

Convolutional Networks for Semantic Heads Segmentation using Top-View Depth Data in Crowded Environment

Daniele Liciotti, Marina Paolanti, Rocco Pietrini, Emanuele Frontoni, Primo Zingaretti
Department of Information Engineering (DII), Università Politecnica delle Marche
Via Brecce Bianche, 12, 60131, Ancona, Italy
Email: d.liciotti@pm.univpm.it, m.paolanti@pm.univpm.it, r.pietrini@pm.univpm.it,
e.frontoni@univpm.it, p.zingaretti@univpm.it

Abstract—Detecting and tracking people is a challenging task in a persistent crowded environment (i.e. retail, airport, station, etc.) for human behaviour analysis of security purposes. This paper introduces an approach to track and detect people in cases of heavy occlusions based on CNNs for semantic segmentation using top-view depth visual data. The purpose is the design of a novel U-Net architecture, U-Net3, that has been modified compared to the previous ones at the end of each layer. In particular, a batch normalization is added after the first ReLU activation function and after each max-pooling and up-sampling functions. The approach was applied and tested on a new and public available dataset, TVHeads Dataset, consisting of depth images of people recorded from an RGB-D camera installed in top-view configuration. Our variant outperforms baseline architectures while remaining computationally efficient at inference time. Results show high accuracy, demonstrating the effectiveness and suitability of our approach.

I. INTRODUCTION

Detecting and tracking people is an important and challenging task for many interactive and intelligent systems, such as visual surveillance and human-computer interaction. In recent years, this has been a topic of considerable research and robust methods have been developed to track isolated or a small number of humans for which only transient occlusion exists [1]. However, tracking in more crowded situations, where a lot of people are present, is a problem that remains largely open due to many serious issues, such as the exhibition of persistent occlusion, change of appearance, dynamic and complex background [2]. In fact, all these issues cause extreme problems when encountered with the crowded environment, since conventional surveillance technologies have difficulty in understanding the image [3]. In these situations, sensors viably adopted are RGB-D cameras because of their reliability, availability and affordability. Researchers have demonstrated the great value (both in efficiency and accuracy) of depth camera in coping with severe occlusions among humans and complex background [4]. The combination of high-resolution depth and visual information opens up new challenges and opportunities for activity recognition and people tracking

for many applications. Reliable depth maps can provide valuable additional information to significantly improve tracking and detection results [5]. When dealing with 2D images, video, and even 3D data the semantic segmentation is proved to be of the high-level task. It paves the way towards the complete scene understanding and it is being tackled using deep learning architectures, usually Convolutional Neural Networks (CNNs), which are surpassing other approaches by a large margin in terms of accuracy and sometimes even efficiency [6], [7], [8], [9]. Semantic segmentation is understanding an image at pixel level. For this reason, each pixel of the image has to be assigned to an object class. In addition to object recognition, this technique is used to define the boundaries of each object.

Silberman et al. [10] were the first to deal with the task of semantic full image labeling using depth information. After the release of the NYU dataset, the indoor RGB-D pixel-wise semantic segmentation has gained popularity [11]. This dataset has proved the benefit of the depth channel to improve segmentation. Their approach used features such as RGB-SIFT, depth-SIFT and pixel location as input to a neural network classifier to predict pixel. Other datasets employed for semantic image segmentation present the objects centered into the images, under nice lightening conditions [12], [13]. A common thread in the approaches adopted by Gupta et al. [14] and Hermans et al. [15] is the use of hand engineered features for classification of either RGB or RGBD images. The first cited authors above use boundary detection and hierarchical grouping before performing category segmentation. The second work uses Random Forest as classifier and it focuses on real-time joint reconstruction and semantic segmentation. As previously stated, the success of CNNs for object classification has led researchers to exploit their feature learning capabilities for segmentation. Three sub-categories were developed: patch-wise training [16], [17], end-to-end learning of pixel-wise classification [18], [19], [20] and multiscale semantic segmentation [21], [19], [22], [23]. In [16], patch-wise classification is used.

This paper outlines a method for the detection and tracking of people in depth images, acquired with a low-

resolution Time-of-Flight (ToF) camera. This depth sensor is placed perpendicular to the ground in order to provide distance information from a top-view position. The paper introduces an approach to track and detect people in the presence of persistent and temporarily heavy occlusion with CNNs for semantic segmentation using top-view depth data in crowded environment. While many studies have performed semantic segmentation for RGB-D images, most existing methods focus on either only medical images or only object recognition in indoor environment. To the best of our knowledge, this is the first exploitation of data acquired from an RGB-D camera in top-view configuration for people detection and tracking in a crowded environment. Our work differs from the references above in both our approach to segmentation and recognition. The approach has been applied to a newly collected dataset (TVHeads) of depth images of people recorded from an RGB-D camera installed in top-view configuration. The goal is to design a novel U-Net architecture, the so-called U-Net3, that has been modified compared with the previous ones, at the end of each layer. In particular, a batch normalization is added after the first ReLU activation function and after each max-pooling and up-sampling functions. The approach presented in this paper does not require persons to be isolated, i.e. not occluded, when they first enter the scene. In order to identify a person, we require that at least the head region must be visible. We assume an RGB-D camera in top-view configuration and people passing under this camera walk on a known ground plane with no significant false alarms due to shadows, reflections, or other reasons.

To briefly summarize, the main contributions of this work are: (i) a demonstration that semantic segmentation can be applied in real-time for people detection and tracking in crowded environment; (ii) a novel U-Net3 design for semantic segmentation; (iii) a challenging new dataset of depth images in a heavy crowded environment (more than 3 people per square meter) collected by an RGB-D camera in top-view configuration, hand-labelled with ground truth semantic segmentations. The application of our approach to TVHeads dataset yields good results in terms of precision, recall and F1-score and demonstrates the effectiveness of the proposed approach.

The paper is organized as follows: Section II introduces our approach consisting of a modified U-Net (Subsection II-A1), a state-of-art CNNs comparison (Subsections II-A2, II-A3, II-A4) and gives details on the “TV-Heads Dataset” (Subsection II-B). Final sections present results (Section III) and conclusions (Section IV) with future works.

II. METHODS

In this section, we introduce a novel modified U-Net architecture, the so called *U-Net3*, for heads images segmentation. We also present the Convolutional Networks for image Segmentation taken from the literature as well

as the dataset used for the comparison. Further details are given in the following subsections. The architecture proposed is comprehensively evaluated on the TVHeads dataset. Details on the data collection and ground truth labelling are discussed in Subsection II-B.

A. Convolutional Networks for image Segmentation

1) *U-Net3*: The U-Net3¹ architecture is illustrated in Figure 1. We build upon a more elegant architecture, modifying and extending the ones proposed in [23], [24] such that it works with very few training images and yields more precise segmentations. It is composed of two main parts: a contracting path (left side) and an expansive path (right side). The first path follows the typical architecture of a convolutional network. It consists of the repeated application of two 3×3 convolutions (unpadded convolutions), each followed by a *ReLU* and a 2×2 max pooling operation with stride 2 for downsampling. At each downsampling step, the number of feature channels is doubled. Every step in the expansive path consists of an upsampling of the feature map followed by a 2×2 convolution (“up-convolution”). It halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting path, and two 3×3 convolutions, each followed by a *ReLU*. At the final layer a 1×1 convolution is used to map each 32-component feature vector to the desired number of classes. Compared with [23], [24], U-Net3 architecture proposed in this work has been modified at the end of each layer. The main idea is to add a batch normalization after the first *ReLU* activation function and after each max-pooling and up-sampling functions.

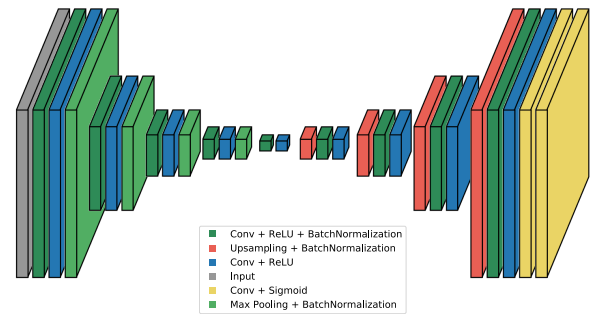


Fig. 1: Illustration of implemented U-Net3 architecture.

2) *SegNet*: SegNet was presented in Vijay *et al.* in [20]. The architecture consists of a sequence of non-linear processing layers (encoders) and a corresponding set of decoders followed by a pixelwise classifier. Typically, each encoder consists of one or more convolutional layers with batch normalization and a *ReLU* non-linearity, followed by non-overlapping max pooling and sub-sampling. The sparse encoding, due to the pooling process, is up-sampled in the decoder using the max pooling indices in the

¹<https://github.com/danielelic/deep-segmentation>

encoding sequence. One key ingredient of the SegNet is the use of max-pooling indices in the decoders to perform upsampling of low resolution feature maps. This has the important advantages of retaining high frequency details in the segmented images and also reducing the total number of trainable parameters in the decoders. The entire architecture can be trained end-to-end using stochastic gradient descent. The raw SegNet predictions tend to be smooth even without a CRF based post-processing.

3) *ResNet*: A ResNet is composed of a sequence of residual units [25]. The output x_n of the n -th RU in a ResNet is computed as:

$$x_n = x_{n-1} + F(x_{n-1}; W_n) \quad (1)$$

where $F(x_{n-1}; W_n)$ is the residual, which is parametrized by W_n . In this way, instead of computing the output x_n directly, F only computes a residual that is added to the input x_{n-1} . This design can be referred as skip connection, since there is a connection from the input x_{n-1} to the output x_n that skips the actual computation F . It has been empirically observed that ResNets have superior training properties over traditional feedforward networks. This can be explained by an improved gradient flow within the network.

4) *FractalNet*: Fractal network is introduced by Larsson *et al.* in [26]. Let C denotes the index of a truncated fractal $f_C(\cdot)$ (i.e., a few stacked layers) and the base case of a truncated fractal is a single convolution:

$$f_1(z) = \text{conv}(z)$$

According to the expansion rule:

$$z' = \text{conv}(z)$$

$$f_{C+1}(z) = \text{conv}(\text{conv}(z') \oplus f_C(z'))$$

can be defined recursively the successive fractals, where $\oplus$ is a join operation and $\text{conv}(\cdot)$ is a convolution operator. Two blobs are merged by the join operation $\oplus$. As these two blobs contain features from different visual levels, joining them can enhance the discrimination capability of our network. Generally, this operation can be summation, maximization and concatenation. In order to enlarge the receptive field and enclose more contextual information, down-sampling and up-sampling operations are added in the above expansion rule. In particular, a max pooling with a stride of 2 and a deconvolution also with a stride of 2 are added. After the down-sampling operation, the receptive field of a fractal becomes broader. When combining different receptive fields through the join operation, the network can harness multi-scale visual cues and promote itself in discriminating.

B. TVHeads Dataset

The *TVHeads*² dataset contains depth images of people recorded from an RGB-D camera installed in top-view

configuration. The heavy crowded dataset is defined as a scenario where at least 3 people per square meter are passing in the area of the camera, with some cases of physical contacts between people entering the area of interest. In particular, the purpose of this dataset is to localize the heads of people who come across the view area of the camera. A total of 1815 depth images (16 bit) with a dimension of 320×240 pixels was collected. Additionally, after the image preprocessing phase, the depth images are also converted with an appropriate scaling. In this way it is possible to obtain an image (8 bit) where the heads silhouette is highlighted, by improving image contrast and brightness. The ground truth was manually labelled by 6 human annotators with a majority policy. Figure 2 shows an example of a dataset instance that includes the three images described above, i.e. 16 bit depth image, 8 bit depth image and the corresponding ground truth.

III. RESULTS AND DISCUSSIONS

In this section, the results of the experiments conducted on TVHeads dataset are reported. In addition to the performance of U-Net3, here is also presented the performance of different approaches taken from literature and based on CNNs in order to solve the problem of heads image segmentation. Each *CNN* implementation is trained with two types of depth images in order to highlight the heads silhouette, improving the images contrast and brightness: 16-bit (original depth images) and 8-bit (scaled images). The general procedure for training neural network based models is to take the dataset and split it into three parts: training, validation, and test. In this case, train, test and validation are chosen respectively to learn model parameters. Once this is complete, the best model is also evaluated over the never before seen test set. Furthermore, different combinations of hyper-parameters are tested, but in this case a learning rate equal to 0.001 and an Adam optimization algorithm have been used.

1) *Performance Evaluation*: Semantic segmentation performances are divided into two different tables. Table I shows Jaccard [27] and Dice [28] indices for training and validation respectively. Table II reports the results in term of accuracy, precision, recall and F1-score. Both tables refer to a learning process conducted during 200 epochs.

In Table I, the best *CNN* architecture is the U-Net3 8-bit version. Indeed, Jaccard index reaches a value equal to 0.962520 regarding train performances. The second best is Fractal Net 16-bit version also obtaining higher values regarding validation performances. For each network is highlighted in red color the best bit version model.

As in the previous case, also in the Table II the best *CNN* architecture, in terms of accuracy, is U-Net3 8-bit version, while Fractal Net 16-bit version exceeds slightly in precision, recall and F1-score metrics.

In Figure 3 is shown the trend of Jaccard index for each CNN during the fit procedure. It is easy to see how the FractalNet and ResNet reach high values immediately

²<http://vrai.dii.univpm.it/tvheads-dataset>



(a) 8 bit Depth image.

(b) Ground truth.

Fig. 2: TVHeads Dataset. It consists of 8 bit scaled depth image (Figure 2a) and the corresponding ground truth (Figure 2b).

TABLE I: Jaccard and Dice indices of different CNN architectures.

Net	Bit	Jaccard Train	Jaccard Validation	Dice Train	Dice Validation
Fractal [26]	8	0.960464	0.948000	0.979833	0.973306
	16	0.961636	0.947762	0.980443	0.973180
U-Net [23]	8	0.896804	0.869399	0.945595	0.930138
	16	0.894410	0.869487	0.944262	0.930188
U-Net2 [24]	8	0.923823	0.939086	0.960403	0.968586
	16	0.923537	0.938208	0.960249	0.968119
U-Net3	8	0.962520	0.931355	0.980902	0.964458
	16	0.961540	0.929924	0.980393	0.963690
SegNet [20]	8	0.884182	0.823731	0.938531	0.903347
	16	0.884162	0.827745	0.938520	0.905756
ResNet [25]	8	0.932160	0.856337	0.964889	0.922609
	16	0.933436	0.848240	0.965572	0.917889

TABLE II: Semantic segmentation results of different ConvNet architectures.

Net	Bit	Accuracy	Precision	Recall	F1-Score
Fractal [26]	8	0.994414	0.991400	0.993120	0.992235
	16	0.994437	0.992667	0.993297	0.992970
U-Net [23]	8	0.992662	0.946475	0.950483	0.948408
	16	0.992569	0.945083	0.948957	0.946938
U-Net2 [24]	8	0.993156	0.970013	0.969206	0.969568
	16	0.993165	0.967884	0.970557	0.969123
U-Net3	8	0.994572	0.990451	0.990387	0.990419
	16	0.994559	0.989382	0.989411	0.989396
SegNet [20]	8	0.992683	0.946304	0.953136	0.949625
	16	0.992699	0.946237	0.953342	0.949658
ResNet [25]	8	0.993789	0.968399	0.968374	0.968359
	16	0.993819	0.968765	0.969256	0.968992

after a few epochs. Instead, the U-Net3 increases its value more slowly. The classic U-Net is always below all other networks.

2) *Qualitative Evaluation:* Table III presents qualitative semantic segmentation results. The table shows the predicted images for each architecture. The best results are obtained by using the U-Net3 (8 and 16 bit), while typical U-Nets and SegNet provide more smooth and rounded

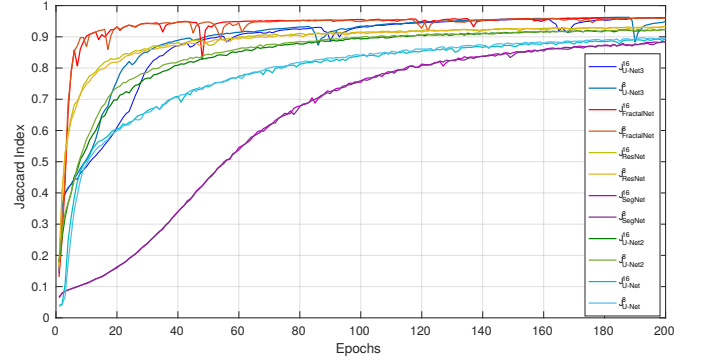


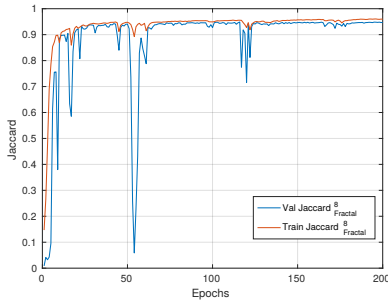
Fig. 3: Jaccard index trends.

shapes of the heads.

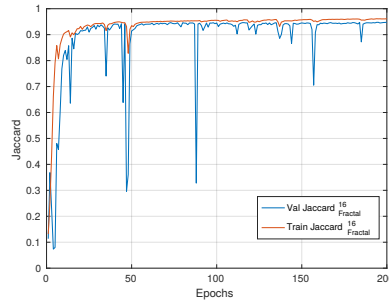
As can be seen, the performance of all networks is good, with F1-Scores ranging that exceed the 90%, thus demonstrating the effectiveness and the suitability of the proposed approach. The comparison shows that U-Net3 performs better than the previous U-Nets design. The system is able to perform at 10 fps on a single core CPU ensuring real-time performances.

IV. CONCLUSION

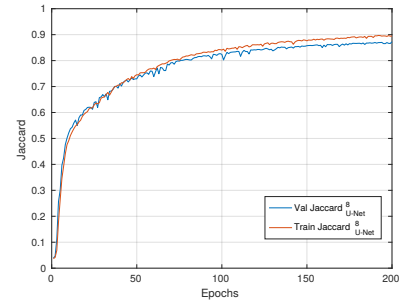
Humans detection and tracking in a crowded situation, where large number of people are present and exhibit persistent occlusion, represents a challenging but rewarding task. In this paper, U-Net3 for semantic heads segmentation has been introduced, using depth images acquired from an RGB-D camera installed in top-view configuration. We compare the performance of the novel network with different approaches taken from literature based on CNNs in order to solve the problem of heads images segmentation: U-Net, U-Net2, FractalNet, SegNet and ResNet. U-Net3 approach is able to learn a high level representation of image content and to achieve high precision and recall. The experiments on TVHeads dataset yield high accuracies and demonstrate the effectiveness and suitability of our approach. Further investigation will be devoted to improve our approach by employing a larger dataset. Moreover, we will extend the evaluation by



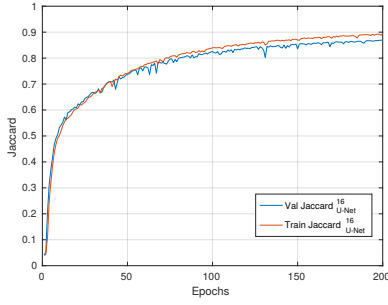
(a) Fractal Net 8-bit.



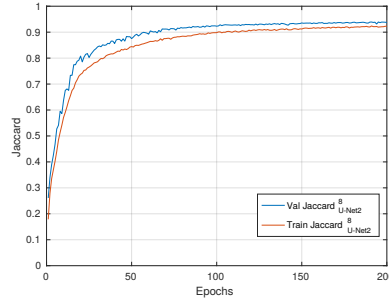
(b) Fractal Net 16-bit.



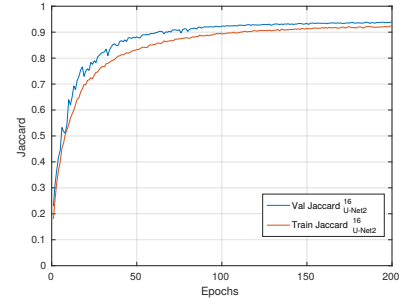
(c) U-Net 8-bit.



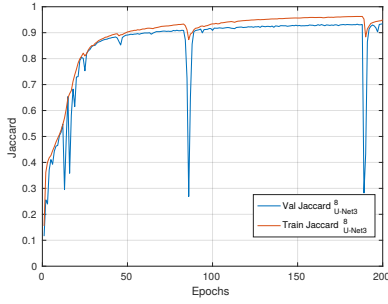
(d) U-Net 16-bit.



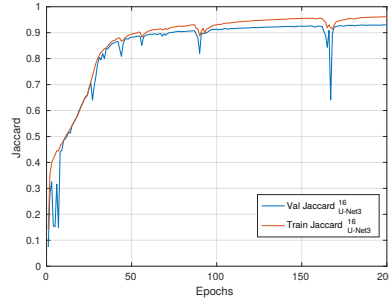
(e) U-Net2 8-bit.



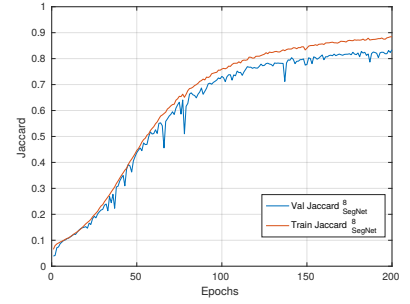
(f) U-Net2 16-bit.



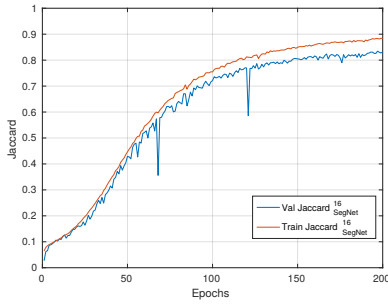
(g) U-Net3 8-bit.



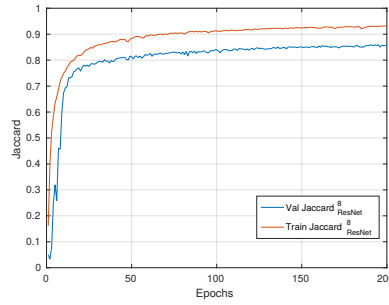
(h) U-Net3 16-bit.



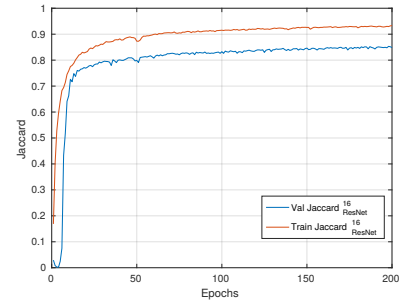
(i) SegNet 8-bit.



(j) SegNet 16-bit.



(k) ResNet 8-bit.



(l) ResNet 16-bit.

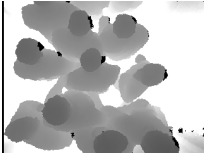
Fig. 4: Jaccard index results.

comparing our U-Net3 with other CNNs existing in the literature. A more complex dataset with other crowded scene will be also collected to propose a more challenging test environment.

REFERENCES

- [1] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and tracking of multiple humans in crowded environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 7, pp. 1198–1211, 2008.
- [2] J. Liu, Y. Liu, G. Zhang, P. Zhu, and Y. Q. Chen, "Detecting and tracking people in real time with rgb-d camera," *Pattern Recognition Letters*, vol. 53, pp. 16–23, 2015.
- [3] W. Ye and Z. Zhong, "Robust people counting in crowded environment," in *Robotics and Biomimetics, 2007. ROBIO 2007. IEEE International Conference on*. IEEE, 2007, pp. 1133–1137.
- [4] M. Sturari, D. Liciotti, R. Pierdicca, E. Frontoni, A. Mancini,

TABLE III: Qualitative result of prediction.

	Label	
		
FractalNet [26]		
U-Net [23]		
U-Net2 [24]		
U-Net3		
SegNet [20]		
ResNet [25]		

M. Contigiani, and P. Zingaretti, “Robust and affordable retail customer profiling by vision and radio beacon sensor fusion,” *Pattern Recognition Letters*, 2016.

- [5] D. Liciotti, M. Paolanti, E. Frontoni, and P. Zingaretti, “People detection and tracking from an rgb-d camera in top-view configuration: Review of challenges and applications,” in *International Conference on Image Analysis and Processing*. Springer, 2017, pp. 207–218.
- [6] M. Paolanti, C. Kaiser, R. Schallner, E. Frontoni, and P. Zingaretti, “Visual and textual sentiment analysis of brand-related social media pictures using deep convolutional neural networks,” in *International Conference on Image Analysis and Processing*. Springer, 2017, pp. 402–413.
- [7] M. Sturari, M. Paolanti, E. Frontoni, A. Mancini, and P. Zingaretti, “Robotic platform for deep change detection for rail safety and security,” in *Mobile Robots (ECMR), 2017 European Conference on*. IEEE, 2017, pp. 1–6.
- [8] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, “Simultaneous detection and segmentation,” in *European Conference on Computer Vision*. Springer, 2014, pp. 297–312.
- [9] M. Paolanti, M. Sturari, A. Mancini, P. Zingaretti, and E. Frontoni,

toni, “Mobile robot for retail surveying and inventory using visual and textual analysis of monocular pictures based on deep learning,” in *Mobile Robots (ECMR), 2017 European Conference on*. IEEE, 2017, pp. 1–6.

- [10] N. Silberman and R. Fergus, “Indoor scene segmentation using a structured light sensor,” in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 601–608.
- [11] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from rgbd images,” *Computer Vision–ECCV 2012*, pp. 746–760, 2012.
- [12] S. Gould, R. Fulton, and D. Koller, “Decomposing a scene into geometric and semantically consistent regions,” in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1–8.
- [13] C. Liu, J. Yuen, and A. Torralba, “Sift flow: Dense correspondence across scenes and its applications,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5, pp. 978–994, 2011.
- [14] S. Gupta, P. Arbeláez, 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.
- [15] A. Hermans, G. Floros, and B. Leibe, “Dense 3d semantic mapping of indoor scenes from rgb-d images,” in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2631–2638.
- [16] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, “Learning hierarchical features for scene labeling,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1915–1929, 2013.
- [17] D. Grangier, L. Bottou, and R. Collobert, “Deep convolutional networks for scene parsing,” in *ICML 2009 Deep Learning Workshop*, vol. 3, 2009.
- [18] 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.
- [19] H. Noh, S. Hong, and B. Han, “Learning deconvolution network for semantic segmentation,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1520–1528.
- [20] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *CoRR*, 2015.
- [21] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” *arXiv preprint arXiv:1511.07122*, 2015.
- [22] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, “Attention to scale: Scale-aware semantic image segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3640–3649.
- [23] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” *arXiv preprint arXiv:1505.04597*, 2015.
- [24] H. Ravishanker, R. Venkataramani, S. Thiruvankadam, P. Sudhakar, and V. Vaidya, “Learning and incorporating shape models for semantic segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 203–211.
- [25] 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.
- [26] G. Larsson, M. Maire, and G. Shakhnarovich, “Fractalnet: Ultra-deep neural networks without residuals,” *arXiv preprint arXiv:1605.07648*, 2016.
- [27] P. Jaccard, “Étude comparative de la distribution florale dans une portion des alpes et des jura,” *Bull Soc Vaudoise Sci Nat*, vol. 37, pp. 547–579, 1901.
- [28] L. R. Dice, “Measures of the amount of ecologic association between species,” *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.