Driver Attention Assistance by Pedestrian/Cyclist Distance Estimation from a Single RGB Image: A CNN-based Semantic Segmentation Approach

Angelo Genovese*, Vincenzo Piuri*, Francesco Rundo†, Fabio Scotti*, Concetto Spampinato‡
* Università degli Studi di Milano, Department of Computer Science, 20133 Milan (MI), Italy
{*angelo.genovese, vincenzo.piuri, fabio.scotti*}*@unimi.it*
† STMicroelectronics, ADG, Central R&D, 95121 Catania (CT), Italy *francesco.rundo@st.com*
‡ Università degli Studi di Catania, Department of Electrical, Electronic and Computer Engineering, 95125 Catania (CT), Italy *concetto.spampinato@dieei.unict.it*

***Abstract*—Automotive companies are investing a relevant amount of resources for designing autonomous driving systems, driver assistance technologies, and systems for assessing the driver's attention. In this context, two important applications consist in processing images of the surrounding environment to respectively separate the different objects in the scene (semantic segmentation) and to estimate their distances. In both applications, methods based on Deep Learning (DL) and Convolutional Neural Networks (CNN) are being increasingly used, considering LiDAR scans or RGB images. However, LiDAR scanners require dedicated sensors, high costs, and post-processing algorithms to estimate a dense depth map or a three-dimensional representation of the surrounding environment. Moreover, current methods in the literature based on RGB images do not consider the combination of semantic segmentation and depth estimation for assessing the distances of specific objects in the scene. In this paper, we propose the first method in the literature able to estimate the distances of pedestrians/cyclists from the vehicle by using only an RGB image and CNNs, without the need for any LiDAR scanner or any device designed for the three-dimensional reconstruction of the scene. We evaluated our approach on a public dataset of RGB images captured in an automotive scenario, with results confirming the feasibility of the proposed method.**

***Index Terms*—Deep Learning, CNN, Semantic Segmentation, Driver Attention, LiDAR.**

I. Introduction

Autonomous driving systems, driver assistance technologies, and systems for assessing the driver's attention can drastically reduce the number of crashes, thus increasing the road safety. To assist the driver's attention, technologies for estimating the distances of pedestrians/cyclists from the vehicle are widely adopted. These technologies are commonly based on acquisition devices composed of RGB cameras and hardware components designed for estimating a three-dimensional representation of the surrounding environment, like LiDAR scanners.

Estimating the distances of pedestrians/cyclists from the vehicle can be considered as a two-step problem: *i)* extracting the regions of the images/three-dimensional representations in which pedestrians/cyclists are present; *ii)* estimating the distances of the detected pedestrians/cyclists from the vehicle.

Extracting the regions of the images/three-dimensional representations with pedestrians/cyclists is a semantic segmentation problem, in which it is necessary to separate the different objects and classify them at the same time [1]. To perform the semantic segmentation, Deep Learning (DL) techniques and Convolutional Neural Networks (CNN) are currently the de-facto standard [2], due to their capability of recognizing shapes by analyzing spatial relationships between the pixels [3], [4]. In autonomous driving systems, the majority of the approaches perform the semantic segmentation by processing an RGB image captured under visible light illumination from a front-facing camera [5].

To estimate the distance from each object of the scene, many systems also use a LiDAR scanner, in addition to determining the type and boundaries of the object using RGB images [6]. In fact, the regions of the image extracted using semantic segmentation techniques are not sufficient for providing an effective instrument for assisting the driver's attention. By using also information from LiDAR scans, it is possible to determine the presence as well as the distance of a pedestrian/cyclist in the image [7] and assist the driver's attention by raising an alarm if such distance falls below a certain threshold.

While LiDAR scanners enable to obtain depth information, their installation on vehicles requires dedicated sensors, high costs, and specialized personnel for further maintenance. Furthermore, it is necessary to process the sparse measurements acquired by LiDAR scanners the to obtain a dense depth map of the surrounding environment [8]. On the other hand, low-cost, single-view, and visible-light cameras are increasingly

This work was supported in part by the EC within the H2020 Program under project MOSAICrOWN, by the Italian Ministry of Research within the PRIN program under project HOPE, and by the Università degli Studi di Milano under project "Artificial Intelligence for image analysis in Forensic Anthropology and Odontology". We thank the NVIDIA Corporation for the GPU donated within the project "Deep Learning and CUDA for advanced and less-constrained biometric systems".

(a) (b)

Fig. 1. Example of semantic segmentation: (a) original image; (b) corresponding semantic segmentation superimposed on the original image.

available to the general public for capturing RGB images [9]. Such images do not allow to directly obtain the depth information, but it is possible to process them with ad-hoc CNNs trained to estimate the depth without additional information [10]. However, while in the literature there are several methods for semantic segmentation and depth estimation using RGB images, there is no method that combines these two approaches to estimate the distance of specific objects in the scene.

In this paper, we propose an innovative method for assisting the driver's attention by automatically estimating the distances of pedestrians/cyclists from a single RGB image[1]. The proposed method is composed of two CNNs. The first CNN performs a semantic segmentation of the image, while the second CNN (AutoDepthNet) estimates a depth image of the scene. A pseudo-image representing the distances of the pedestrians/cyclists from the vehicle is obtained by fusing the estimated depth map in the regions extracted using the semantic segmentation technique.

The paper has the following contributions: *i)* it is the first method that combines semantic segmentation and depth estimation to process the distance of specific objects in the scene (in our case, pedestrians/cyclists) using a single RGB image; *ii)* differently from the methods in the literature proposed for estimating the depth from a single RGB image [11], [12], our work uses existing CNN architectures both for semantic segmentation and for depth estimation, without the need for ad-hoc architectures.

We evaluated the accuracy of the proposed method for a dataset of RGB images coupled by the corresponding LiDAR scans, that have only been used to extract the ground truth information for training the CNNs and assessing the accuracy of the proposed method. The obtained results proved the feasibility of the proposed method.

The paper is structured as follows. Section II introduces the related works. Section III describes the proposed method. Section IV presents the experimental evaluation. Finally, Section V concludes the work.

II. Related Works

This section reviews the related works, with a specific focus on DL-based methods for semantic segmentation, dense depth completion, and single-image depth estimation.

[1] http://iebil.di.unimi.it/DistPedCNN/index.htm

A. Semantic Segmentation

Following the increasing popularity of CNNs for object detection and classification, several architectures based on CNNs have been proposed to address the problem of semantic segmentation, which consists in segmenting as well as classifying the objects in an image [1]. Fig. 1 shows an example of a semantic segmentation performed using the CNN-based approach described in [13].

The method described in [14] is the first approach that uses a CNN for semantic segmentation. The method consists in modifying existing CNN architectures (e.g., AlexNet, VGG16) by removing the fully connected layer, with the purpose of having as output the image instead of an integer number representing the class. The first CNN architecture dedicated to semantic segmentation is described in [15]. The approach proposes an encoder-decoder architecture, in which the encoder is structured as a CNN for object classification and has the purpose of extracting the distinctive features of the objects in an image. Then, the decoder reconstructs the output of the encoder to obtain a pixelwise map describing the different classes.

The approaches proposed in [13] also adopt the encoder-decoder structure, at the same time introducing a spatial pyramid pooling, which improves the segmentation accuracy by processing the images at different scales. A method based on pyramid pooling has also been proposed in [16], which adopts a CNN based on the ResNet architecture [17].

Recent methods have been increasingly focusing on contextual information, with the purpose of enhancing the segmentation accuracy by considering the semantic relation between the pixels of the image [18].

B. Dense Depth Completion

Dense depth completion refers to the process of reconstructing a dense map from a set of sparse depth measurements, usually obtained from LiDAR scanners, by estimating the depth information also for the regions in the image that do not have an associated measurement. To this purpose, recent methods are mostly based on CNNs, such as the method described in [8], which uses an occluded depth map and the corresponding RGB image to obtain a dense depth map. Similarly, the methods proposed in [19]–[21] use CNNs to obtain dense depth maps by processing RGB images together with the sparse maps obtained by LiDAR scanners.

While many methods require both a sparse depth map and the corresponding RGB image for estimating a dense depth map, some approaches can use only a set of sparse depth measurements and a specific CNN architecture. Anyway, RGB images usually permit to use additional information and achieve better accuracy [22].

C. Single-Image Depth Estimation

Several methods in the literature have been proposed to compute a dense depth map using only a single RGB image as input, in the cases when no depth measurements of any kind are available (e.g., LiDAR data or multiple-view images). The

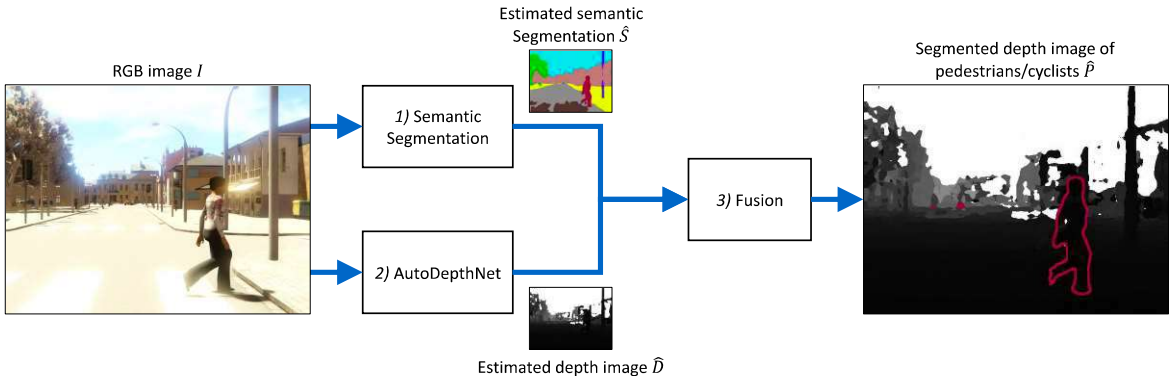


Fig. 2. Outline of the proposed methodology.

method described in [12] performs both a depth estimation and a semantic segmentation by means of an ad-hoc CNN. More recently, several methods rely on existing deeper network architectures (e.g., ResNet) to improve the depth estimation accuracy, such as the approach introduced in [23]. Rather than using deeper models, other approaches combine global and local processing for the same image [24].

To cope with the limited availability of real-world data encompassing labeled pairs of RGB images and dense depth maps, some methods consider synthetic databases for training, then apply domain transfer approaches to apply the trained CNN to real-world scenarios [11]. Other methods collect images from the internet, adopting structure-from-motion to estimate the depth ground truth, when multiple images of the same subject can be collected [10].

However, in the literature there are no DL-based methods for estimating depth maps of pedestrians/cyclists from a single image.

III. Proposed Method

The proposed method determines the presence of pedestrians/cyclists and their distances by using only a single RGB image. Our method is based on the following steps: *i)* a CNN performs the semantic segmentation of the RGB image; *ii)* the proposed AutoDepthNet estimates a dense depth map from the RGB image; *iii)* a fusion algorithm combines the semantic segmentation with the depth map to extract only the regions corresponding to pedestrians/cyclists from the depth map. Fig. 2 outlines the proposed method.

A. Semantic Segmentation of RGB Images

To perform the semantic segmentation of RGB images, we use a CNN with an encoder-decoder architecture. In particular, the encoder is a ResNet50 [17] and the decoder is the Pyramid Pooling Module (PPM) of PSPNet [16][2]. For each RGB image I in the dataset, we apply the CNN without performing any training or fine tuning. Fig. 3 shows an example of semantic segmentation $\hat{S}$.

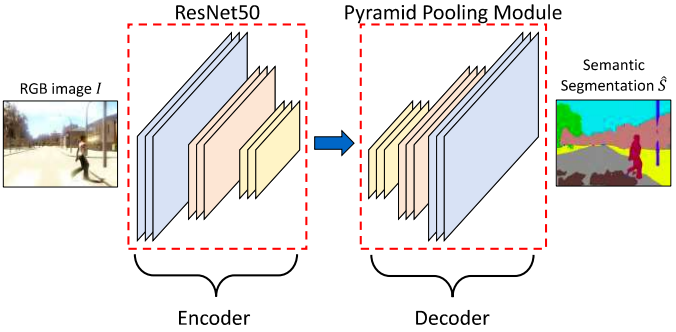


Fig. 3. Schema of the CNN used to perform the semantic segmentation $\hat{S}$ of an RGB image $\hat{I}$, consisting in a ResNet50 encoder and a Pyramid Pooling Module (PPM) decoder.

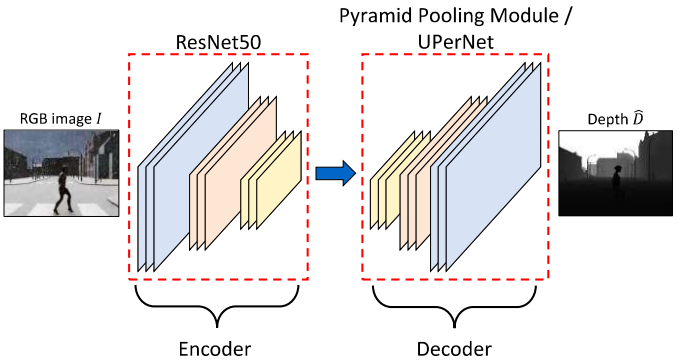


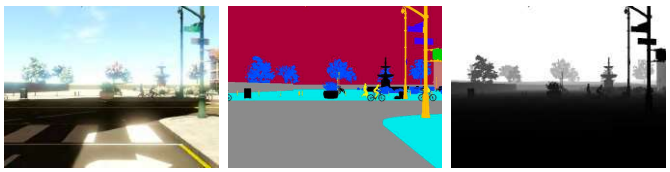
Fig. 4. Schema of the AutoDepthNet used to estimate a dense depth map $\hat{D}$ from an RGB image I. As encoder, the AutoDepthNet uses the ResNet50, while as decoder it can use two models: the Pyramid Pooling Module (PPM) or the UPerNet.

B. AutoDepthNet

We propose the AutoDepthNet to estimate a dense depth map $\hat{D}$ from a RGB image I. This CNN learns the relation between the image I and the ground truth depth map D, obtained using a LiDAR scanner.

AutoDepthNet is structured using the encoder-decoder architecture, with ResNet50 as encoder. We chose this architecture since CNNs based on the ResNet can achieve a high

[2]https://github.com/CSAILVision/semantic-segmentation-pytorch



(a) (b) (c)

Fig. 5. Examples of images in the SYNTHIA-AL database: (a) RGB image I; (b) semantic segmentation S; (c) depth image D.

accuracy in several fields and are implemented in numerous toolboxes for DL-based image processing [1], [25]. As decoder, it can use two models: the PPM [16] or the UPerNet [26], respectively.

We trained the AutoDepthNet considering as input a set of RGB images $\{I\}$ and the corresponding ground truth depth maps $\{D\}$. The trained AutoDepthNet can then estimate the dense depth image $\hat{D}$ from an input RGB image I. Fig. 4 shows an example of depth estimation $\hat{D}$.

C. Fusion

This step combines the semantic segmentation $\hat{S}$ with the estimated depth map $\hat{D}$ to compute a pseudo-image representing the distances of pedestrians/cyclists from the vehicle.

First, we process the estimated semantic segmentation $\hat{S}$ to create a mask M considering only the regions of I corresponding to pedestrians/cyclists. We compute M using the following formula:

$$M(x,y) = \begin{cases} 1 & \text{if } \hat{S}(x,y) = pedestrian \\ 0 & \text{otherwise} \end{cases} , \qquad (1)$$

where the class "pedestrian" can be expressed as a categorical variable, a gray intensity level, or an RGB color scheme, depending on the considered dataset and CNN implementation.

We compute the pseudo-image $\hat{P}$ as the values of the depth map $\hat{D}$ in the coordinates determined by the segmentation mask M. The regions of the coordinates that do not represent pedestrians/cyclists are set to 0.

Since the scene can present more than one pedestrian/cyclist, it is possible to estimate the distance of every single pedestrian/cyclist by considering the 8-connected regions of M and performing a computation similar to the one used to compute $\hat{P}$.

IV. Experimental Results

This section presents the performed experimental analysis, by introducing the used dataset, the evaluation procedure, the parameters, the procedure for training AutoDepthNet, and the obtained accuracy. Lastly, we perform a qualitative analysis on the results.

A. Database

To evaluate the accuracy of the proposed method, we used "The SYNTHetic collection of Imagery and Annotations (SYNTHIA)" [27]. In particular, we used the SYNTHIA-AL subset [28], which is a synthetic dataset resembling acquisitions performed in an automotive scenario. The dataset is generated at 25 FPS. Every sample includes an RGB image I, the corresponding semantic segmentation S, and the depth information encoded as an RGB image (R_D, G_D, B_D). All the images have a size $W \times H = 640 \times 480$. To obtain every ground truth depth map D from the depth information provided by the dataset, we processed the RGB image as follows [27]:

$$D = 5000 \frac{(R_D + G_D \cdot 256 + B_D \cdot 256 \cdot 256)}{(256 \cdot 256 \cdot 256 - 1)}. \qquad (2)$$

Fig. 5 shows examples of images of the SYNTHIA-AL database.

B. Evaluation Procedure

To compute the semantic segmentation of the RGB images, we applied a CNN with an encoder-decoder architecture, as described in Section III-A. The CNN is pretrained on the ADE20K scene parsing dataset [29], [30].

To train the AutoDepthNet, we considered the *train* subset of the SYNTHIA-AL database, which is composed by 178 video sequences, with $\approx$ 250 frames each. Due to limited time and computational constraints, we selected the first 21 sequences, for a total of $4,480$ frames. In particular, we divided the obtained set of frames by randomly selecting 80% of them for creating the training set and 20% for the validation set (for a total of 3584 and 896 frames, respectively).

To test the proposed method, we selected the first 21 sequences of the *test* subset of the SYNTHIA-AL database, for a total of 6725 frames.

C. Parameters and CNN Training of AutoDepthNet

We trained the AutoDepthNet using the Stochastic Gradient Descent (SGD) algorithm for a total of 200 epochs, with a batch size of 2. The encoder is pretrained on the ImageNet database, while the weights of the decoder are randomly initialized. We applied an initial learning rate $lr = 0.02$, with momentum $m = 0.9$. After each epoch, we reduced the learning rate as $lr' = lr^{0.9}$. We used the negative log likelihood loss as a loss function. The target classes are the integer numbers in the range $[0, 255]$, representing all the possible intensity values of the depth image D. The size of the input layer is 640×480. At the end of each epoch, we evaluated the accuracy using the validation subset. Lastly, we selected the values of the weights for which we obtained the highest accuracy on the validation subset.

D. Accuracy

To evaluate the accuracy obtained in estimating the distances of pedestrians/cyclists, we considered the regression error as a figure of merit. Specifically, for each frame i, we computed the pedestrian/cyclist regression error (Ep) as follows:

$$\text{Ep}_i = \frac{\sum_{j=1}^{n} \left| \hat{P}_i(j) - P_i(j) \right|}{\sum_{j=1}^{n} M_i(j)} , \qquad (3)$$

TABLE I
ACCURACY OF THE PROPOSED METHOD FOR ESTIMATING THE DISTANCES PEDESTRIANS/CYCLISTS

CNN	Mean Ep (%)
ResNet50 [17] + PPM [16]	0.01
ResNet50 [17] + UPerNet [26]	0.02

where n is the total number of pixels and P_i corresponds to the pseudo-image representing the real distances of pedestrians/cyclists, which is computed by using the algorithm applied to obtain $\hat{P}_i$, but considering the ground truth dept map D_i instead of $\hat{D}_i$.

Table I summarizes the results obtained by the proposed method for estimating the distances of pedestrians/cyclists. The configuration based on the PPM achieved the best accuracy, with a mean Ep $< 0.1\%$. The achieved accuracy is satisfactory and hence the proposed method could represent an additional driver assistance tool to be integrated in automotive application.

E. Qualitative Analysis

We perform a qualitative analysis of the results by visually comparing the original RGB images, the results of the semantic segmentation, and the output of the AutoDepthNet.

Fig. 6 shows examples of the results of the semantic segmentation and the depth estimation, by comparing the original RGB image I, the obtained segmentation $\hat{S}$, and the estimated depth image $\hat{D}$ obtained using AutoDepthNet. In the estimated depth images $\hat{D}$, the class "pedestrian" is outlined in red. From the figure, it is possible to observe that the CNN achieves a satisfactory accuracy in detecting the position and estimating the depth of pedestrians/cyclists in the RGB images.

V. CONCLUSIONS

In this paper we proposed a method for assisting the driver's attention by at the same time detecting the presence of pedestrians/cyclists in the scene and estimating their distance from the vehicle, using only an RGB image. The method performs the semantic segmentation of the scene using a pre-trained CNN and extracts the "pedestrian" class. Then, it applies the proposed AutoDepthNet to estimate a dense depth map from the input RGB image. Lastly, our method performs the fusion by combining the output of the AutoDepthNet with that of the semantic segmentation to estimate the distances of the pedestrians/cyclists. We evaluated the proposed approach on the SYNTHIA-AL dataset, containing both RGB images and LiDAR scans captured in an automotive scenario. We obtained satisfactory results, with a mean regression error $< 0.1\%$ in estimating the average distance of the pedestrian/cyclist. Future works should consider databases captured in real and heterogeneous applications scenarios and CNN architectures able to process data captured in the different operational and environmental conditions typical of real-life conditions.

REFERENCES

[1] F. Lateef and Y. Ruichek, "Survey on semantic segmentation using deep learning techniques," *Neurocomputing*, vol. 338, pp. 321–348, 2019.

[2] L. Chan, M. Hosseini, and K. Plataniotis, "A comprehensive analysis of weakly-supervised semantic segmentation in different image domains," *Int. J. Comput. Vis.*, 2020.

[3] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018.

[4] R. Donida Labati, A. Genovese, E. Muñoz, V. Piuri, and F. Scotti, "A novel pore extraction method for heterogeneous fingerprint images using Convolutional Neural Networks," *Pattern Recognition Letters*, vol. 113, pp. 58–66, 2018.

[5] W. Zhou, J. S. Berrio, S. Worrall, and E. Nebot, "Automated evaluation of semantic segmentation robustness for autonomous driving," *IEEE Trans. on Intelligent Transportation Systems*, vol. 21, no. 5, pp. 1951–1963, 2020.

[6] K. El Madawi, H. Rashed, A. El Sallab, O. Nasr, H. Kamel, and S. Yogamani, "RGB and LiDAR fusion based 3D semantic segmentation for autonomous driving," in *Proc. of the 2019 IEEE Intelligent Transportation Systems Conf. (ITSC)*, 2019, pp. 7–12.

[7] L. Pang, Z. Cao, J. Yu, S. Liang, X. Chen, and W. Zhang, "An efficient 3D pedestrian detector with calibrated RGB camera and 3D LiDAR," in *Proc. of the 2019 IEEE Int. Conf. on Robotics and Biomimetics (ROBIO)*, 2019, pp. 2902–2907.

[8] Y. Zhang and T. Funkhouser, "Deep depth completion of a single RGB-D image," in *Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 175–185.

[9] Techradar, "Best dash cam 2020: 10 car-ready cameras for peace of mind," 2020. [Online]. Available: https://www.techradar.com/best/best-dash-cam

[10] Z. Li and N. Snavely, "MegaDepth: Learning single-view depth prediction from internet photos," in *Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2041–2050.

[11] A. Atapour-Abarghouei and T. P. Breckon, "Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer," in *Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2800–2810.

[12] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. of the 2015 IEEE Int. Conf. on Computer Vision (ICCV)*, 2015, pp. 2650–2658.

[13] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation." in *Proc. of the European Conference on Computer Vision (ECCV)*, 2018, pp. 833–851.

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[15] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[16] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 6230–6239.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[18] Z. Huang, X. Wang, Y. Wei, L. Huang, H. Shi, W. Liu, and T. S. Huang, "Ccnet: Criss-cross attention for semantic segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2020.

[19] Z. Chen, V. Badrinarayanan, G. Drozdov, and A. Rabinovich, "Estimating depth from RGB and sparse sensing," in *Proc. of the European Conf. on Computer Vision (ECCV)*, 2018, pp. 176–192.

[20] A. Eldesokey, M. Felsberg, and F. S. Khan, "Confidence propagation through CNNs for guided sparse depth regression," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2423–2436, 2020.

[21] P. Hambarde and S. Murala, "S2DNet: Depth estimation from single image and sparse samples," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 806–817, 2020.

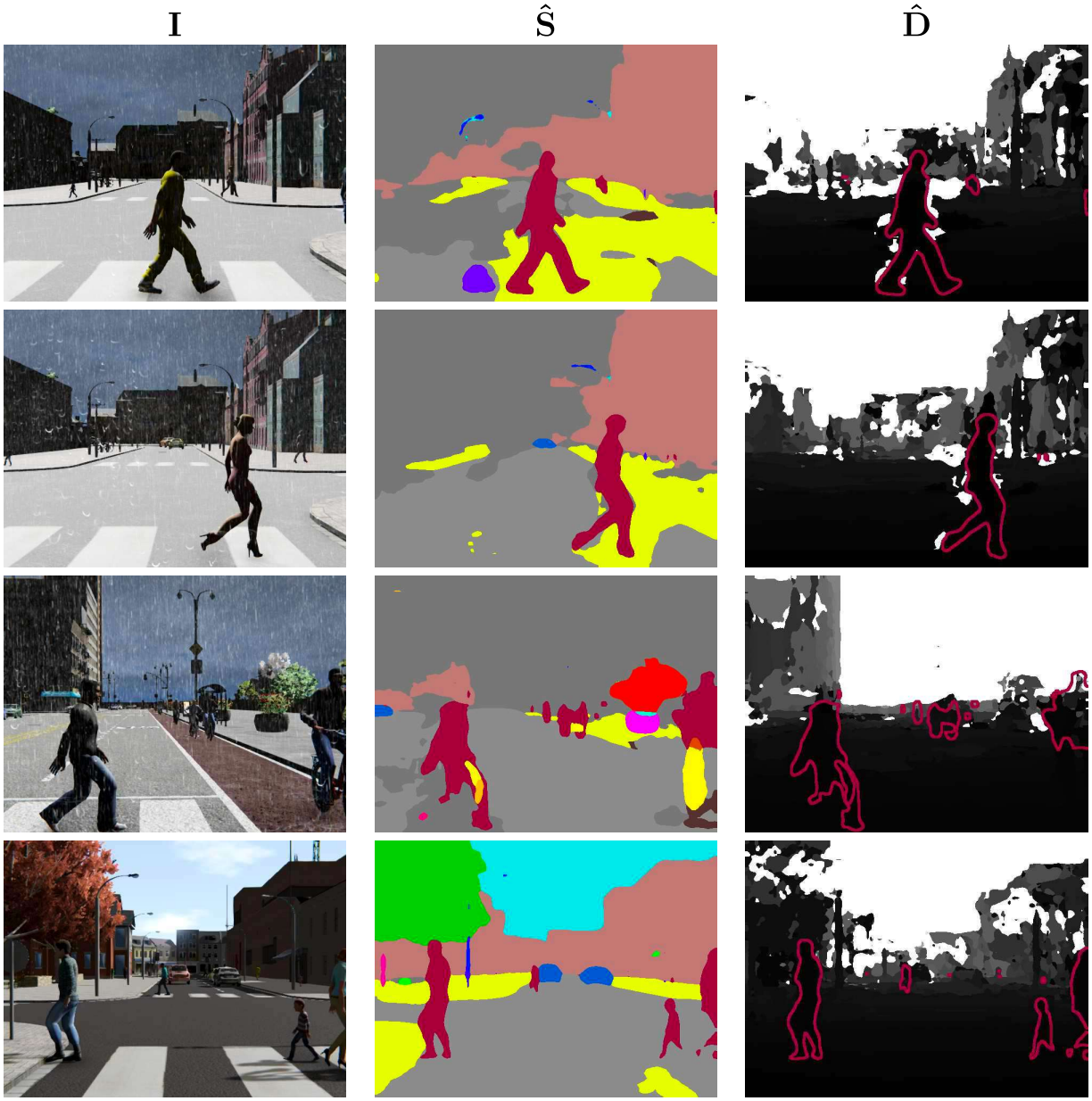


Fig. 6. Qualitative analysis of the semantic segmentation and the estimated depth. First column: original RGB images I. Second column: obtained segmentations $\hat{S}$ (pedestrians/cyclists are shown in red). Third column: estimated depth images $\hat{D}$ (pedestrians/cyclists are outlined in red). It is possible to observe that the CNN achieves a satisfactory accuracy in detecting the position and estimating the depth of pedestrians/cyclists.

[22] Z. Huang, J. Fan, S. Cheng, S. Yi, X. Wang, and H. Li, "HMS-Net: Hierarchical multi-scale sparsity-invariant network for sparse depth completion," *IEEE Trans. on Image Processing*, vol. 29, pp. 3429–3441, 2020.

[23] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully Convolutional Residual Networks," in *Proc. of the 2016 Fourth Int. Conf. on 3D Vision (3DV)*, 2016, pp. 239–248.

[24] Y. Kim, H. Jung, D. Min, and K. Sohn, "Deep monocular depth estimation via integration of global and local predictions," *IEEE Trans. on Image Processing*, vol. 27, no. 8, pp. 4131–4144, 2018.

[25] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep Learning for generic object detection: A survey," *Int. J. Comput. Vis.*, 2019.

[26] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. of the European Conference on Computer Vision (ECCV)*, 2018, pp. 432–448.

[27] Universitat Autònoma de Barcelona, "The SYNTHetic collection of Imagery and Annotations (SYNTHIA) dataset," 2017. [Online]. Available: http://synthia-dataset.net/

[28] J. Zolfaghari Bengar, A. Gonzalez-Garcia, G. Villalonga, B. Raducanu, H. Habibi Aghdam, M. Mozerov, A. M. Lopez, and J. van de Weijer, "Temporal coherence for active learning in videos," in *Proc. of the 2019 IEEE/CVF Int. Conf. on Computer Vision Workshop (ICCVW)*, 2019, pp. 914–923.

[29] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ADE20K dataset," *Int. J. Comput. Vis.*, vol. 127, p. 302–321, 2019.

[30] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5122–5130.