Computer Science and Application 计算机科学与应用, 2023, 13(12), 2372-2378
Published Online December 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.1312237

Hans汉斯

基于特征提取的RGBD语义分割算法研究

徐薇蓉

同济大学电子信息与工程学院，上海

收稿日期：2023年11月25日；录用日期：2023年12月21日；发布日期：2023年12月27日

摘　要

RGBD语义分割是近年来备受关注的研究领域。该领域的挑战在于有效地利用RGB和深度图像的不同信息特征。RGB图像具有全局颜色变化的特点，而深度图像则提供了关于对象局部位置的信息。因此，深度图像被认为是更具代表性的局部语义信息源，有助于后续的编码和分割。然而，目前的方法往往将RGB和深度图像通过相同的卷积运算符进行处理，忽略了它们之间的固有差异。为了解决这一问题，本文对传统的重叠补丁嵌入方法进行了修改，以更好地利用深度信息，实现更高精度的语义分割。通过修改传统的重叠补丁嵌入方法，本文能够更好地利用深度信息。具体来说，本文提出了一种改进的方法，通过对深度图像进行处理，更好的进行特征提取。通过在数据集上进行实验，本文的方法在RGBD语义分割任务中取得了较高的准确性和鲁棒性。

关键词

语义分割，特征提取，多模态

Research on RGBD Semantic Segmentation Algorithm Based on Feature Extraction

Weirong Xu

School of Electronic and Information Engineering, Tongji University, Shanghai

Received: Nov. 25th, 2023; accepted: Dec. 21st, 2023; published: Dec. 27th, 2023

Abstract

RGBD semantic segmentation is a research area that has received much attention in recent years. The challenge in this area is to effectively utilize the different information characteristics of RGB and depth images.RGB images feature global color changes, while depth images provide information about the local location of an object. As a result, depth images are considered to be a more

文章引用: 徐薇蓉. 基于特征提取的 RGBD 语义分割算法研究[J]. 计算机科学与应用, 2023, 13(12): 2372-2378.
DOI: 10.12677/csa.2023.1312237

representative source of local semantic information, which helps in subsequent coding and segmentation. However, current approaches tend to process RGB and depth images by the same convolution operator, ignoring the inherent differences between them. To address this problem, this paper modifies the traditional overlapping patch embedding method to better utilize the depth information and achieve higher precision semantic segmentation. By modifying the traditional overlapping patch embedding method, this paper is able to better utilize the depth information. Specifically, this paper proposes an improved method for better feature extraction by processing depth images. By conducting experiments on the dataset, the method in this paper achieves high accuracy and robustness in RGBD semantic segmentation tasks.

Keywords

Semantic Segmentation, Feature Extraction, Multimodality

1. 引言

近年来，计算机视觉领域在处理图像和视频等媒体信息方面取得了巨大的发展。其中，语义分割[1]作为计算机视觉的基本任务之一，能够将图像中的物体进行分类和解释，为提供结构化信息提供了重要的支持。在各个领域中，语义分割的应用也越来越广泛，包括自动驾驶、人流量分析、医学影像诊断等。

在语义分割任务中，人们通常将图像分为前景和背景两部分，其中前景代表感兴趣的区域，背景则是其余部分。为了更深入地分析和理解图像，人们对语义分割进行了标签化处理，赋予了图像一定的语义含义。这样可以帮助人们更好地理解图像的含义，提高其实用价值。

卷积神经网络的引入加速了语义分割领域的进步。通过使用深度神经网络对图像进行像素级和语义级的分割，展示了神经网络在学习特征方面的潜力。随着摄像头技术的不断发展，提供了更理想的条件进行语义分割。然而，在实际场景中，语义分割任务面临着许多挑战，如不同天气和光照条件的影响，相同场景在图像上可能存在巨大的差异。为了解决这些问题，研究人员提供了许多标准数据集，如cityscapes [2]和 NYUv2 [3]等，以便对现实场景进行捕捉和标注。

尽管语义分割领域已经提出了许多优秀的方法，但仍然存在一些待解决的问题。图像不仅仅有表面的含义，还蕴含着推理知识，从具体到抽象的跨越对于语义分割至关重要。因此，许多专家学者致力于研究更准确和高效的算法，进一步提升图像分割的性能。传统方法已经在一定程度上解决了实际问题，而深度学习的发展使得解决计算机视觉问题的趋势变得更加明显。

研究表明，在复杂的场景中，仅依靠 RGB 图像进行分割的结果往往不够理想[4]。因此，利用除 RGB 图像以外的深度图像来获取更多信息是有益的。RGB 图像提供了物体的颜色和纹理等信息，而深度图像提供了物体在空间中的位置信息，同一物体在深度图像上通常呈现连续性。通过充分利用这两种互补的图像信息，可以对场景进行更精细的分析。然而，目前仍然存在着对这两种信息利用不足或不当，以及对几何信息利用不足等问题，导致语义分割的精度仍然较低。为了充分利用不同图像中的有效信息，本文将研究基于深度的 RGBD 语义分割算法，以改进现有算法中存在的相关缺陷，主要改变网络中特征提取的方式，对深度信息进行更加精准的学习，使其能够正确反应物体的轮廓信息。

2. 相关工作

2.1. RGBD 语义分割

随着深度传感器的发展，将 RGB 图像和深度数据结合起来进行语义分割(称为 RGB-D 语义分割)的方法备受关注[5] [6] [7]。虽然全卷积网络(FCN) [8]为端到端密集语义分割铺平了道路，但大多数现有算法仍然严重依赖 RGB 图像，这使得分割结果高度依赖于 RGB 图像的质量。在 RGB 图像受到影响的情况下，如高动态区域或低照度条件下，分割的准确性会大打折扣。深度图像作为一种补充信息源，对光照具有鲁棒性，能提供稳定的补充信息。

目前基于 RGB-D 的方法大致可分为两个阶段：主干阶段和分割阶段。骨干阶段主要是从 RGB 和深度数据中提取特征。这一阶段常用的模型包括 Segformer [5]和 Swin Transformer [7]，这两种模型都是在 ImageNet [9]数据集上预先训练好的。目前已提出了多种特征提取方法。例如，Gupta [10]等人引入了一种深度图像地心嵌入算法，对每个像素的地面高度和重力角进行编码。Li 等人[11]分别使用卷积层和长期短期记忆层捕捉光度和深度通道中的上下文。记忆层对垂直方向的短期和长期空间依赖性进行编码。Chen 等人[12]提出了一种空间信息引导的卷积(S-Conv)算法，该算法将 RGB 特征与三维空间信息整合在一起，使网络能够根据三维空间信息为卷积核推导采样偏移，从而适应几何变换。Cheng 等人[13]改进了每种模式的边界分割，并整合了来自 RGB 和深度数据的局部视觉和几何线索，以增强物体边界的清晰度。

2.2. 重叠补丁嵌入(Overlap Patch Embedding)

Overlap patch embedding 是一种用于图像处理和计算机视觉任务的技术，用于将图像划分为重叠的小块，并将这些小块转换为嵌入向量。这种方法可以帮助提取图像中的局部特征，并在不丢失全局信息。在传统的图像处理任务中，常常采用固定大小的非重叠块来处理图像。然而，这种方法可能会导致信息的丢失，特别是对于边缘和细节等重要的局部特征。为了解决这个问题，overlap patch embedding 将图像划分为重叠的小块，这样每个小块都可以包含一些相邻区域的信息。具体而言，overlap patch embedding 首先将图像划分为固定大小的小块，通常是正方形或矩形。然后，这些小块以一定的步幅进行滑动，使得它们之间有一定的重叠。对于每个小块，可以将其转换为一个嵌入向量，该向量可以表示该小块的特征。这种嵌入向量可以使用各种方法生成，如卷积神经网络(CNN)或自注意力机制等。使用 overlap patch embedding 的好处是可以提取图像中的局部特征，并保留了全局信息。由于小块之间的重叠，相邻区域的信息可以被多次利用，从而提高了特征的丰富性和表达能力。此外，overlap patch embedding 还可以在处理大型图像时减少计算量，因为只需对小块进行处理，而不是整个图像。overlap patch embedding 在许多计算机视觉任务中得到了广泛应用，如图像分类、目标检测、语义分割等。通过提取局部特征并保留全局信息，它可以提高这些任务的性能和准确性。

3. 算法原理与模型结构

3.1. ShapeConv

若想要在特征提取阶段识别到精确的深度信息，需要对深度信息进行另外的处理，识别深度相关的物品信息。在本文中我们发现 ShapeConv [14]具有识别物体形状的功能，用其来提取深度信息中的有效语义。给定一个输入部分 $\mathbb{P} \in \mathbf{R}^{K_h \times K_w \times C_{in}}$，$K_h$ 和 K_w 分别是卷积核的高度和宽度。C_{in} 是输入特征图的通道，普通卷积计算为：

$$F_{c_{out}} = \sum_{i}^{K_h \times K_w \times C_{in}} \left(\mathbb{K}_{i,c_{out}} \times \mathbb{P}_i \right), \tag{1}$$

其中 $\mathbb{K} \in \mathbf{R}^{K_h \times K_w \times C_{in} \times C_{out}}$ 表示卷积层中内核的可学习权重，为了简洁本文忽略了式中的偏置项，C_{out} 表示输出特征图的通道。

从上面的等式中，我们可以看到，当相同的物品放置在不同的距离时，使用常规卷积获得的特征通常是不同的，因此我们最终可能会得到不同的预测结果。例如，当物体靠近和原理时，深度信息所表示的距离信息时不定的，所以很有可能在这两种情况下被识别成不同的两种物体。因此，普通的卷积层不能很好地处理这种情况。其实我们可以根据它的形状来判断，利用相对深度来很好的规避这个问题。基于上述分析，ShapeConv 将输入分解为两个分量 $\mathbb{P}_B$ 和 $\mathbb{P}_S$，其中基础分量 $\mathbb{P}_B$ 描述输入部分的位置，而轮廓分量 $\mathbb{P}_S$ 描述输入部分的形状。根据定义，将该输入部分的平均值称为 $\mathbb{P}_B$，将其相对值称为 $\mathbb{P}_S$，有如下式的定义：

$$\begin{aligned} \mathbb{P}_{\mathrm{B}} &= m(\mathbb{P}), \\ \mathbb{P}_S &= \mathbb{P} - m(\mathbb{P}). \end{aligned} \tag{2}$$

$m(\mathbb{P})$ 是 $\mathbb{P}$ 上的平均函数，并且 $\mathbb{P}_B \in \mathbf{R}^{1\times 1\times C_{in}}$，$\mathbb{P}_S \in \mathbf{R}^{K_h \times K_w \times C_{in}}$。ShapeConv 定义 W_B 和 W_S 和大小分别设置为 $W_B \in \mathbf{R}^1$，$W_S \in \mathbf{R}^{K_h \times K_w \times K_h \times K_w \times C_{in}}$。最终的卷积公式如下：

$$\begin{aligned} \mathbb{F} &= \mathrm{Shapeconv}(\mathbb{K}, \mathbb{W}_B, \mathbb{W}_S, \mathbb{P}) \\ &= \mathrm{Conv}\left(\mathbb{W}_B \lozenge m(\mathbb{K}) + \mathbb{W}_S * (\mathbb{K} - m(\mathbb{K})), \mathbb{P}\right) \\ &= \mathrm{Conv}\left(\mathbb{W}_B \lozenge \mathbb{K}_B + \mathbb{W}_S * \mathbb{K}_S, \mathbb{P}\right) \\ &= \mathrm{Conv}\left(K_B + K_S, \mathbb{P}\right) \\ &= \mathrm{Conv}\left(K_{BS}, \mathbb{P}\right) \end{aligned} \tag{3}$$

其中 $\lozenge$ 和 $*$ 分别表示基础和形状积运算符，可以得到两个卷积分量计算方式分别为：

$$\begin{cases} K_B = \mathbb{W}_B \lozenge \mathbb{K}_B \\ K_{B_{1,1,c_{in},c_{out}}} = \mathbb{W}_B \times \mathbb{K}_{B_{1,1,c_{in},c_{out}}} \end{cases} \tag{4}$$

$$\begin{cases} K_S = \mathbb{W}_S * \mathbb{K}_S \\ K_{S_{k_h,k_w,c_{in},c_{out}}} = \sum_i^{K_h \times K_w} \left(\mathbb{W}_{S_{i,k_h,k_w,c_{in}}} \times \mathbb{K}_{S_{i,c_{in},c_{out}}} \right) \end{cases} \tag{5}$$

最后的卷积核需要将这二者相加，从而达到提取深度信息的效果。

3.2. 网络结构

如图 1 所示，本文采用编码－解码器结构，其中编码部分共有四个阶段，RGB 使用常规的重叠补丁嵌入提取特征，而深度图使用 ShapeConv 作为特征提取层，通过双流网络对不同特征进行分别提取，再使用 Transformer 层进行处理后输入下一阶段。

本文采用了编码－解码器结构，该结构包含四个阶段的编码部分。在这个结构中，RGB 图像的特征提取采用了常规的重叠补丁嵌入方法，而深度图像的特征提取则使用了 ShapeConv 作为特征提取层。双流网络分别对 RGB 和深度图像的特征进行提取。这意味着两种图像的特征被分别处理，充分利用了它们之间的固有差异。通过分别提取 RGB 和深度图像的特征，可以更好地捕捉到它们各自的信息特征。

常规的重叠补丁嵌入被用于提取 RGB 图像的特征。这种方法利用了重叠补丁的方式，将图像分割为多个小块，并通过卷积运算提取每个补丁的特征。这样可以捕捉到图像的全局颜色变化特征。深度图像的特征提取采用了 ShapeConv。ShapeConv 是一种特殊的卷积层，它能够更好地处理深度图像的信息。通过 ShapeConv，深度图像能够提供关于对象局部位置的信息。Transformer 层是一种能够处理序列数据

的神经网络层，通过自注意力机制和位置编码，能够更好地捕捉到特征之间的依赖关系。

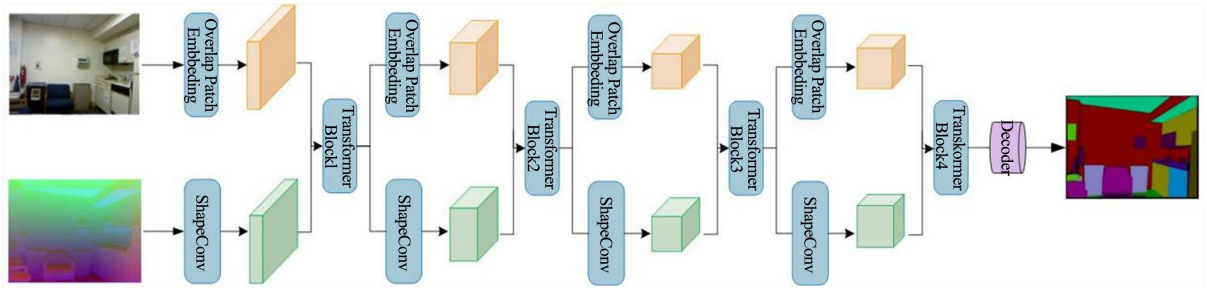

Figure 1. Overall network structure
图 1. 整体网络结构图

4. 实验

4.1. 实验数据集与评价指标

本文实验中主要使用 NYUv2 数据集，该数据集由 1449 张带有逐像素标签的 RGBD 图像组成，每张图像大小为 640 × 480。数据集标注了 40 个语义类别，包括 795 张训练图像和 654 张测试图像。

为了评估我们方法的性能，我们采用了两个常用指标：像素准确度(Pixel Acc.)和平均交并比(Mean Intersection Over Union, mIoU)。这些指标可全面评估分割准确度和预测分割图的整体质量。

4.2. 实验设置

我们的网络架构遵循编码器–解码器结构，由两个分支组成。对于深度分支，我们利用 HHA 图像或行深度图像作为输入。编码器组件采用在 ImageNet 上预训练的 Mix Transformer 编码器(MiT)作为骨干，而解码器则使用 MLP (多层感知器)解码器实现。为了训练网络，我们利用交叉熵损失函数进行监督。我们采用了一个“poly”学习率方案，裁剪大小为 480 × 640。原始学习率设置为 6e−5，批量大小设置为 8。为了优化，我们使用 AdamW 优化器，动量为 0.9，权重衰减为 0.01。我们对 NYUv2 数据集进行了 500 个 epoch 的训练，并使用多个尺度的水平翻转来评估结果，并将其与其他最先进的方法进行比较。为了与其他方法进行公平比较，我们在实验过程中采用了单尺度和多尺度两种测试策略。若文中没有特别说明，均表示实验是单尺度测试，表格中的带“*”表示采用多尺度测试实验。

4.3. 实验结果

通过图 2 可以得知，我们的方法能够更好的捕捉物体的形状特征。如第四行中，我们的算法成功识别出了整个沙发的形状，而其他行的分割结果也更接近实际的物体形状。与原本的网络主干相比，我们通过以上的编码部分，能够使网络模型更好地利用深度信息，并实现更高精度的语义分割。通过对深度图像的特征提取和 RGB 图像的特征提取的不同处理方式，双流网络能够更好地捕捉到两者的信息特征。

从表 1 中可以看出本文的方法和其他先进的方法相比也获得了比较优越的分割性能。在像素准确率和平均交并比上均表现出色。综上所述，通过实验验证，本文的方法在 RGBD 语义分割任务中取得了较高的准确性。这也说明了并通过不同的特征提取方法，充分利用了 RGB 和深度图像的不同信息特征是非常有效的。通过对深度信息的更好利用，本文的方法实现了更高精度的语义分割，并在实验中取得了较好的结果。这对于解决 RGBD 语义分割领域的挑战，有效利用不同信息源的特征具有重要意义。

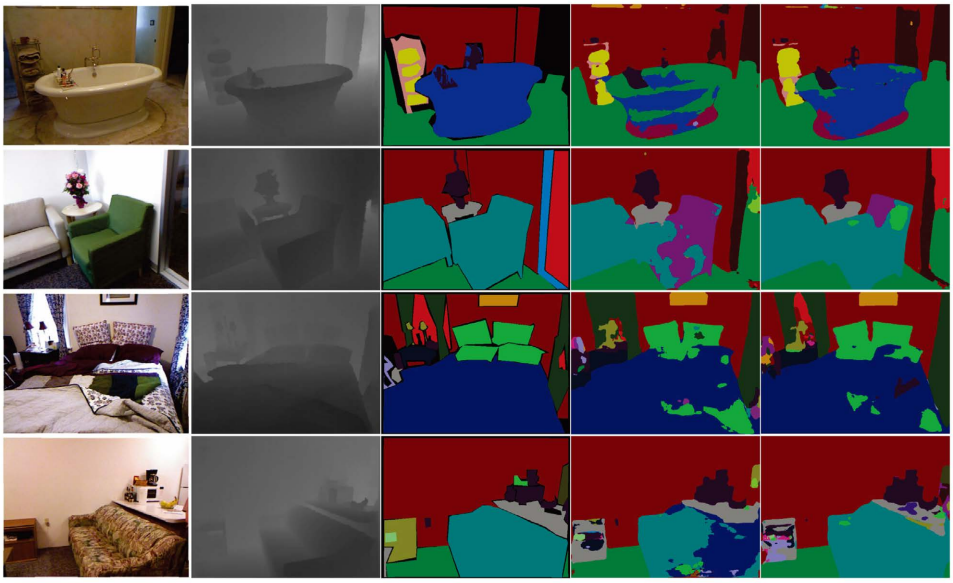

Figure 2. Visualization of segmentation results on the NYUv2 dataset
图 2. NYUv2 数据集上的分割结果可视化

Table 1. Comparison with each state-of-the-art algorithm on NYUv2 dataset
表 1. 在 NYUv2 数据集上与各先进算法的比较

Method	Pixel Acc (%)	mIoU (%)
SGNet [12]	76.8	51.1
NANet [15]	77.9	52.3
InvPT [16]	-	53.5
SA-Gate [17]*	77.9	52.4
InverseForm [18]*	78.1	53.1
Ours*	78.4	53.6

5. 总结

本文主要介绍了 RGBD 语义分割领域的研究现状和挑战，并提出了一种改进的重叠补丁嵌入方法来更好地利用深度信息，以提高语义分割的准确性。传统的方法往往将 RGB 和深度图像通过相同的卷积运算符进行处理，忽略了它们之间的固有差异。为了解决这个问题，文章对传统的重叠补丁嵌入方法进行了修改，以更好地利用深度信息。具体而言，文章提出了一种改进的方法，通过对深度图像进行处理，更好地进行特征提取。通过在数据集上进行实验，本文验证了该方法在 RGBD 语义分割任务中的有效性和鲁棒性。实验结果表明，该方法能够提高语义分割的准确性，并充分利用深度信息的优势。

参考文献

[1] Chen, L.-C., *et al*. (2014) Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. arXiv: 1412.7062. https://doi.org/10.48550/arXiv.1412.7062

[2] Cordts, M., *et al*. (2016) The Cityscapes Dataset for Semantic Urban Scene Understanding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 3213-3223. https://doi.org/10.1109/CVPR.2016.350

[3] Silberman, N., *et al*. (2012) Indoor Segmentation and Support Inference from RGBD Images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y. and Schmid, C., Eds., *Computer Vision—ECCV* 2012, Springer, Berlin. https://doi.org/10.1007/978-3-642-33715-4_54

[4] Hu, X., *et al*. (2019) Acnet: Attention Based Network to Exploit Complementary Features for RGBD Semantic Segmentation. 2019 *IEEE International Conference on Image Processing* (*ICIP*), Taipei, 22-25 September 2019, 1440-1444. https://doi.org/10.1109/ICIP.2019.8803025

[5] Xie, E., *et al*. (2021) SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Advances in Neural Information Processing Systems*, **34**, 12077-12090.

[6] Wang, W. and Neumann, U. (2018) Depth-Aware CNN for RGBD Segmentation. *Proceedings of the European Conference on Computer Vision* (ECCV), Munich, 8-14 September 2018, 144-161. https://doi.org/10.1007/978-3-030-01252-6_9

[7] Liu, Z., *et al*. (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, 10-17 October 2021, 9992-10002. https://doi.org/10.1109/ICCV48922.2021.00986

[8] Long, J., Shelhamer, E. and Darrell, T. (2015) Fully Convolutional Networks for Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 3431-3440. https://doi.org/10.1109/CVPR.2015.7298965

[9] Russakovsky, O., *et al*. (2015) ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, **115**, 211-252. https://doi.org/10.1007/s11263-015-0816-y

[10] Gupta, S., *et al*. (2014) Learning Rich Features from RGB-D Images for Object Detection and Segmentation. *Computer Vision—ECCV* 2014: 13*th European Conference*, Zurich, 6-12 September 2014, 345-360. https://doi.org/10.1007/978-3-319-10584-0_23

[11] Li, Z., *et al*. (2016) LSTM-CF: Unifying Context Modeling and Fusion with LSTMs for RGB-D Scene Labeling. *Computer Vision—ECCV* 2016: 14*th European Conference*, Amsterdam, 11-14 October 2016, 541-557. https://doi.org/10.1007/978-3-319-46475-6_34

[12] Chen, L.-Z., *et al*. (2021) Spatial Information Guided Convolution for Real-Time RGBD Semantic Segmentation. *IEEE Transactions on Image Processing*, **30**, 2313-2324. https://doi.org/10.1109/TIP.2021.3049332

[13] Cheng, Y., *et al*. (2017) Locality-Sensitive Deconvolution Networks with Gated Fusion for RGB-D Indoor Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 1475-1483. https://doi.org/10.1109/CVPR.2017.161

[14] Cao, J., *et al*. (2021) Shapeconv: Shape-Aware Convolutional Layer for Indoor RGB-D Semantic Segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, 10-17 October 2021, 7068-7077. https://doi.org/10.1109/ICCV48922.2021.00700

[15] Wang, J., *et al*. (2016) Learning Common and Specific Features for RGB-D Semantic Segmentation with Deconvolutional Networks. *Computer Vision—ECCV* 2016: 14*th European Conference*, Amsterdam, 11-14 October 2016, 664-679. https://doi.org/10.1007/978-3-319-46454-1_40

[16] Ye, H. and Xu, D. (2022) Inverted Pyramid Multi-Task Transformer for Dense Scene Understanding. *European Conference on Computer Vision*, Tel Aviv, 23-27 October 2022, 514-530. https://doi.org/10.1007/978-3-031-19812-0_30

[17] Chen, X., *et al*. (2020) Bi-Directional Cross-Modality Feature Propagation with Separation-and-Aggregation Gate for RGB-D Semantic Segmentation. *European Conference on Computer Vision*, Glasgow, 23-28 August 2020, 561-577. https://doi.org/10.1007/978-3-030-58621-8_33

[18] Borse, S., *et al*. (2021) Inverseform: A Loss Function for Structured Boundary-Aware Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, 20-25 June 2021, 5897-5907. https://doi.org/10.1109/CVPR46437.2021.00584