Lane Segmentation Based on Convolution Neural Network and Conditional Random Field

Qianghuang Huang*, Fuxin Xu and Su Wang

Central South University 932 Lushan South Road, Yuelu District, Changsha, China

*Corresponding author: hqh3511@csu.edu.cn

Keywords: full convolution neural network-conditional random field-pattern recognition-lane segmentation.

Abstract: Common lane segmentation methods have disadvantages such as complicated pretreatment process, rough detail segmentation, low robustness and large computation. In view of these shortcomings, we propose a new lane segmentation method that combine full convolution neural network and conditional random field (CRF). The front end is feature represented by U-Net network, and the image is segmented into two parts, lane and background, while the back end adopts CRF for smooth segmentation of the lane edge. The main improvements made in this paper include: in order to improve the accuracy and robustness of lane segmentation method, we give up migration learning and adopt the latest, largest and most complete database -BDD100K for model training; LeakyReLU activation function is used to replace the original relu function to avoid parameter necrosis. In order to make the segmentation method able to deal with the complex lane environment, we add the CRF to the back end for edge segmentation. We achieve 98.86% ACU and the processing speed of 0.09s per image, which is improved compared with other methods. The test results prove that our method is effective.

1. Introduction

In recent years, with the continuous development of social technology, automatic driving began to enter our lives. Lane identification and segmentation are the foundation of the field of autonomous driving. Accurate and real-time lane recognition algorithms not only keep cars on the right road but also provide other information, such as lane markings, pedestrians, vehicles and even other animals. Traditional road segmentation methods are usually based on texture, color, shape and other inherent attributes of the image. Traditional lane segmentation methods are generally through image preprocessing, edge detection, Hough transformation, lane line fitting and lane segmentation. Although the segmentation algorithm based on lane texture represented by text is of high real-time performance (Graovac S et al,2012), it has disadvantages of complex pretreatment and poor robustness. Moreover, the algorithm is not effective in detecting fuzzy lane edges. There are k-nearest neighbor algorithm and unsupervised learning k-means algorithm based on image color. These algorithms rely on the color information of the image itself and is sensitive to shadow, potholes, water accumulation and weather conditions. Therefore, the algorithm has the disadvantages of high requirement for data set preprocessing and poor robustness.

With the continuous improvement of computer computing power, convolutional neural network (CNN) (Lecun Y et al,1998), as the most popular learning tool, has been applied in the field of natural language processing and computer vision on a large scale. Compared with the traditional lane segmentation method, convolutional neural network directly processes two-dimensional images and avoids the complex feature extraction process, which has the advantages of autonomous learning. The convolutional neural network adopts the structure of convolutional layer-pooling layer, which makes CNN have a large sense field and leads to a rough edge of the image semantic segmentation result (Ganin Y,2014). The classical VGG-16 (Wang B,2014) neural network has five pooling layers, and the final pooling image is 1/32 of the input image. The method of down-sampling can enlarge the receptive field to obtain larger scale features, but reduce the spatial information. In order to retain or restore spatial information, continuous convolution-deconvolution up-sampling operation is usually carried out after a series of down-sampling. Fully convolution networks (FCN) and U-Net are improved image semantic segmentation algorithms based on CNN (Szegedy C,2015). FCN replaces the full connection layer of CNN with the upper sampling layer to restore the reduced image to the same size as the input image, so as to achieve end-to-end structure and achieve pixel level semantic segmentation (Mendes CCT,2016). FCN algorithm has achieved good results in image segmentation, but it also has rough edge segmentation. K-means clustering algorithm belongs to unsupervised algorithm, which can find out the similarity between unannotated data set instances and divide it into K categories. Conditional random fields (CRF) (Lafferty JD,2001) is a classical discriminant probability undirected graph learning model, which is used to solve similar feature classification problems. In this paper, we combine the advantages of full convolution network and conditional random airport, and propose a lane segmentation method based on U-Net network with CRF processing at the back end (HE K M, et al,2016). Experimental results show that the proposed method can maintain the real-time performance and improve the segmentation accuracy.

2. Combined Model

The combined model is composed of U-Net network and CRF.The whole lane segmentation process is shown in figure 1. Firstly, the gray-scale RGB image is input into the u-net network for training, and then the output results and the original RGB image are input into the CRF of the back end for edge smooth segmentation. The gray image output by CRF is combined with the original RGB image to get the final result.

2.1 U - Net Network

The structure of the u-net network is shown in figure 2. It is improved from the classic VGG-16 network. We removed one layer of convolution-pooling structure from VGG-16 network, that is, only the sampling structure under 4 layers was used, and then added 4 layers of deconvolution structure. After four convolution and pooling operations, the input image was compressed to 1/16 of the original size. This down-sampling process not only gains more feature semantic information, but also loses a lot of target spatial information. In order to solve this problem (J. Hur,2013), the latter part carries out 4 upper sampling operations, then superimposes and convolves the feature graphs of the same size. This operation maximally extracts semantic feature information while preserving target spatial information.

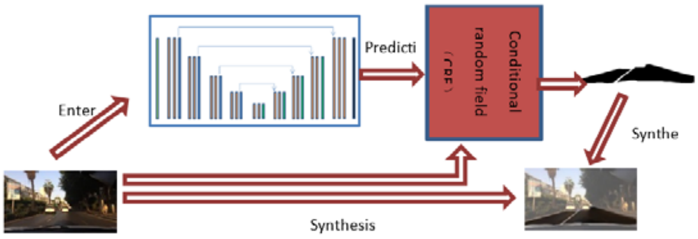

Figure 1: Lane segmentation process

From left to right, U-Net has 28 layers, including 19 convolution layers, 4 pooling layers, 4 deconvolution layers and 1 Softmax layer. The convolution kernel number of each set of convolution layers is the same, the convolution kernel size is 3x3, and the step length is 1. In order to simplify the feature image clipping process, the convolution process is filled with the “same”. The left encoding section is 64, 128, 256, 512, and 1024 in order from left to right. On the right side, the number of convolution kernels of each deconvolution layer is 512, 256, 128 and 64 respectively. Each convolution layer is followed by a batch of normalized layers and rectified linear units (LeakyReLU). The network structure adopts the maximum pooling layer pooling, with the size of 2x2 and the stride length of 2.

In order to complete the task of lane segmentation better, this paper improves the structure of the traditional U-Net network (V. Badrinarayanan,2017). We replaced the original relu activation function with the LeakyReLU function.

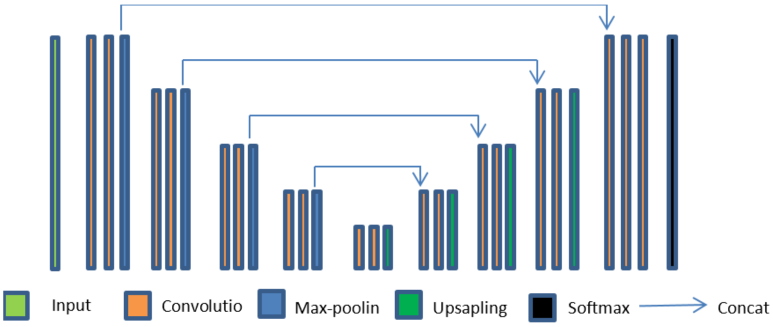

Figure 2: Structure of U-Net

Relu function not only has no gradient saturation problem, but also has a faster convergence rate than sigmoid function and tanh function. However, during the training of relu function (L.-C. Chen,2016), there exists the phenomenon of neuronal "necrosis", that is, when the neuronal input is negative, the gradient will be 0, so that the neuron cannot update. In order to solve the neuronal "necrosis" phenomenon of relu function, LeakyReLU activation function was adopted, and its expression was as follows:

$$f(x) = \alpha x, x < 0 \tag{1}$$

$$f(x) = x, x \geq 0 \tag{2}$$

Where is the offset, which is generally α small number obtained through training and learning. This allows the input to be negative without loss of information (SZEGEDY C, et al,2015). In the process of up-sampling, deconvolution operation is adopted, and the specific operation principle is shown in figure 3. As can be seen from figure 3, the deconvolution process retains the original spatial information of the image to a certain extent (SHI W Z, et al,2016).

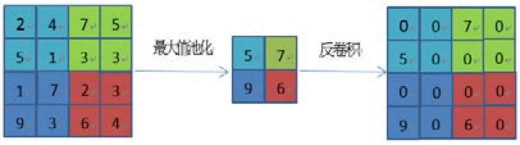

Figure3. Deconvolution process.

In the training, this paper adopts Adam algorithm. Adam is essentially RMSprop (root mean square prop) with momentum term, which adjusts the learning rate of each parameter through the first-order estimation matrix and the second-order estimation matrix of the gradient. After each correction, the iteration learning rate changes within a certain range, which makes the parameter change relatively stable. The main formula is as follows:

$$m_t = \mu * m_{t-1} + (1 - \mu) * g_t \tag{3}$$

$$n_t = \nu * n_t + (1 - \nu) * g_t^2 \tag{4}$$

$$\hat{m}_t = \frac{m_t}{1 - u^t} \tag{5}$$

$$\hat{n}_t = \frac{n_t}{1 - \nu^t} \tag{6}$$

$$\varDelta\theta_t = -\frac{\widehat{m}_t}{\sqrt{\widehat{n}}+\epsilon} * \eta \tag{7}$$

Where, equations (3) and (4) respectively represent the first-order moment estimation and second-order moment estimation, and equations (5) and (6) represent the correction of the first-order second-order moment estimation. Formula (7) is the rate at which the adaptive learning rate changes, keeping the learning rate within a certain range.

Compared with stochastic gradient descent methods (SGD) and RMSprop, Adam algorithm has obvious advantages:

1) less need for computer memory.

2) combining the advantages of Adagrad and RMSprop, it performs very well in dealing with sparse gradients and non-stationary targets.

3) can calculate different adaptive learning rates in the training process.

4) it also has good effects on big data sets, non-convex optimization and high-dimensional space.

2.2 CRF

Although the full convolutional network can complete the lane segmentation task by training a large enough data set, the study found that the performance of the pure neural network in the target edge segmentation was not good enough (Zheng S, et al,2015), and the introduction of CRF back-end processing was a good solution.

CRF is a discriminant undirected probability graph model that classifies each frame in a sequence. $P(Y|X)$ is a linear condition random field, set $X = \{x_1, x_2, \cdots x_n\}$ is a data set of random variables, $Y = \{y_1, y_2, \dots, y_n\}$ is the label of the data set, N is the number of pixels of the image, then the CRF model formula is as follows:

$$P(Y|X; \varTheta) = \frac{1}{Z(X)} exp\big(-E(Y, X; \varTheta)\big) \tag{8}$$

Among them: $E(Y, X; \varTheta)$ is the energy function, $Z(X)$ is a normalized factor, $Z(X) = \varSigma\, exp\big(-E(Y, X; \theta)\big)$. After minimizing the energy function, the optimal pixel classification results can be obtained. The energy function formula is as follows:

$$E(Y, X; \varTheta) = \sum\nolimits_{P \in N} \varPhi^{(1)}(y^p, x; \theta) + \sum \varPhi^{(2)}(y^p, y^q, x, \theta) \tag{9}$$

Type: $\varPhi^{(1)}$ and $\varPhi^{(2)}$ are a potential function and dual potential function, which is suitable for the distribution of x and y to θ, p and q are corresponding nodes. The unary potential function is the characteristic of a single pixel. The binary potential function is not only related to its own single pixel, but also includes 4 adjacent pixels and 4 diagonal pixels. In order to reduce the complexity of the model and computation (D. Eigen ,2015), we only use the adjacent 4 pixels. We use maximum likelihood estimates to assign the most likely label to the random variable X.Unary potential function and binary potential function are shown as follows:

$$\varPhi^{(1)}(y^p, x) = \lambda_m I(y, y^p) x(l) \tag{10}$$

$$\varPhi^{(2)}(y^p, y^q, x; w) = \mu_m I(y, y^p) I(y', y^q) \tag{11}$$

Where: I is the indicator function of 0 or 1, and λ_m、μ_m are the parameters of state function and transition function respectively. w is the distribution parameter that x and y obey. The binary potential function is the relation between pixels, and the relation between pixels is as follows:

$$\sum\nolimits_{(p_3 y) \in S} \varPhi^{(2)}(y^p, y^q, x, \theta) = \sum_{i=1}^{4} \sum_{(p,q) \in s_i} \Phi_i^{(2)}(y^p, y^q, x, \theta) \tag{12}$$

Where: s_1, s_2, s_3 and s_4 represent the adjacent pixels in the top, bottom, left and right directions respectively.

3. Analysis of Results

3.1 The Data Set

The data set used in this paper is the largest autonomous driving data set BDD100K launched by Berkeley university AI laboratory (BAIR) in 2018.The data set contains 100,000 images and tags with 720pixels×1280pixels.

Considering the training time cost of the model, 10,000 pictures and tags in BDD100K were compressed to 160pixels×320pixels for training, and 2700 pictures were randomly selected from the remaining 90,000 pictures as the verification set. In order to make the model can better adapt to all kinds of road environment and weather and light conditions, we selected 2000 photos of five weather conditions: sunny, cloudy, rainy, foggy and snowy. These images include six scenes of residential areas, highways, city streets, gas stations, tunnels, parking lots and three times of dawn/dusk, day and night. In order to better compare with other algorithms, we also made a training of k-means, FCN-16s, U-Net.

3.2 Performance Evaluation

The lane area segmentation problem can be regarded as a dichotomy problem. All the pixels in the picture are classified as the target or background. The predicted classification results were compared with the annotated images to evaluate the results for each pixel. The results can be divided into four cases: true positive (TP), false positive (FP), true negative (TN) and false negative (FN).We can calculate the true case rate (TPR) and false positive case rate (FPR) of each pixel:

$$TPR = \frac{TP}{TP+FN} \tag{13}$$

$$FPR = \frac{FP}{TN+FP} \tag{14}$$

The “ROC curve” can be obtained by taking FPR as the horizontal axis and TPR as the vertical axis. AUC (Area Under ROC Curve) can be obtained through calculation. We take AUC as the criterion to evaluate the quality of lane segmentation model. The greater the AUC, the better the segmentation model accuracy (Ronneberger.O,2015).

3.3 Model Training

The lane model in this paper is developed with python, with Windows 10 operating system, Keras development framework and cuDNNv7.6 experimental backend. The hardware device adopts Intel(R) Core (TM)i7-3770 CPU @3.40ghz, 10GB memory, and NVIDIA GeForce GTX TITAN X. The algorithm adopts Adam optimizer, and each training is conducted by randomly selecting 80% of the training set and 20% of the test set. The training weight attenuation was set as 1e-5, the batch size was 2, the loss function was “binary_crossentropy”, the Dropout layer was set as 0.5, and the activation function was LeakyRuLU. 40 times of training was conducted for each image, with a total of 400,000 iterations. The back-end CRF is developed in python, and the third-party database used by CRF is pydensecrf. Figure 4 shows the loss value and AUC changes of U-Net + CRF. After 400,000 iterations, both the loss curve and AUC tend to be stable. It can be seen that the loss value is stable about 0.038, and the AUC is about 98.86%.

3.4 Result Analysis

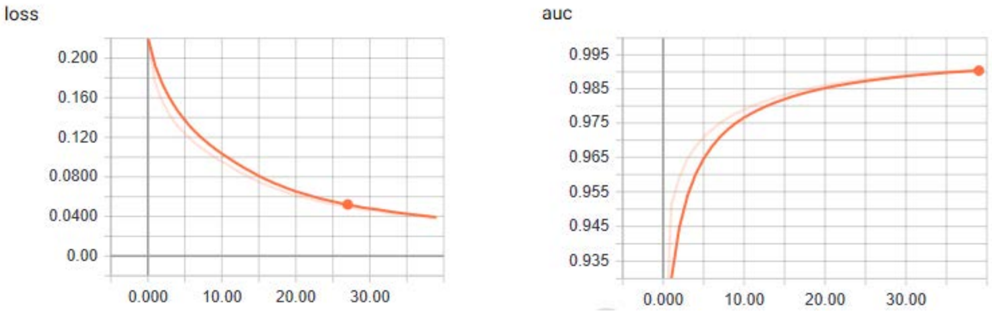

Figure 4: Loss function curve and AUC curve of U-net + CRF.

Figure 5 is the result of partial image segmentation. The first line is the annotated image, the second line is the image segmentation result of U-Net network, the third line is the image segmentation result of FCN-16s network, and the fourth line is the result obtained by the method in this paper. Figure 5 shows the straight road in the city on the left and the curved road in the country on the right. Compared with the result of the first line and the second line, it is obvious that the simple U-Net network can better identify the lane position (G. Ros,2016), but it is very rough in lane edge segmentation. The disadvantage of poor edge segmentation ability is more obvious in the curve lane segmentation. Comparing the images of the second, third and fourth lines, we can find that FCN-16s and U-Net + CRF are relatively better than U-Net in edge segmentation. Among them, U-Net + CRF performs better in detail segmentation than FCN-16s.The U-Net + CRF treated curves extend even further than the marker image. This shows that the method in this paper is really useful in improving lane edge segmentation.

Table 1 shows the results of the algorithm in this paper and other classical segmentation algorithms. Each algorithm is tested under the same data set we selected. As can be seen from the table, U-Net + CRF improved by about 5% compared with k-means, and by about 3% compared with pure U-Net. U-net + CRF took only 0.01 seconds longer to process a test image than U-Net.

4. Conclusion

This paper proposes a new lane segmentation method based on full convolution neural network and probability graph. We regard lane segmentation as an object-background dichotomy problem, and use U-Net to extract image features to initially segment lane and background, and then use CRF to conduct smooth segmentation on the edges of segmentation results. The method proposed in this paper combines the autonomous learning and excellent feature extraction ability of full convolutional neural network, and reasoning ability of conditional random field to further improve lane detection results in detail segmentation. In order to make the model more robust, we give up the practice of migration learning and obtained our own model under the training of the latest data set BDD100K.The method has been trained with a lot of data and has good results in various complex lane environments. This method has achieved 98.86 AUC and 0.09s processing speed, and achieved expected detection results, which is a good segmentation method.

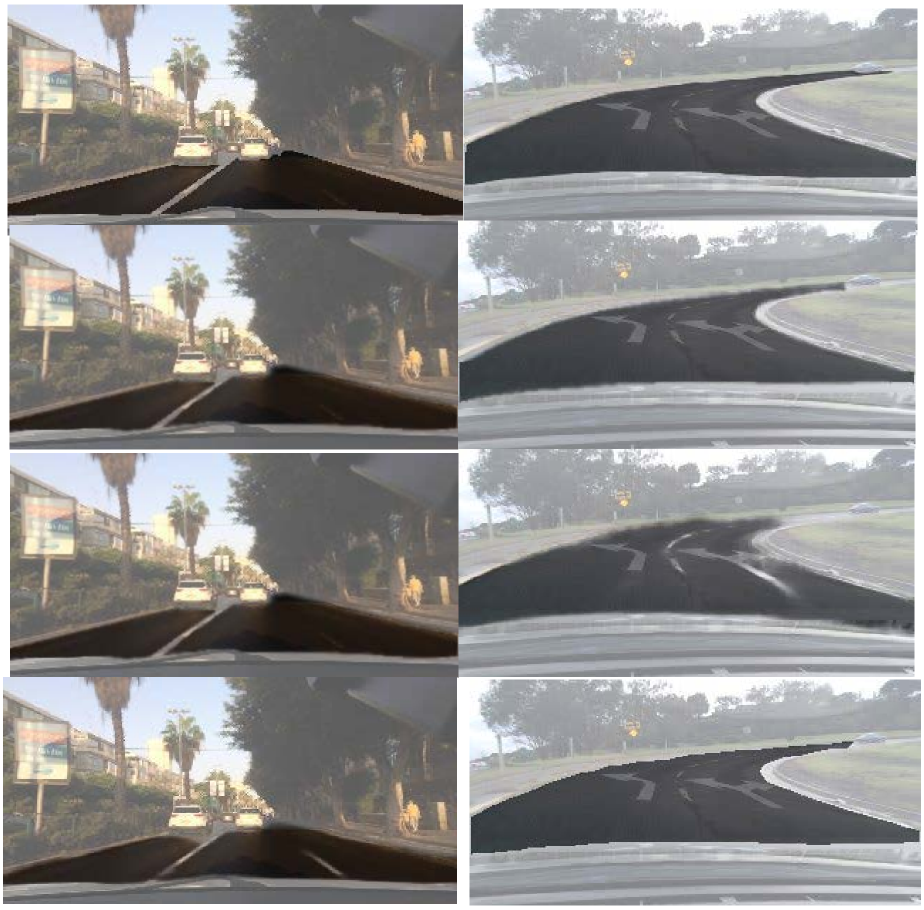

Figure 5: Part of the road segmentation results.

Table 1: The test results of each algorithm

methods	AUC/%	Runtime/s
k-means	93.00	0.16
FCN-16s	96.85	0.04
U-Net	95.56	0.08
U-Net+CRF	98.86	0.09

References

[1] B. Wang, Frémont V, Rodríguez S A. (2014) ‘Color-based road detection and its evaluation on the KITTI road benchmark’[C]// IEEE Intelligent Vehicles Symposium Proceedings, Dearborn, MI, USA, 31–36.

[2] HE K M, ZHANG X Y, REN S Q, et al. (2016) ‘Deep residual learning for image recognition’, IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 770-778.

[3] D. Eigen and R. Fergus, (2015) ‘Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture’, in ICCV, pp. 2650–2658.

[4] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, (2016) ‘The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes’, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[5] J. Hur, S. N. Kang, and S. W. Seo, (2013) ‘Multi-lane detection in urban driving environments using conditional random fields’, in Intelligent Vehicles Symposium (IV).

[6] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. (2001) ‘Conditional random fields: Probabilistic models for segmenting and labeling sequence data’[C]//Proceedings of the 18th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann Publishers Inc., 282-289.

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, (2016) ‘Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs’, arXiv preprint arXiv:1606.00915.

[8] LECUN Y, BOTTOU L, BENGIO Y, et al. (1998) ‘Gradient-based learning applied to document recognition[J]’, Proceedings of the IEEE, 86(11): 2278-2324.

[9] Mendes C C T, Frémont V, Wolf D F. (2016) ‘Exploiting fully convolutional neural networks for fast road detection’[C]//2016 IEEE International Conference on Robotics and Automation, Stockholm, Sweden.

[10] O. Ronneberger, P. Fischer, and T. Brox, (2015) ‘U-net: Convolutional networks for biomedical image segmentation’, in MICCAI.

[11] S. Graovac, and Goma, A. (2012) ‘Detection of Road Image Borders Based on Texture Classification’, International Journal of Advanced Robotic Systems. doi: 10.5772/54359.

[12] SHI W Z, CABALLERO J, HUSZÁR F, et al. (2016) ‘Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network’, IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Nevada, USA: IEEE, 2016: 1874-1883.

[13] SZEGEDY C, LIU W, JIA Y Q, et al. (2015) ‘Going deeper with convolutions’, IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015: 1-9.

[14] SZEGEDY C, LIU W, JIA Y Q, et al. (2015) ‘Going deeper with convolutions’, IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015: 1-9.

[15] V. Badrinarayanan, Kendall A and Cipolla, R. (2017) ‘SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation’, IEEE Transactions on Pattern Analysis and Machine Intelligence. 39 (12) p.e 95.doi:10.2481.

[16] Y. GANIN, LEMPITSKY V. (2014) N4-fields: ‘Neural network nearest neighbor fields for image transforms’ [C]//Proceedings of the 12th Asian Conference on Computer Vision. Singapore: Springer, 536-551.

[17] ZHENG S, JAYASUMANA S, ROMERA-PAREDES B, et al. (2015) ‘Conditional random fields as recurrent neural networks’, IEEE International Conference on Computer Vision, Santiago, Chile: IEEE, 2015: 1529-1537.