

Review on Indoor RGB-D Semantic Segmentation with Deep Convolutional Neural Networks

Sami Barchid

Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRISTAL
F-59000 Lille, France
sami.barchid@univ-lille.fr

José Mennesson

IMT Lille-Douai, Institut Mines-Télécom,
Centre for Digital Systems
Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRISTAL
F-59000 Lille, France
jose.mennesson@imt-lille-douai.fr

Chaabane Djéraba

Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRISTAL
F-59000 Lille, France
chabane.djeraba@univ-lille.fr

Abstract—Many research works focus on leveraging the complementary geometric information of indoor depth sensors in vision tasks performed by deep convolutional neural networks, notably semantic segmentation. These works deal with a specific vision task known as "RGB-D Indoor Semantic Segmentation". The challenges and resulting solutions of this task differ from its standard RGB counterpart. This results in a new active research topic. The objective of this paper is to introduce the field of Deep Convolutional Neural Networks for RGB-D Indoor Semantic Segmentation. This review presents the most popular public datasets, proposes a categorization of the strategies employed by recent contributions, evaluates the performance of the current state-of-the-art, and discusses the remaining challenges and promising directions for future works.

Index Terms—RGB-D Indoor Semantic Segmentation, Deep Convolutional Neural Networks, Deep Learning

I. INTRODUCTION

Semantic segmentation is a fundamental task in computer vision. It is required for many applications such as robot navigation, AR/VR, etc. Semantic segmentation in indoor context is challenging due to cluttered scenes and variation of illumination, camera poses, and object's appearances. Over the last decade, computer vision has shown great advances thanks to deep learning and Deep Convolutional Neural Networks (DCNN) [1], including semantic segmentation [2]. With the advent of precise depth sensors in indoor environments, semantic segmentation models were able to leverage the depth information of a scene in addition to the standard RGB image in order to improve the segmentation performance. These models resolve a specific vision task known as "*RGB-D (epth) indoor semantic segmentation*". The objective of this paper is to introduce the field of RGB-D indoor semantic segmentation using DCNNs, from the main aspects to the current state-of-the-art solutions.

This paper is organized as follows: Section III formulates the basic notions of semantic segmentation. Section IV analyses the main datasets used in RGB-D segmentation papers. An overview and categorization of state-of-the-art approaches

are given in Section IV. Section V reports the quantitative performance of the current state-of-the-art. Finally, Section VI concludes our work.

II. PRELIMINARY NOTIONS

This section discusses the basic concepts related to semantic segmentation. We introduce a formulation and the commonly used metrics. A short overview of RGB semantic segmentation is also presented, given that the RGB-D segmentation field is strongly related to its RGB counterpart.

A. Formulation of Semantic Segmentation

We define the semantic segmentation task as follows: given an input RGB image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, the objective is to produce an output semantic segmentation map $\mathbf{S} \in \mathbb{R}^{H \times W \times C}$ where C is the number of semantic classes. In other words, for each of the $H \times W$ pixels of an RGB image, the semantic segmentation task produces a probability distribution over C categories. In an RGB-D context, a depth map $\mathbf{D} \in \mathbb{R}^{H \times W}$ is available in addition to the RGB input so as to enhance the accuracy of the predicted segmentation map.

B. Metrics

The two most popular metrics used to evaluate a segmentation model's accuracy is the Pixel Accuracy (PA) and the mean Intersection over Union (mIoU). The PA can roughly be described as the ratio of pixels in \mathbf{S} that are correctly predicted. The mIoU is the mean value of all the intersections between the predicted \mathbf{S} and the ground truth over their unions. Because of the ability to compare the similarities between two sets, the mIoU is considered a better metric and is used in Section V to evaluate state-of-the-art models.

C. Overview of RGB Semantic Segmentation with DCNNs

Most recent state-of-the-art segmentation networks can be classified into two paradigms, depending on the kind of architecture used to design the DCNN.

The first paradigm is the encoder-decoder architecture [3]. It is composed of two main modules: the encoder and the decoder. The encoder is usually a standard backbone network [4] and aims to extract features that will be fed to the decoder part. The decoder recovers the spatial information lost by the deep parts of the encoder to reconstruct a semantic segmentation map.

The second paradigm [5] is based on atrous convolution [6]. Atrous convolution is a variant of the standard convolution that introduces another parameter known as the dilation rate. The dilation rate determines the spacing between values in the kernel of the convolution. It expands the receptive field of the resulting feature maps and maintains high resolution, even in the late stages of the network.

III. EXISTING BENCHMARKS

Various public datasets are available in order to evaluate the performance of indoor semantic segmentation models. In this section, we introduce the most popular semantic segmentation RGB-D datasets and analyze the main challenges related to these datasets (and indoor datasets in general). For simplification purposes, we do not mention the additional annotations (for pose estimation, 3D reconstruction, etc) that may be available in the presented datasets. More details can be found in Table 3 of [7].

NYUv2 [8]: this dataset is the most popular for RGB-D indoor segmentation. It contains 1449 images with pixel-wise labels and depth maps captured from a Microsoft Kinect depth sensor with a resolution of 640×480 . The dataset is split into a training set of 795 images and a testing set of 654 images. NYUv2 originally has 13 different categories. However, the recent models mostly evaluate their performance with the more challenging 40-classes settings [9].

SUN-RGBD [10], [11]: this dataset provides 10335 RGB-D images with the corresponding semantic labels. It contains images captured by different depth cameras (Intel RealSense, Asus Xtion, Kinect v1/2) since they are collected from previous datasets. Therefore, the image resolutions vary depending on the sensor used. SUN-RGBD has 37 classes of objects. The training set consists of 5285 images and the testing set consists of 5050 images.

SceneNet RGB-D [12]: this dataset is composed of 5 million photo-realistic 240×320 images of synthesized indoor scenes. These synthetic scenes are randomly generated with physically simulated objects among 255 different classes, which are usually regrouped into the same 13-classes settings as NYUv2. Due to the high quantity of annotated data, SceneNet RGB-D is well suited for pre-training segmentation models before fine-tuning on sparser, real-world datasets.

Stanford 2D-3D-S [13]: it is a large-scale dataset that consists of 70496 RGB images with the associated depth maps. The images are in 1080×1080 resolution and are collected in a 360° scan fashion. The usual class setting employed is 13 classes.

Matterport3D [14]: Similar to Stanford 2D-3D-S, this dataset is a recent large dataset composed of 194 400

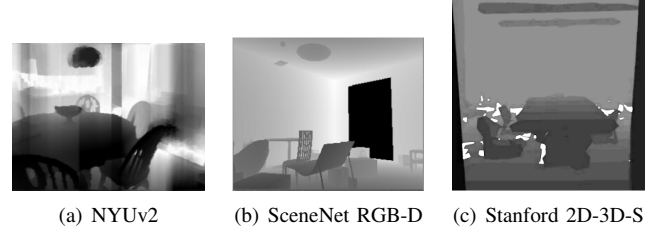


Fig. 1. Example depth maps of RGB-D datasets.

panoramic RGB-D data with a resolution of 1024×1280 . The dataset contains a total of 50811 instance annotations that are regrouped in 40 semantic classes.

The main problem to mention is the important unbalanced distribution of classes in indoor datasets. Some categories (e.g. 'Wall' or 'Floor') cover almost the whole dataset while labels have very few samples [7]. This leads to an important bias to over-represented classes and poor performances for rare objects (usually rare objects found in specific scenes such as TVs or boards). On the other hand, the quality of depth sensors is another important feature to take into account. Compared to the current depth sensor's performance, the depth maps collected by less recent datasets (NYUv2 or SUN-RGBD) are not as accurate. Fig. 1 illustrates examples of depth maps from different datasets. As seen in the NYUv2 example, the early depth sensors provide non-smooth depth maps with many artifacts, as opposed to the more recent 2D-3D-S example. The perfectly-annotated example of SceneNet RGB-D is unreachable in practice because of the synthetic nature of the data. Therefore it can lead to poor feature extraction. Finally, we can also observe that most research papers only focus on NYUv2 and SUN-RGBD even if they have all the drawbacks mentioned above. Their other problem is the limited number of images available, particularly not suited for data-hungry machine learning algorithms such as deep learning.

IV. OVERVIEW OF RGB-D SEGMENTATION MODELS

Depth provides additional geometric information that can benefit an RGB semantic segmentation model [15]. However, there is no established methodology to perfectly merge these two modalities inside a DCNN. Consequently, many research papers propose different methodologies to solve this question, mainly based on standard DCNNs following the encoder-decoder paradigm (see Section II for more details). This section proposes a classification of the current state-of-the-art papers depending on the way depth features are incorporated into a standard DCNN and discusses the pros and cons of each category. Fig. 2 illustrates the three discussed policies.

A. Depth as Input

This approach [15]–[20] is the most popular and was the first attempt to leverage depth in DCNNs. It uses the depth map as an additional input with the RGB image in order to extract more features. Depth and RGB images are fed into separated

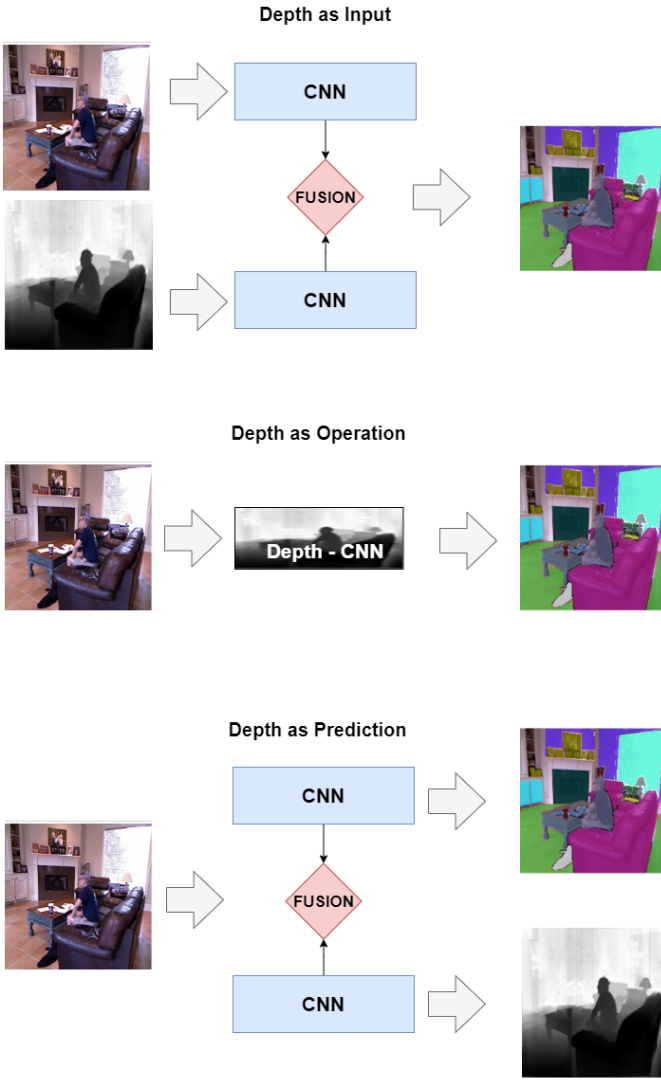


Fig. 2. Illustration of the three defined strategies followed by RGB-D semantic segmentation DCNNs.

branches of a DCNN, then the extracted features are fused to produce the segmentation mask. Research works based on this strategy vary according to the fusion of the designed model [7]. Although this method is intuitive, the main problem is an increase in computational complexity and memory cost because of the need to duplicate the DCNN's modules for each modality.

B. Depth as Operation

Originally designed by [21], the main idea of this paradigm [22]–[24] is to modify some operations (e.g. convolutions and pooling) from the DCNN to take the depth information into account. Instead of using the depth map as an input, the DCNN's operations are directly modified with respect to the depth. For instance, [21] designs a convolution and a pooling operations that adjust their weight with respect to a depth similarity term with the assumption that neighbor pixels of the same depth generally belong to the same class. The

Method	Backbone	Category	NYUv2	SUN-RGBD	FPS
RedNet [16]	ResNet-34×2	DaI	-	46.8	26.0
RedNet [16]	ResNet-50×2	DaI	-	47.8	22.1
ACNet [17]	ResNet-50×3	DaI	48.3	48.1	16.5
IdemPotent [18]	ResNet-101×2	DaI	49.9	47.6	-
RDFNet [15]	ResNet-152×2	DaI	50.1	47.7	5.8
ESANet [19]	ResNet-50 ×2	DaI	50.3	48.17	22.6
SA-Gate [20]	ResNet-50×2	DaI	50.4	49.4	11.9
ESANet* [19]	ResNet-34×2	DaI	51.58	48.04	29.7
3DN-Conv [24]	ResNet-101×1	DaO	48.2	-	-
DA-CNN [21]	ResNet-152×1	DaO	48.4	-	-
2.5-Conv [23]	ResNet-101 ×1	DaO	48.5	48.2	-
Malleable 2.5D [22]	ResNet-101×1	DaO	50.9	-	-
GAD [25]	ResNet-50 ×2	DaP	59.6	54.5	-

TABLE I

PERFORMANCE COMPARISON OF STATE-OF-THE-ART METHODS IN mIoU (%) AND FPS FOR NYUv2 AND SUN-RGBD DATASETS. DaI, DaO AND

DAP ARE ABBREVIATIONS FOR "DEPTH AS INPUT", "DEPTH AS OPERATION" AND "DEPTH AS PREDICTION" CATEGORIES, RESPECTIVELY. BEST AND SECOND BEST PERFORMANCE ARE RESPECTIVELY MARKED IN BOLD AND UNDERLINED. *: PRE-TRAINED ON SCENENET RGB-D [12]

main advantage of this approach is to reduce the additional complexity needed to process both modalities in parallel, while still exploiting the geometric relations of pixels in a depth map.

C. Depth as Prediction

As opposed to the previous paradigms, this recent strategy [24], [25] does not use the depth map during inference but only on the training step. The objective is to design a DCNN that will predict both segmentation and depth maps from an RGB image. In this way, the model learns to implicitly extract the complementary geometric information with the auxiliary depth prediction task. Then the two task-related features can be merged together to improve both predictions, including the targeted segmentation task. Like the "Depth Map as Input" policy, it requires additional complexity due to duplication of some parts in the DCNN. However, unlike the two previous strategies, it does not require any depth sensor and autonomously predicts depth. Hence it enables the use of cheaper RGB cameras for indoor applications that need depth images for additional tasks.

V. PERFORMANCE ANALYSIS

In this section, we report the performance of state-of-the-art models with the two most popular benchmarks: NYUv2 [8] and SUN-RGBD [10], [11]. Table I lists the performance results (in terms of mIoU) of each model in NYUv2 and SUN-RGBD (if available). The classification defined in Section IV is also included. Furthermore, we include the FPS measure reported in Section IV of [19] taken with an NVIDIA Jetson AGX Xavier when it is available. The type and number of backbone networks in the encoder's part are also reported.

The results show that "Depth as Input" and "Depth as Prediction" strategies use several backbone networks instead of one, confirming the problem of computational and memory complexity due to duplicate parts in the model. The recent "Depth as Prediction" strategy seems to be a promising policy, with [25] achieving state-of-the-art results by a large margin. As for the inference speed, few papers achieve real-time

performance (i.e. ≥ 24.0 FPS). However, indoor applications usually run on low-power devices and hence need lightweight and fast models, which is not possible with lots of the reported methods. To solve this issue, "Depth as Operation" seems to be a good solution due to the unique encoder's backbone and the efficient use of depth information inside the DCNN. Another solution that is not explored by the reported methods is to use lightweight backbones such as Mobilenetv2 [26] in order to reduce the encoder's complexity.

VI. CONCLUSION

In this paper, we briefly introduced the field of RGB-D indoor semantic segmentation so as to have a good understanding of the current state-of-the-art. We presented the basic notions of semantic segmentation. We reviewed the most popular RGBD datasets and discussed their main challenges. We proposed a categorization of the recent works based on the way the depth features are exploited inside the DCNN. In addition, we reported the performance found in state-of-the-art models. Finally, during this review, we observe that many recent state-of-the-art models still focus on smaller, older datasets of lower resolution. We believe that future works must exploit the advantages of recent large-scale datasets in order to achieve better results by a large margin.

ACKNOWLEDGMENT

This work was partly supported by IRCICA USR 3380 (CNRS, Univ. Lille, F-59000 Lille, France).

REFERENCES

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.
- [2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
- [3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [5] 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 transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
- [6] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.
- [7] F. Fooladgar and S. Kasaei, "A survey on indoor rgb-d semantic segmentation: from hand-crafted features to deep convolutional neural networks," *Multimedia Tools and Applications*, vol. 79, no. 7, pp. 4499–4524, 2020.
- [8] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgb-d images," in *European conference on computer vision*. Springer, 2012, pp. 746–760.
- [9] S. Gupta, P. Arbelaez, and J. Malik, "Perceptual organization and recognition of indoor scenes from rgb-d images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 564–571.
- [10] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 567–576.
- [11] 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*. Springer, 2013, pp. 141–165.
- [12] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison, "Scenenet rgb-d: 5m photorealistic images of synthetic indoor trajectories with ground truth," *arXiv preprint arXiv:1612.05079*, 2016.
- [13] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese, "Joint 2d-3d-semantic data for indoor scene understanding," *arXiv preprint arXiv:1702.01105*, 2017.
- [14] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *arXiv preprint arXiv:1709.06158*, 2017.
- [15] S.-J. Park, K.-S. Hong, and S. Lee, "Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4980–4989.
- [16] 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.
- [17] X. Hu, K. Yang, L. Fei, and K. Wang, "Acnet: Attention based network to exploit complementary features for rgb-d semantic segmentation," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1440–1444.
- [18] Y. Xing, J. Wang, X. Chen, and G. Zeng, "Coupling two-stream rgb-d semantic segmentation network by idempotent mappings," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1850–1854.
- [19] D. Seichter, M. Köhler, B. Lewandowski, T. Wengelfeld, and H.-M. Gross, "Efficient rgb-d semantic segmentation for indoor scene analysis," *arXiv preprint arXiv:2011.06961*, 2020.
- [20] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, and G. Zeng, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation," in *ECCV*, 2020.
- [21] W. Wang and U. Neumann, "Depth-aware cnn for rgb-d segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 135–150.
- [22] Y. Xing, J. Wang, and G. Zeng, "Malleable 2.5d convolution: Learning receptive fields along the depth-axis for RGB-D scene parsing," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIX*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12364. Springer, 2020, pp. 555–571. [Online]. Available: https://doi.org/10.1007/978-3-030-58529-7_33
- [23] Y. Xing, J. Wang, X. Chen, and G. Zeng, "2.5 d convolution for rgb-d semantic segmentation," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1410–1414.
- [24] Y. Chen, T. Mensink, and E. Gavves, "3d neighborhood convolution: Learning depth-aware features for rgb-d and rgb semantic segmentation," in *2019 International Conference on 3D Vision (3DV)*. IEEE, 2019, pp. 173–182.
- [25] J. Jiao, Y. Wei, Z. Jie, H. Shi, R. W. Lau, and T. S. Huang, "Geometry-aware distillation for indoor semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2869–2878.
- [26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.