意味着在模型的训练过程中，模型在对实例分割进行预测时会尝试使用环境信息，比如说模型会特别希望会议室中会出现一把椅子，或者在厕所中会出现一个马桶。环境信息与物体之间的关联虽然一定程度上帮助模型产生较为准确的结果，但是在一定程度上导致了模型的过拟合现象。但是在实际的应用中，一些较为罕见的场景也可能会对最终的决策产生较大的影响。因此使用数据增强技术对输入的数据进行旋转，镜像和混合等过程，可以帮助我们减少过拟合现象，帮助我们更好的进行全局环境信息和局部特征信息之间的平衡，在图 3中使用了将两个场景进行混合的技术。

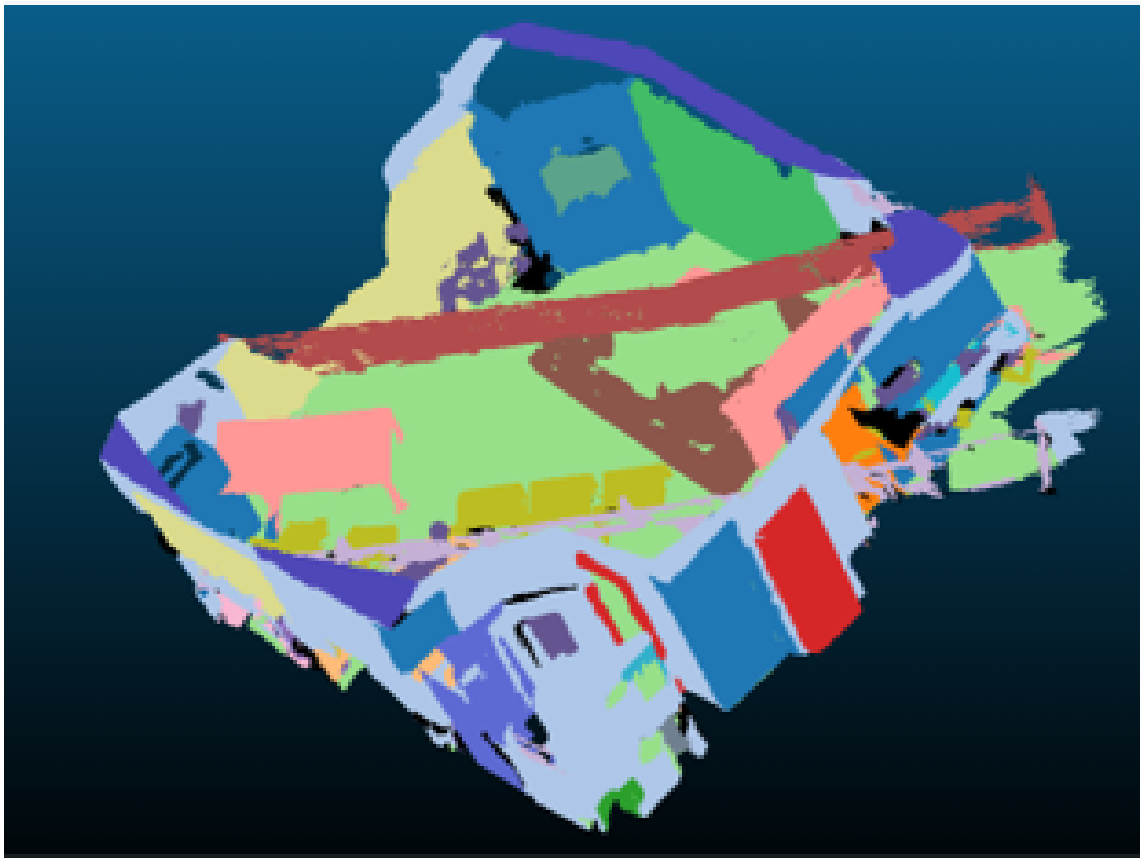

图 3. 数据混合

4.2 实验环境搭建

实验环境使用了 ubuntu20.04 和 cuda11.3 版本，其中 pytorch==1.3。硬件条件 GPU 为 RTX A4000。训练的过程中的 Loss 值如下图所示。

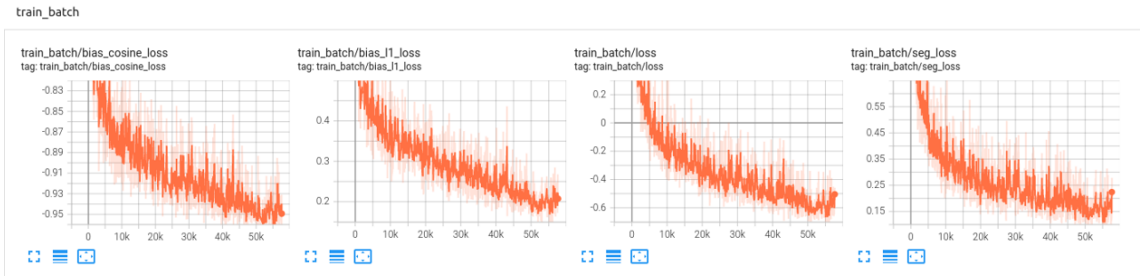

图 4. Loss

4.3 创新点

原始的数据集是在真实的现实场景中取得，这种场景信息会使得模型在学习的过程中将全局环境信息与局部信息相关联，因此很容易产生过拟合现象。本次复现实现数据增强技术，

5 实验结果分析

本次实验的结果如图所示，在于原文中的结果进行比较中可以发现，进行数据混合后的结果在诸如椅子，图片等这种较小的物体上取得了比原文较好的分割结果。因此我们可以的出结论，进行不同场景的混合后的数据可以在一定程度上减少模型的过拟合现象。同样的，不可避免的存在着一些结果变得很差，这是由于使用的场景的融合是过程是一个随机的过程，这种随机导致了一些完整结构场景之间的破坏比如两个椅子之间的相互重叠现象。其中 AP50 表示 IoU 阈值设为 50% 时的 AP 得分。换句话说，对于一个检测结果的判定，只有当检测框与真实标注框的 IoU 大于等于 50% 时，才会被认为是正确的检测结果。而 AP50 则是在考虑了所有 IoU 大于等于 50% 的检测结果后，计算得到的平均精确率。其中椅子这一数据平均 AP50 达到了 0.991，相对于原文提升了 0.194。椅子在我们的日常生活中比较常见，因此在数据集中出现的次数也会比较多，因此在椅子数据的训练过程中很容易导致过拟合现象。使用数据增强技术可以减少这种过拟合现象。

Method	Avg AP50	bathtub	bed	bookshelf	cabinet	chair	counter	curtain	desk	door	otherfur.	picture	refridge.	s.curtain	sink	sofa	toilet	window
paper	0.636	1.00	0.765	0.624	0.505	0.797	0.116	0.696	0.384	0.441	0.559	0.476	0.596	1.000	0.666	0.756	0.997	0.513
my	0.608	0.870	0.734	0.520	0.503	**0.911**	**0.189**	0.550	**0.410**	**0.491**	0.584	**0.555**	**0.634**	0.681	0.699	0.583	0.996	0.384

图 5. 实验结果图

6 总结与展望

在本次复现研究中实现了 PointGroup 论文方法的复现，并且通过实现数据增强技术有效的提高了模型在较小物体上的预测精度。文章中使用 U-Net 网络作为特征提取的骨干网络，在复现的过程中尝试使用 R2U-Net 来实现骨干网络，从而提升模型最终结果的预测准确度，但是由于设备硬件条件限制导致模型无法进行训练，因此使用数据增强技术来改进 PointGroup 方法。在日后的改进工作中也可以通过改进 U-Net 网络来提高模型的预测的准确度，或者使用其他的网络模型。从数据增强这一角度来看，目前也存在着一些主要的问题，在场景的混合过程中使用了随机混合，也就是将两个场景进行随机的混合。这可能导致模型在混合的过程中，不同的物体实例之间可能存在着大量的相互之间的重叠现象，这也在一定程度上导致了模型分割的精度下降，这是由于 PointGroup 方法在进行分割的过程中使用了实例与实例之间的空白间隙来进行实例的预测，即 PointGroup 假设物体实例之间存在这一定的间隔，因此可以有利于分割实例。在日后的工作中可以尝试在模型的混合过程中减少实例之间的重叠现象，即我们可以为每个点设置一个虚拟的力场，这些点会对不属于自己实例的点产生排斥效应，实例与实例之间的重叠现象约大，这种力就越大，这样可以帮助我们实现场景的融合过程中有效防止重叠现象。

参考文献

[1] Anurag Arnab and Philip H. S. Torr. Pixelwise instance segmentation with a dynamically instantiated network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[2] Min Bai and Raquel Urtasun. Deep watershed transform for instance segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2858–2866, 2017.

[3] R. Qi Charles, Hao Su, Mo Kaichun, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, 2017.

[4] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3070–3079, 2019.

[5] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3150–3158, 2016.

[6] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3d-mpa: Multi-proposal aggregation for 3d semantic instance segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9028–9037, 2020.

[7] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 297–312, Cham, 2014. Springer International Publishing.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[10] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4416–4425, 2019.

[11] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[12] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on X-transformed points. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 828–838, Red Hook, NY, USA, 2018. Curran Associates Inc.

[13] Chen Liu and Yasutaka Furukawa. Masc: Multi-scale affinity with sparse convolution for 3d instance segmentation. *arXiv preprint arXiv:1902.04478*, 2019.

[14] Shih-Hung Liu, Shang-Yi Yu, Shao-Chi Wu, Hwann-Tzong Chen, and Tyng-Luh Liu. Learning gaussian instance segmentation in point clouds. *arXiv preprint arXiv:2007.09860*, 2020.

[15] Shu Liu, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. Sgn: Sequential grouping networks for instance segmentation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3516–3524, 2017.

[16] Quang-Hieu Pham, Thanh Nguyen, Binh-Son Hua, Gemma Roig, and Sai-Kit Yeung. Jsis3d: Joint semantic-instance segmentation of 3d point clouds with multi-task pointwise networks and multi-value conditional random fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8827–8836, 2019.

[17] Jordi Pont-Tuset, Pablo Arbeláez, Jonathan T. Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1):128–140, 2017.

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[19] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–779, 2019.

[20] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 945–953, 2015.

[21] Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2569–2578, 2018.

[22] Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. *Learning Object Bounding Boxes for 3D Instance Segmentation on Point Clouds*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[23] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J. Guibas. Gspn: Generative shape proposal network for 3d instance segmentation in point cloud. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3942–3951, 2019.

[24] Ziyu Zhang, Sanja Fidler, and Raquel Urtasun. Instance-level segmentation for autonomous driving with deep densely connected mrfs. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 669–677, 2016.