Enhancing Semantic Segmentation for Indoor Environments: Integrating Depth Information into Neural Networks

Kristiyan Velikov, University of Twente, The Netherlands

ABSTRACT

Image segmentation is part of computer vision tasks, with significant implications for the accurate identification and classification of indoor settings. Enhancing the performance of neural network-based image segmentation models can be achieved through depth information integration. This study moves beyond a survey to experimentally evaluate the integration of depth data within these models and pinpoint the most effective methods. The research delves into the primary sources of depth data and their incorporation into neural network models for image segmentation. Furthermore, this study probes the impact of depth information on the performance of DeepLabV3 Plus architecture, specifically with the implementation of a Shape-aware Convolution (ShapeConv) layer on the ResNext101 backbone. The research was conducted using the NYU Depth Dataset V2, with a critical focus on addressing the intricacies, challenges, and limitations inherent to depth information integration. In doing so, the study offers insights into the optimization of image segmentation models, particularly in the context of indoor environment analysis.

Keywords

Depth information, image segmentation, neural networks, indoor environments, shapeconv, deeplab,

1. INTRODUCTION

In our increasingly digitized world, computer vision has emerged as a branch of artificial intelligence [1]. This field of study has influenced the way machines perceive, interpret, and make sense of visual data. The significance of computer vision stretches across a diverse array of applications, from medical imaging [2] autonomous vehicles [3], and robotics [18] to security systems [4] construction [5].

Diving deeper into the realm of computer vision, image segmentation [6] emerges as one of its components. It refers to the process of dividing an image into multiple segments, often with the aim of simplifying or altering the representation of an image into something more meaningful and easier to analyze. Image segmentation, by partitioning an image into non-overlapping regions, provides a granular analysis of each constituent object, helping machines to recognize, track, and categorize these elements.

TScIT 39, July 7, 2023, Enschede, The Netherlands

Semantic segmentation [7], a particular type of image segmentation, takes this analysis a step further. Instead of just identifying regions of the image, semantic segmentation seeks to assign each pixel in the image a class label, such as table, 'bed', 'chair', and so forth. This process of pixel-level classification enables a more comprehensive understanding of an image's content, making semantic segmentation valuable in applications that demand a high level of detail and precision.

However, it has been observed that the effectiveness of image segmentation, including semantic segmentation, can be potentially enhanced with the inclusion of depth data [8], [17], [21], [24]. Depth information, which offers insight into the relative distances of objects within a scene, provides additional context that aids in the accurate identification and classification of objects, particularly in complex indoor environments.

2. PROBLEM STATEMENT

Integrating depth information into neural network-based image segmentation models can potentially enhance their performance in accurately identifying and classifying indoor environments. Nonetheless, it is essential to comprehend the most effective methods for including depth data, its impact on the performance of various neural network architectures, and the challenges tied to the process. This study aims to explore these aspects to contribute to the development of more effective image segmentation models for indoor environments. The main research question in relation to the study is:

How does incorporating depth information into neural network-based semantic segmentation models influence their accuracy and effectiveness in identifying and classifying indoor environments?

The question is supported by the following sub-questions, each addressing an aspect of the primary research query:

1. What are the main sources of depth information?
   - Identifying the most effective methods to integrate depth information into these models can assist in maximizing the benefit of incorporating this additional data, leading to potentially improved accuracy and effectiveness.
2. What are the methods for integrating depth information into neural network models for image segmentation?
   - Identifying the most effective methods to integrate depth information into these models can assist in maximizing the benefit of incorporating this additional data, leading to potentially improved accuracy and effectiveness.

3. What challenges and limitations are associated with integrating depth information into neural network-based image segmentation models, particularly in the context of indoor environment recognition?
    - Identifying the challenges and limitations related to the use of depth information can help in devising strategies to overcome them, paving the way for the broader adoption of depth information in image segmentation models.

3. RELATED WORK

This section presents a preliminary overview of literature in the realms of image segmentation, depth information, and neural network-based models. The primary emphasis is on studies that have delved into the inclusion of depth information in image segmentation models and assessed the impact of diverse methods on model performance, as well as the use of RGB-D datasets.

3.1. Sources of Depth Information

As computer vision tasks gain increasing complexity and detail, the integration of depth data becomes valuable to their performance. Depth information augments the perceptual capability of these models by providing spatial understanding of the scenes and objects in focus. Different sources and methods of acquiring depth data have been explored and used in the field. Here, we will delve into the state-of-the-art techniques employed for gathering depth information, highlighting those which incorporate well with computer vision tasks.

3.1.1. Stereo Vision

Stereo vision [9] involves the use of two cameras positioned at different locations to capture the same scene, mimicking human binocular vision. By measuring disparity between corresponding points in the two images, depth information can be derived. While this technique has been widely used in computer vision tasks, it poses challenges including accurate image correspondence and susceptibility to texture-less or occluded regions [10].

3.1.2. Time-of-Flights Sensors

Time-of-Flight (ToF) sensors [11] estimate depth information by measuring the round-trip time of an artificial light signal provided by a laser or an LED. The depth map obtained from ToF sensors is dense and can provide real-time data, making them popular in fields like robotics. However, they can struggle in outdoor environments due to interference from ambient light and often exhibit noise on reflective or absorptive surfaces. The technology is used in devices like the Microsoft's Kinect v2 [12]. The device projects modulated infrared light onto the scene and measures the phase shift between the emitted and reflected light to estimate depth.

Light Detection and Ranging (LIDAR) [15], a type of ToF sensor, uses pulsed lasers to measure the distance between the sensor and the object, creating a 3D representation of the scene. LIDAR offers high accuracy and resolution but is generally more expensive and can struggle with detecting small or thin objects.

3.1.3. Structured Light Sensors

Structured Light sensors [13] project a specific light pattern onto a scene and measure the deformation of this pattern to estimate depth. While offering high resolution, structured light sensors can struggle with rapid motion or sunlight interference. This technology is used in popular devices like Microsoft's Kinect v1, which combines RGB and depth cameras to provide high-resolution depth data [14].

3.1.4. Depth from Monocular Images

An increasingly popular source of depth information comes from monocular images through the use of deep learning. Techniques have been developed to estimate depth from a single 2D image [16], largely powered by Convolutional Neural Networks (CNNs). Despite being more susceptible to errors than hardware-based methods, the progress in this area is rapid and the accessibility of monocular cameras makes this method highly appealing.

3.2. Integrating Depth Information into Neural Networks for Image Segmentation

The incorporation of depth information into image segmentation models can enhance their performance in accurately identifying and classifying objects in both indoor and outdoor environments. One of the ways in which depth information can be integrated is through data augmentation, a technique that can increase the robustness of models to variations in the data [20]. Shorten and Khoshgoftaar [20] discuss the importance of image data augmentation for deep learning, including the use of depth information for improved performance.

In the following sub-sections, we delve deeper into specific techniques for integrating depth information into neural networks for image segmentation.

3.2.1. Depth as Additional Input Channel

The use of depth maps as additional input channels to color channels in the input data has shown to be beneficial for image segmentation. A prominent example of this approach is Fully Convolutional Networks (FCNs) [17]. FCNs treat depth information as an extra channel in the input image, passing it through the same convolutional and pooling layers as the color channels.

A noteworthy extension of this approach involves encoding the depth data into different forms, such as the HHA (Height above ground, Horizontal disparity, and Angle with gravity) encoding [18]. The HHA encoding can be directly concatenated with the RGB image to form a 6-channel input, which can be processed by a standard CNN. The HHA encoding provides different perspectives of depth information, which has been found to improve the performance of segmentation tasks, especially for indoor scenes [18].

3.2.2. Multi-Modal Fusion Techniques

There are multiple techniques for fusing RGB and depth information in neural networks [19]. Early fusion involves concatenating RGB and depth data in the input layer and processing them jointly throughout the network. Late fusion, on the other hand, processes RGB and depth data through separate branches of the network and combines their features

in a later stage. Slow fusion is a balance between the two, fusing the data at multiple stages of the network. These techniques aim to leverage the complementary information in RGB and depth data to improve the segmentation performance.

3.3. Existing Neural-Network Based Segmentation Models Incorporating Depth

The use of depth information has been a trend in the evolution of neural network-based models for image segmentation. This section examines the impact of incorporating depth information into these models, especially convolutional neural networks (CNNs), and assesses the improvements in segmentation performance observed in various studies.

In their study, Couprie et al. [21] adapted a multiscale convolutional network to exploit the depth information in addition to the standard RGB data for indoor scene labeling. Their work utilized the NYU depth dataset, which contains RGB images paired with depth maps, as well as labeled ground truth data for a wide range of indoor scenes such as offices, stores, and home rooms. The dataset is particularly challenging due to the diverse range of object categories, varied lighting conditions, and occluded objects it includes. The team's innovation lay in training a multiscale convolutional network with both the RGB and depth data. They compared the performance of their model with that of a model trained only on RGB data and found that the addition of depth information significantly improved the recognition of certain classes of objects, such as floors, ceilings, and furniture. They found a 15% or more gain in accuracy for those classes of objects.

Chen et al. [22] expanded the field with the introduction of DeepLab, a sophisticated method for semantic image segmentation that employs deep convolutional nets, atrous convolution, and fully connected Conditional Random Fields (CRFs). Where "deep convolutional nets" refers to neural networks with multiple layers designed for image analysis. "Atrous convolution" is a technique that captures contextual information at different scales by introducing gaps in convolutional filters. "Fully connected Conditional Random Fields (CRFs)" are probabilistic models used for post-processing to improve segmentation results by considering spatial dependencies.

Building upon these initial studies, Cao et al. [24] proposed a novel model called ShapeConv which introduced a shape-aware convolutional layer for indoor RGB-D semantic segmentation. They validated their model using three popular indoor RGB-D benchmarks: NYU-DepthV2, SUN-RGBD, and the Stanford Indoor Dataset (SID). Their ShapeConv model showed significant improvements over baseline models including DeepLab across different architectures including ResNet and ResNext, outperforming the baselines in metrics like Pixel Accuracy, Mean Accuracy, Mean IoU, and Frequency Weighted Intersection Over Union.

The integration of depth information into neural network-based image segmentation models contributes to their performance. Theoretically, depth information should provide a more robust spatial understanding, helping the model to distinguish between foreground and background objects more effectively, as well as recognize the spatial relationships between objects. The studies substantiate this theory, demonstrating improvements in segmentation accuracy upon the inclusion of depth data. This highlights the potential and adaptability of incorporating depth information into neural network-based image segmentation models.

4. METHODOLOGY

4.1. Dataset

This study utilizes the NYU Depth Dataset V2, primarily chosen due to its specific focus on indoor environments [23]. This dataset includes a wide range of indoor scene types, providing corresponding RGB images and depth maps, making it well-suited for training and validating models aimed at indoor scene semantic segmentation. Importantly, it offers real depth data, enabling an authentic assessment of models trained to incorporate depth data. The dataset contains 1449 labeled pairs of RGB and depth images for both 13 and 40 classes. Furthermore, to maximize the utility of the depth data, it is preprocessing HHA encoding.

4.2. Model

An open-source implementation of DeepLabV3 Plus with a ResNext101 backbone was utilized for the semantic segmentation task [24]. Four separate models were created, each trained on different combinations of data channels: RGB, RGB + HHA (Horizontal disparity, Height above ground, and Angle with gravity), RGB + Depth, and RGB + Depth + HHA. DeepLabV3 Plus was chosen for its proven performance on several benchmark datasets. Its use of atrous convolution and fully connected Conditional Random Fields (CRFs) improves segmentation outcomes. The ResNext101 backbone was selected due to its ability to extract complex features from a high number of input channels. The ShapeConv layer, integrated into the ResNext101 backbone, utilizes shape cues derived from depth information, thus providing a more refined understanding of indoor scenes [24]. This model setup aims to answer the question of the impact of integrating depth information into the model.

4.3. Implementation

The dataset was divided into 60% for training, 20% for validation, and 20% for testing. Firstly, the training phase involves teaching the model to make accurate predictions using labeled data. Then, the validation phase helps fine-tune the model's settings and assess its performance on unseen data. Finally, the testing phase evaluates the model's performance on completely new data to measure its real-world effectiveness.

All models were trained with the same splits and configurations to minimize bias. The best epochs, as determined by the highest mean Intersection over Union (mIoU), were saved for each model, and used in the testing phase.

4.4. Evaluation Metrics

To assess the performance of the models, several metrics were computed. These metrics include Intersection over Union (IoU), Mean IoU (mIoU), and Accuracy. IoU measures the overlap between predicted and ground truth segmentation masks, while mIoU calculates the average IoU across different

classes or categories. Accuracy represents the overall pixel-level classification accuracy.

To optimize the computation process, the implementation leveraged the processing power of GPUs. Moreover, CUDA's mode was used for convolutional operations, ensuring optimized computation performance.

By utilizing these evaluation techniques and computational optimizations, the performance of the semantic segmentation models trained on the NYU Depth Dataset V2 could be effectively measured and compared.

5. RESULTS

This section provides an overview of the experimental results obtained from testing the four variants of the DeepLabV3 Plus model integrated with the ResNext101 backbone on the NYU Depth Dataset V2.

All the models had very similar times for training and inference given the same dataset.

Table 1. Performance of the four models on the 13 classes of the NYU Depth Dataset V2. *The values represent the scores of the evaluation metrics: Equal mIoU, Frequency Weighted mIoU, Pixel Accuracy, and Class Accuracy. The highest values for each evaluation metric are marked with green color.*

Model	Equal mIoU	Frequency Weighted mIoU	Pixel Accuracy	Class Accuracy
RGB	0.6407	0.7225	0.8339	0.7646
RGB + Depth	0.6677	0.7411	0.8454	0.7805
RGB + HHA	0.6684	0.7423	0.8479	0.7818
RGB + Depth + HHA	0.6740	0.7431	0.8480	0.7933

Table 2. Performance of the four models on the 40 classes of the NYU Depth Dataset V2. *The values represent the scores of the evaluation metrics: Equal mIoU, Frequency Weighted mIoU, Pixel Accuracy, and Class Accuracy. The highest values for each evaluation metric are marked with green color.*

Model	Equal mIoU	Frequency Weighted mIoU	Pixel Accuracy	Class Accuracy
RGB	0.5128	0.6368	0.7669	0.6342
RGB + Depth	0.5383	0.6512	0.7768	0.6617
RGB + HHA	0.5394	0.6518	0.7772	0.6547
RGB + Depth + HHA	0.5430	0.6562	0.7811	0.6677

From **Table 1**, for the 13 classes model, while the RGB alone had an mIoU of 0.6407, adding depth information increased the mIoU to 0.6677. Further integration of HHA information, with or without depth, led to extremely slight improvements, increasing the mIoU to 0.6740 and 0.6684 respectively.

The performance enhancements were also consistent in the 40 classes model as we can observe from **Table 2** Here, the RGB alone model had a mIoU of 0.5128. Adding depth increased this score to 0.5383, while adding HHA and the combination of depth and HHA information pushed the score very little further to 0.5394 and 0.5430 respectively.

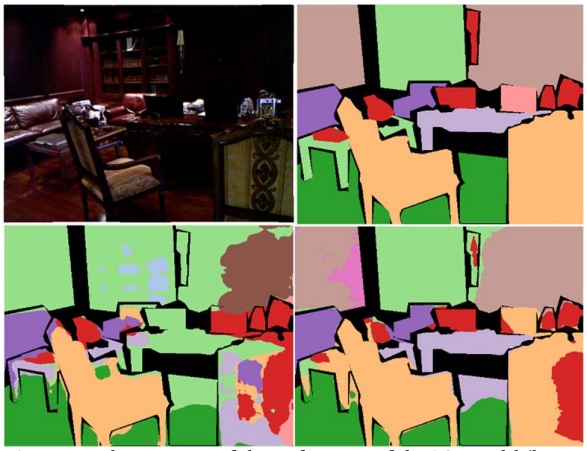

Fig. 1. *Visual comparison of the performance of the RGB model (bottom left) and the RGB, Depth, and HHA model (bottom right), contrasted against the original RGB image (top left) and the ground truth (top right).*

The segmentation produced by the model trained on RGB data appears to struggle with classifying and separating areas where colors are similar or overlap due to the lighting conditions. This is especially notable in the darker areas of the image where there is less contrast between objects as we can observe from **Fig. 1**.

The results obtained from the experiments clearly indicate that the inclusion of depth information consistently improved the performance of the models, both in terms of equal mIoU, frequency weighted mIoU, pixel accuracy, and class accuracy especially when similar colors overlap and in darker light conditions. Models that utilized HHA and Depth in combination with RGB data outperformed the model that only used RGB data across all classes.

Notably, the combination of RGB, Depth, and HHA data yielded the best results, albeit with a slight increase in performance. Importantly, the training and inference times for all four models (RGB, RGB+DEPTH, RGB+HHA, RGB+HHA+DEPTH) were the same. This suggests that utilizing RGB+HHA+DEPTH channels may be the most beneficial approach since it does not incur additional time costs and delivers superior performance. However, it should be noted that the performance improvement is relatively small and insignificant which suggests further exploration.

These results align with the hypothesis that depth data can provide an additional layer of context, facilitating more accurate identification and classification of objects, especially in intricate indoor environments.

6. REFLECTION

6.1. Challenges

Through the process of model implementation, a key challenge was ensuring the correct and meaningful integration of depth information. This required a thorough understanding of the dataset and its structure, as well as the ability to process and incorporate the depth information effectively into the model. Additionally, while the actual training and inference times

were comparable for all models, the preprocessing of HHA data, which involves converting depth maps to HHA encoding, can be computationally intensive. This could pose a challenge in environments where computational resources are limited. However, once the preprocessing is done, there's no significant difference in the time taken for training and inference of the different models.

6.2. Limitations

As with any study, there were limitations to this research. One key limitation was the use of a single dataset, the NYU Depth Dataset V2, for training and validating the models. While it is comprehensive in terms of the variety of indoor environments it represents, it includes only 1,449 labeled pairs of RGB and depth images. For deep learning models, and particularly for complex tasks such as semantic segmentation, larger datasets generally enable models to learn more generalized features and thus achieve better performance. This could be the reason of the only slight performance increase from the results when incorporating depth information.

Additionally, only one model architecture (DeepLabV3 Plus with ResNext101 backbone) was used in the study. While this architecture has proven performance [22][24], other architectures may respond differently to the inclusion of depth data.

7. CONCLUSION

The integration of depth information into semantic segmentation models, as explored in this study, builds on the existing body of work, and extends it by providing an empirical evaluation.

In conclusion, the results have consistently shown enhanced performance across models when depth information is incorporated, confirming the advantages presented in previous literature [17, 21, 24]. It should be noted that even though benefits of incorporating depth and HHA data are evident, the gains are relatively small especially in comparison to the previous literature.

The findings provide insights for researchers and developers working on advanced computer vision tasks, emphasizing the need for depth data incorporation to augment the perceptual capabilities of neural network models. This observation invites further investigation into optimizing the integration of depth data to maximize performance improvements. Additionally, exploring the use of depth and HHA data across a wider range of datasets and model architectures is a promising avenue for future work.

REFERENCES

(1) Paneru, S., & Jeelani, I. (2021). Computer vision applications in construction: Current state, opportunities & challenges. Automation in Construction, 132, 103940. https://doi.org/10.1016/j.autcon.2021.103940

(2) Esteva, A., Chou, K., Yeung, S. *et al.* (2021). Deep learning-enabled medical computer vision. *npj Digit. Med.* **4**, 5 https://doi.org/10.1038/s41746-020-00376-2

(3) Joel Janai, Fatma Güney, Aseem Behl and Andreas Geiger (2020), "Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art", Foundations and Trends® in Computer Graphics and Vision: Vol. 12: No. 1–3, pp 1-308. http://dx.doi.org/10.1561/0600000079

(4) Aydin, I., & Othman, N. A. (2017). A new IoT combined face detection of people by using computer vision for security application. In 2017 International Artificial Intelligence and Data Processing Symposium (IDAP). 2017 International Artificial Intelligence and Data Processing Symposium (IDAP). IEEE. https://doi.org/10.1109/idap.2017.8090171

(5) Seo, J., Han, S., Lee, S., & Kim, H. (2015). Computer vision techniques for construction safety and health monitoring. In Advanced Engineering Informatics (Vol. 29, Issue 2, pp. 239–251). Elsevier BV. https://doi.org/10.1016/j.aei.2015.02.001

(6) Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. (2020). Image Segmentation Using Deep Learning: A Survey. ArXiv Comput. Vis. Pattern Recognition. https://doi.org/10.1109/tpami.2021.3059968

(7) Guo, Y., Liu, Y., Georgiou, T., & Lew, M. S. (2017). A review of semantic segmentation using deep neural networks. In International Journal of Multimedia Information Retrieval (Vol. 7, Issue 2, pp. 87–93). Springer Science and Business Media LLC. https://doi.org/10.1007/s13735-017-0141-z

(8) Muhammad Muzammal Naseer, Salman H. Khan, and Fatih Porikli. 2019. Indoor Scene Understanding in 2.5/3D for Autonomous Agents: A Survey. IEEE Access (2019). https://doi.org/10.1109/access.2018.2886133

(9) Scharstein, D., & Szeliski, R. (2002). In A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms International Journal of Computer Vision (Vol. 47, Issue 1/3, pp. 7–42). Springer Science and Business Media LLC. https://doi.org/10.1023/A:1014573219977

(10) Brown, M. Z., Burschka, D., & Hager, G. D. (2003). Advances in computational stereo. In IEEE Transactions on Pattern Analysis and Machine Intelligence (Vol. 25, Issue 8, pp. 993–1008). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/tpami.2003.1217603

(11) Foix, S., Alenya, G., & Torras, C. (2011). Lock-in Time-of-Flight (ToF) Cameras: A Survey. In IEEE Sensors Journal (Vol. 11, Issue 9, pp. 1917–1926). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/jsen.2010.2101060

(12) Lachat, E., Macher, H., Mittet, M.-A., Landes, T., & Grussenmeyer, P. (2015). FIRST EXPERIENCES WITH KINECT V2 SENSOR FOR CLOSE RANGE 3D MODELLING. In The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences: Vol. XL-5/W4 (pp. 93–100). Copernicus GmbH. https://doi.org/10.5194/isprsarchives-xl-5-w4-93-2015

(13) Salvi, J., Pagès, J., & Batlle, J. (2004). Pattern codification strategies in structured light systems. In Pattern Recognition (Vol. 37, Issue 4, pp. 827–849). Elsevier BV. https://doi.org/10.1016/j.patcog.2003.10.002

(14) Sarbolandi, H., Lefloch, D., & Kolb, A. (2015). Kinect range sensing: Structured-light versus Time-of-Flight Kinect. In Computer Vision and Image Understanding (Vol. 139, pp. 1–20). Elsevier BV. https://doi.org/10.1016/j.cviu.2015.05.006

(15) Raj, T., Hashim, F. H., Huddin, A. B., Ibrahim, M. F., & Hussain, A. (2020). A Survey on LiDAR Scanning Mechanisms. In Electronics (Vol. 9, Issue 5, p. 741). MDPI AG. https://doi.org/10.3390/electronics9050741

(16) Zhao, C., Sun, Q., Zhang, C., Tang, Y., & Qian, F. (2020). Monocular depth estimation based on deep learning: An overview. In Science China Technological Sciences (Vol. 63, Issue 9, pp. 1612–1627).

Springer Science and Business Media LLC. https://doi.org/10.1007/s11431-020-1582-8

(17) Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. https://doi.org/10.1109/cvpr.2015.7298965

(18) Gupta, S., Girshick, R., Arbeláez, P., & Malik, J. (2014). Learning Rich Features from RGB-D Images for Object Detection and Segmentation. In Computer Vision – ECCV 2014 (pp. 345–360). Springer International Publishing. https://doi.org/10.1007/978-3-319-10584-0_23

(19) Zhang, Y., Sidibé, D., Morel, O., & Mériaudeau, F. (2021). Deep multimodal fusion for semantic image segmentation: A survey. In Image and Vision Computing (Vol. 105, p. 104042). Elsevier BV. https://doi.org/10.1016/j.imavis.2020.104042

(20) Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. In Journal of Big Data (Vol. 6, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1186/s40537-019-0197-0

(21) Couprie, C., Farabet, C., Najman, L., & LeCun, Y. (2013). Indoor Semantic Segmentation using depth information (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1301.3572

(22) Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. In IEEE Transactions on Pattern Analysis and Machine Intelligence (Vol. 40, Issue 4, pp. 834–848). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/tpami.2017.2699184

(23) Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. ECCV (5), 7576, 746-760.

(24) Cao, J., Leng, H., Lischinski, D., Cohen-Or, D., Tu, C., & Li, Y. (2021). ShapeConv: Shape-aware Convolutional Layer for Indoor RGB-D Semantic Segmentation (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2108.10528