Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Semantic Segmentation Model for Road Scene Based on Encoder-Decoder Structure

Yuanzhe Peng[1,2,3,4] ,Weichao Han[1,3,4]and Yongsheng Ou[1,3,4]

***Abstract*— Semantic segmentation as a pixel-wise segmentation task provides rich object information, which is an important research topic in robotic perception. It has been widely applied in many fields, such as autonomous driving and robot navigation. In the application of understanding road scene, the semantic segmentation model should accurately describe the appearance and shape of different categories of objects. In addition, the semantic segmentation model need to understand the spatial relationships between different categories. In order to improve the performance of semantic segmentation model for road scene, we present a model based on encoder-decoder structure with dilated convolution. We apply this model on the Cityscapes dataset and compare it with other classical models. To assess performance, we rely on the standard Jaccard Index IoU (Intersection over Union) and mIoU (mean Intersection over Union). The experimental results verify that this model can effectively improve the performance of semantic segmentation and meet the requirements for road scene.**

***Index Terms*— Semantic Segmentation, Road Scene, Encoder-Decoder Structure, mIoU.**

I. Introduction

Semantic segmentation as a pixel-wise segmentation task provides rich object information, which is an important research topic in robotic perception. It has been widely applied in many fields, such as autonomous driving and robot navigation. In the automatic driving task, the semantic segmentation technology can be used to extract the road information and help the automatic driving system to make decisions after the current road image is obtained. Aiming at the problem of semantic segmentation for road scene, traditional methods cannot achieve accurate semantic segmentation results. With the improvement of computer processing power, we are increasingly using deep learning to solve the problem of semantic segmentation and have made great progress.

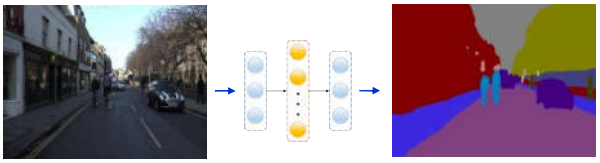
*This work was jointly supported by National Natural Science Foundation of China (Grant No.U1613210), Guangdong Special Support Program (2017TX04X265), Science and Technology Planning Project of Guangdong Province (2019B090915002), and Shenzhen Fundamental Research Program (JCYJ20170413165528221). (*Corresponding author: Yongsheng Ou.*)

[1]Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences (CAS), Shenzhen 518055, China.

[2]University of Chinese Academy of Sciences, Beijing 100049, China.

[3]Guangdong Provincial Key Lab of Robotics and Intelligent System, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China.

[4]CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Shenzhen 518055, China.



(a) The input image (b) Encoder-Decoder structure (c) The output image

Fig. 1. The task of semantic segmentation model design for road scene based on encoder-decoder structure.

Semantic segmentation involves two aspects: segmentation and classification. In the application of road scene, the semantic segmentation model should be able to accurately describe the appearance and shape of different categories of objects, and understand the spatial relationship between different categories, such as vehicle lanes and sidewalks. In a typical road scene, most pixels belong to the large target category, such as roads, buildings and so on. They are required for the model to maintain good smoothness and regional consistency in image segmentation results. Moreover, the segmentation model needs to have a good ability to identify the shape of the object and be able to accurately identify small objects. Therefore, the task of semantic segmentation model design for road scene has the following specific problems:

(1) The segmentation results should ensure the integrity and regional consistency of large targets, such as roads and buildings.
(2) The segmentation results should accurately divide the edges of adjacent objects.
(3) The segmentation results should have a good recognition rate for small targets, such as pedestrians and bicycles.

In this paper, in order to improve the performance of semantic segmentation for road scene, we mainly present a model based on encoder-decoder structure. On the premise of ensuring real-time requirement, the performance of semantic segmentation model is effectively improved, due to the help of encoder-decoder structure using dilated convolution. Finally, we apply this model on the Cityscapes dataset and compare it with other classical models. The experimental results verify that this model can effectively improve the performance of semantic segmentation and meet the requirements for road scene.

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

The remainder of this paper is structured as follows. In Section II some related work is reviewed, including early semantic segmentation methods and some classic methods based on deep neural networks. In this section, we also emphasize that our original contributions are different from previous work. In Section III we formally propose our method for semantic segmentation. We first explain the principle of using dilated convolution based on encoder-decoder structure and then we describe our model in detail. In Section IV experiments have been provided to verify the validity of our semantic segmentation model. Finally, Section V concludes the paper and makes a plan for future work.

II. Related Work

In this section, some semantic segmentation methods will be reviewed. Early semantic segmentation methods relied on handcrafted features. They used Random Decision Forest [1] or Boosting [2] to predict the class probabilities and used probabilistic models known as Conditional Random Fields (CRFs) to handle uncertainties and propagate contextual information across the images. In recent years, Convolutional Neural Networks (CNNs) have made great progress in computer vision thanks to large-scale training datasets and high-performance Graphics Processing Unit (GPU). In addition, excellent open-source deep learning frameworks like Caffe, MXNet, and Tensorflow boosted the development of algorithms. The powerful deep neural network greatly reduced the classification error on ImageNet [3], and semantic segmentation made great progress in this process.

It is crucial to incorporate context information in relevant image regions when making a prediction for the task of semantic segmentation. An extensive receptive field is usually desirable to capture the entire useful information. The downsampling operation can multiply the size of the receptive field. After down-sampling operation, a large number of low-level visual features are lost, which is easy to damage the spatial structure of the scene. Dilated convolution or Atrous convolution [4] [5] is proposed to alleviate this problem by enlarging the receptive field without reducing the spatial resolution. It increases the size of the kernel by introducing zeros into the convolution filter without increasing the parameters. On the other hand, some networks like ResNet [6] theoretically have a large receptive field, even larger than the input image because of the significantly increased depth. But unfortunately, the effective receptive field of a network is much smaller than the theoretical one. Instead of handcrafted designing models, the deformable model [7] learns the shapes of convolution filters conditioned on an input feature map. The receptive field and the spatial sampling locations are adapted according to the scale and shape of objects. The experimental results verify that it is feasible and effective to learn geometric transformation of visual task in neural network [8].

Most previous semantic segmentation frameworks are based on Fully Convolutional Networks (FCN) [9]. FCN successfully improved the performance of semantic segmentation by adapting classification networks into fully convolutional networks. In other words, the fully connected layer of the classification model is replaced by the convolution layer. However, FCN has the following problems: (i) Mismatched Relationship: Contextual matching is important for understanding complex scene; (ii) Confusion Categories: Many labels have associations that can confuse categories; (iii) Inconspicuous Classes: The model may ignore small things and large things may exceed the FCN range, leading to discontinuous predictions.

In order to solve the problem that FCN cannot effectively incorporate suitable global features, Pyramid Scene Parsing Network (PSPNet) [10] is proposed. It can integrate local and global information together and fuse the appropriate global features. Experimental results verify that PSPNet performs well on a plenty of datasets. In addition, Efficient Residual Factorized ConvNet (ERFNet) [11] proposes a deep architecture that is able to run in real-time while providing accurate semantic segmentation. The core of its architecture is a novel layer that uses residual connections and factorized convolutions in order to remain efficient while retaining remarkable accuracy.

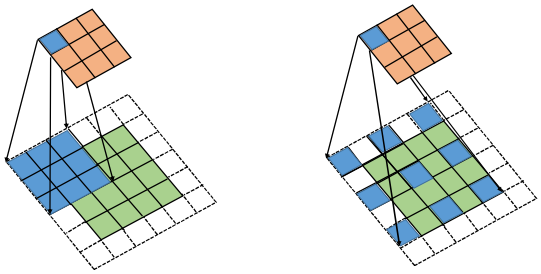
Based on the past achievements, in this paper, we mainly present a semantic segmentation model for road scene based on encoder-decoder structure with dilated convolution. Different from previous work, our original contributions are:

(1) In order to integrate multi-scale information, we use encoder-decoder structure and we can arbitrarily control the resolution of extracted encoder features by dilated convolution to balance accuracy and runtime.
(2) Compare with other classical semantic segmentation models, our proposed model attains a state-of-art performance on Cityscapes dataset and is suitable for road scene.

III. Methodology

A. Encoder-Decoder Structure with Dilated Convolution

Dilated convolutions, also known as atrous convolutions, have been widely explored in deep convolutional neural networks (DCNNs) for various dense prediction tasks. Dilation upsamples convolutional filters by inserting zeros at both left and right sides. Compared with the original normal convolution operation, dilated convolution uses an extra dilation rate to control the interval of the convolution kernel inserting zeros, as illustrated in Fig. 2. It enlarges the receptive field but does not require training extra parameters in DCNNs. Dilated convolutions can be used in cascade to build multi-layer networks. Another advantage of dilated convolutions is that they do not reduce the spatial resolution of responses.



(a) Standard Convolution with a 3×3 kernel (and padding)

(b) Dilated Convolution with a 3×3 kernel and dilation rate 2

Fig. 2. Dilated convolution uses an extra dilation rate to enlarge the receptive field.

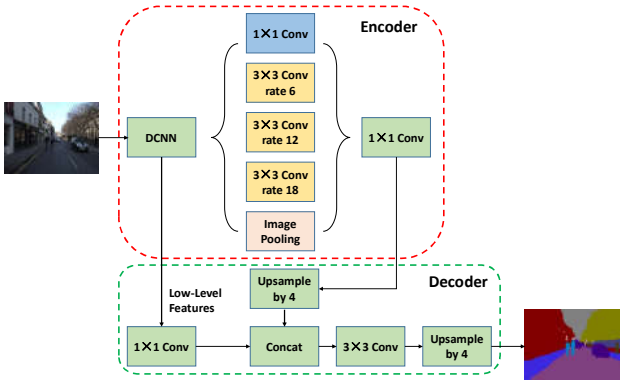


Fig. 3. The semantic segmentation model based on encoder-decoder structure.

This is a key difference from downsampling layers, such as pooling layers or convolutions with stride larger than one.

In the problem of an image needs to rely on global information, dilated convolution can be well applied, such as image semantic segmentation. Dilated convolution can explicitly set the size of the feature graph of the deep convolutional network, it can also change the size of the receptive field by setting the convolution kernel, see Fig. 2.

The concrete implementation of dilated convolution can start from the example of a one-dimensional signal. For a one-dimensional input signal x[i], convolution operation is performed using a filter w[k] of length K and dilation rate r. It can be expressed as:

$$y[i] = \sum_{k=1}^{K} x[i + r \cdot k] w[k] \tag{1}$$

Through the description of the algorithm of dilated convolution, it can be seen that dilated convolution is simple to realize: zero values are inserted into the convolution kernel in a certain span, and only those non-zero values are really involved in the weighted sum operation in the convolution, and dilated convolution will not bring the increase of calculation. Although the pooling layer increases the receptive field and introduces invariance into the network to improve the classification performance of the network, it ignores certain spatial information and also causes the decrease of the resolution of the feature map. For the semantic segmentation task, the greater receptive field obtained by pooling layer is contradictory to the reduced resolution. Therefore, it is feasible to remove some pooling layers and introduce empty convolution to obtain the greater receptive field and maintain the image resolution.

B. Model Design for Road Scene

Model design has always been the most important part of semantic segmentation research. In recent years, the development of semantic segmentation is mainly to optimize the model design.

Deeplabv3+ is the fourth version in the Deeplab series and an upgrade to Deeplabv3. Deeplabv3+ uses dilated convolution to reduce the reduction of sampling rate while keeping the original training parameters unchanged. At the same time, Deeplabv3+ also uses dilated convolution to acquire high-level features at different scales. The segmentation results are obtained by fusion with the low-level features. In the semantic segmentation task, the spatial pyramid pooling module (SPP) can capture multi-scale information, and the encoder-decoder structure can better recover the edge information of the object. Deeplabv3+ takes the original Deeplabv3 as encoder and adds decoder to get a new model, see Fig. 3.

But semantic segmentation requires a lot of computation while it should be minimized for road scene. In order to reduce the computation generated by semantic segmentation, there are generally two methods: (i) reducing the image size and (ii) reducing the complexity of the model. Reducing the size of the image will most directly reduce the amount of computation, but the image will lose a lot of detail and affect the accuracy. Reducing the complexity of the model will weaken the feature extraction ability of the model and affect the segmentation accuracy. Therefore, we present a Road Scene (RS) Deeplabv3+ framework based on encoder-decoder structure in order to minimize the negative effects of both methods according to the semantic segmentation requirements of road scene, see Fig. 4.

In the encoder network, we eliminate the full connection layer structure to reduce the network parameters and retain the maximum pooled index before the pooling operation. The maximum pooled index is reused in the upsampling process of decoder network, which reduces the memory overhead while preserving the image boundary information.

Block1 outputs the feature of four times downsampling, block2 outputs the feature of sixteen times downsampling, and C represents the number of channels of the input feature. RS Deeplabv3+ uses three 3×3 dilated convolution with dilation rate of 6, 12, 18 to extract high-level features of

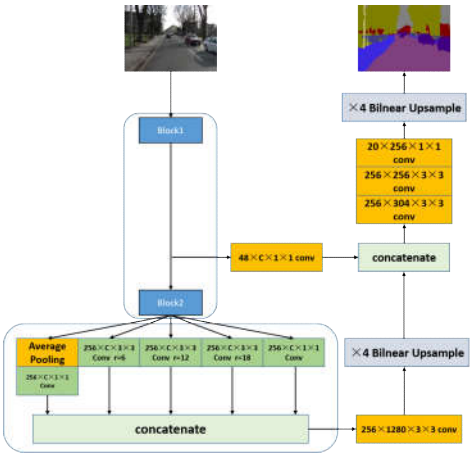


Fig. 4. The framework of Road Scene(RS) Deeplabv3+.

different scales in Atrous Spatial Pyramid Pooling (ASPP). It corresponds to the size of the receptive field, which is 13×13, 25×25 and 37×37 respectively. Not only the dilated convolution is used in ASPP, but also a 1×1 convolution is used to obtain the original scale features and an average pooling is used to obtain the features with global information. After fusing these high level features with different scales, this feature uses four times bilinear interpolation upsampling and fuses with four times downsampling features. Finally, the semantic segmentation result is obtained by three times convolution operations and four times bilinear interpolation upsampling.

IV. Experiments

In order to verify the validity of our proposed semantic segmentation model for road scene, we conducted several experiments. First, we selected Cityscapes datset, which is suitable for the semantic segmentation experiment of road scene. Second, we scaled down the input original images and label images, then we trained the model according to the experimental steps. Finally, we used the mean Intersection over Union (mIoU) as an important indicator to measure the performance of our proposed semantic segmentation model.

A. Cityscapes Datset

Semantic segmentation task for road scene related to autonomous driving, Cityscapes [12] can be said to be the most authoritative dataset at present. The images in Cityscapes dataset come from a large number of video sequences recorded from streets in different cities, covering spring, summer and fall in 50 cities. Through advanced on-board photography technology and post-processing, Cityscapes resulted in 5000 fine images with high-quality pixel-level labels and 20000 additional images with coarse labels. Among the fine images, the number of training samples is 2975, the number of verification samples is 500, and the number of test samples is 1525. In the label of Cityscapes dataset, a total of 30 visual categories have been defined, which can be classified into 8 categories: flat ground, building, nature, vehicle, sky, object, human and others.

TABLE I
Absolute Number And Density Of Annotated Pixels For Cityscapes, CamVid, DUS And KITTI

Dataset	Pixel number (10^9)	Label density (%)
Cityscapes(fine)	9.43	97.1
Cityscapes(coarse)	26.0	67.5
Camvid	0.62	96.2
DUS	0.14	63.0
KITTI	0.23	88.9

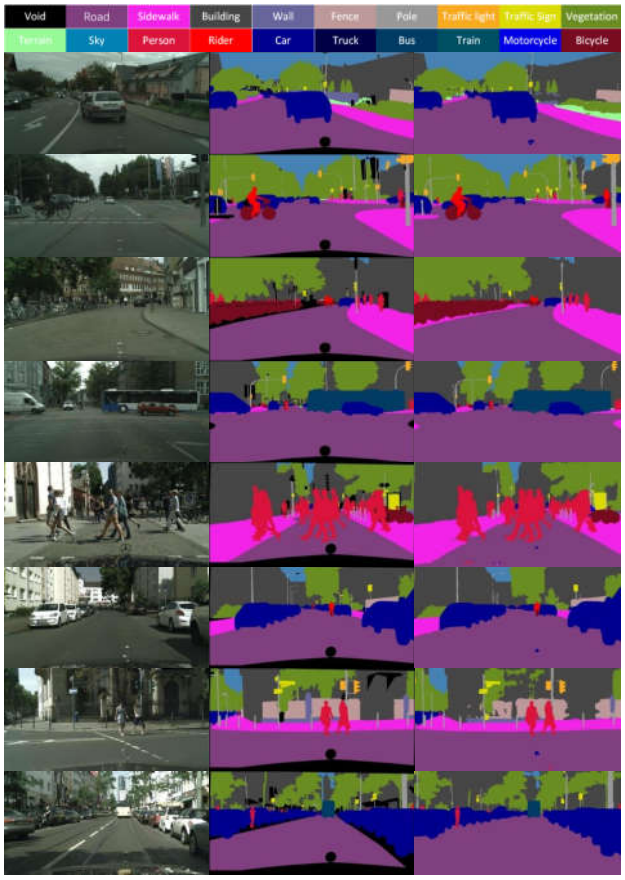
The area covered by Cityscapes dataset is much larger than that of previous road scene segmentation datasets that only contain images from individual cities, such as CamVid [13], DUS [14] and KITTI [15]. In terms of absolute and relative numbers of semantically annotated pixels (training, validation, and test data), Cityscapes compares favorably to CamVid, DUS, and KITTI with up to two orders of magnitude more annotated pixels, as shown in TABLE I.

It can be seen that Cityscapes dataset surpasses other dataset in terms of size and scene complexity, and can accommodate more road scenes in the real world. Therefore, the algorithm applied in Cityscapes dataset needs to take into account a wider range of scales and more complex conditions, which has a great practical value.

B. Experiment with Our Model and Compare It with Other Classical Models

In this paper, we adopts the strategy of training encoder part and decoder part separately, which can reduce the complexity of single training task and get better training results. The steps of this experiment are as follows:

(1) Data preprocessing. In the training stage of encoder and the testing stage on validation set, the resolution of the input original image and label image are scaled down to 1024×512 and 128×64. Then for encoder-decoder, the resolution of the input original image and the label image are both scaled down to 1024×512.
(2) Train the encoder. The encoder is trained on the fine image sets of Cityscapes dataset, and a convolutional layer is added after the encoder in order to get the segmentation result.
(3) Train encoder-decoder structure. Remove the last convolutional layer of the encoder model that has been trained in the previous step, and then initialize the encoder part of the whole encoder-decoder structure with other parts, so as to train the encoder-decoder structure.
(4) Dataset test. Run the trained model on the Cityscapes dataset, scale down the segmentation result to the initial



(a) The input image (b) The label (c) The output image

Fig. 5. The semantic segmentation results based on encoder-decoder structure.

size of the dataset image using the nearest neighbor interpolation method to match the label image of the server, and then submit the results to evaluate the performance of the model.

The semantic segmentation results based on our model with encoder-decoder structure are shown in Fig. 5.

In the field of image segmentation based on deep learning in computer vision, mIoU is an important index to measure the performance of image segmentation. In the field of object detection, it is often used to constrain the coincidence degree between object frame and prediction frame. In the image semantic segmentation task, it can be understood as the ratio between the intersection of the Prediction pixel set P and the corresponding Ground Truth (GT) pixel set GT and their union.

$$IoU = \frac{|P \cap GT|}{|P \cup GT|} \tag{2}$$

Formula (3) is usually used in calculation,

TABLE II
TRUE- FALSE AND POSITIVE- NEGATIVE CONDITION OF THE LABEL AND PREDICTION

Condition	Prediction is true	Prediction is false
The label is true	True Positive, TP	False Positive, FP
The label is false	False Negative, FN	True Negative, TN

$$IoU = \frac{TP}{TP+FP+FN} \tag{3}$$

TP, FP and FN are the number of true positive, false positive and false negative pixels of the predicted result, which can be referred to TABLE II in calculation. It can be seen that formula (3) is equivalent to formula (2), and IoU is 1 when the prediction result of this category is completely coincide with the label result.

mIoU is the average of the IoU of each category in the segmentation result,

$$mIoU = \frac{1}{C}\sum_{i=1}^{C} IoU_i \tag{4}$$

where C is the number of categories of the label results.

In this paper, we compare experimental results between classical models and our model based on encoder-decoder structure using dilated convolution, see Table III. The semantic segmentation model performance is evaluated by mIoU.

C. Results

By comparing and analysing the experimental results, our model has significantly improved IoU in several categories, such as road, sidewalk, sky, etc. However, the IoU value of semantic segmentation decreased when our model encountered smaller target categories (such as traffic lights) and dense crowds. This is because our model completely based on encoder-decoder structure with dilated convolution will create new problems:

(1) The gridding effect. Multiple stacks of multiple convolution kernels with the same dilation rate will result in some pixels in the grid not participating in the operation from beginning to end. In other words, these pixels have no effect during calculation, which has a negative impact for the prediction of pixel level.

(2) Long-ranged information might be not relevant. Although dilated convolution guarantees a larger receptive field with the same parameters, it is defective for some small objects that themselves do not need such a large receptive field. How to deal with the relations between objects of different sizes at the same time is the key to design a dilated convolution network.

However, most importantly, the experimental results verify that our mIoU value is higher than other classical semantic segmentation models, which proves that our model can

TABLE III
COMPARISON RESULTS OF EACH MODEL ON THE CITYSCAPES DATASET

Model	Road	Sidewalk	Building	Wall	Fence	Pole	Traffic Light	Car	Bicycle	Person	Sky	mIoU(%)
FCN-8S [9]	70.2	72.5	71.6	68.5	66.2	58.6	47.2	69.5	62.8	66.6	75.6	66.3
ERFNet [11]	71.3	74.1	70.6	69.8	64.2	60.7	56.3	70.1	66.4	67.8	77.8	68.1
ResNet-38 [6]	84.7	85.6	82.1	75.6	79.5	76.3	72.1	82.6	77.9	79.7	88.3	80.4
PSPNet [10]	83.6	82.7	84.9	85.4	80.1	**78.7**	60.5	83.7	78.2	79.9	88.6	80.6
Deeplabv3 [5]	82.0	81.7	82.1	81.6	**80.4**	77.9	**77.5**	84.8	80.7	78.4	87.2	81.3
Ours	**85.1**	**85.7**	**85.9**	**85.5**	78.5	75.1	67.8	**85.1**	**80.9**	**80.1**	**89.1**	**81.7**

improve the overall segmentation ability and help achieve better segmentation effect.

V. CONCLUSION AND DISCUSSION

Semantic segmentation as a pixel-wise segmentation task provides rich object information, which is an important research topic in robotic perception. It has been widely applied in many fields, such as autonomous driving and robot navigation. Aiming at the problem of semantic segmentation for road scene, traditional methods cannot achieve accurate semantic segmentation results. In this paper, in order to improve the performance of semantic segmentation for road scene, we mainly present a model based on encoder-decoder structure. On the premise of ensuring real-time requirement, the mIoU value of semantic segmentation model is effectively improved, due to the help of encoder-decoder structure using dilated convolution. Finally, we apply this model on the Cityscapes dataset and compare it with other classical models. The experimental results verify that this model can effectively improve the performance of semantic segmentation and meet the requirements for road scene.

Nevertheless, our proposed model of semantic segmentation based on encoder-decoder structure for road scene still needs extensions. When the input image quality is poor, such as bad weather, dark image and so on, there will be some false segmentation in the results. However, they can be avoided, such as a car in the middle of the road. Prior knowledge can be added to the model in the future work to reduce the generation of false segmentation and improve the segmentation quality. In addition, the model proposed in this paper is mainly aimed at image segmentation. For the video datasets, the image of each frame is also segmented separately, and the timing information in video is not considered. Therefore, time sequence information can be combined to improve the segmentation quality in the future work.

REFERENCES

[1] Tin Kam Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, Aug 1998.

[2] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, Dec 2001, pp. I–I.

[3] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 248–255.

[4] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 636–644.

[5] L. 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, April 2018.

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

[7] R. Girshick, F. Iandola, T. Darrell, and J. Malik, "Deformable part models are convolutional neural networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 437–446.

[8] M. Holden, "A review of geometric transformations for nonrigid body registration," *IEEE Transactions on Medical Imaging*, vol. 27, no. 1, pp. 111–128, Jan 2008.

[9] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, p. 640C651, Apr 2017. [Online]. Available: http://dx.doi.org/10.1109/tpami.2016.2572683

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

[11] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, Jan 2018.

[12] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," *CoRR*, vol. abs/1604.01685, 2016.

[13] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88 – 97, 2009, video-based Object and Event Analysis.

[14] T. Scharwächter, M. Enzweiler, U. Franke, and S. Roth, "Stixmantics: A medium-level model for real-time semantic scene understanding," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 533–548.

[15] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.