Combining Convolutional-kernel-based Conditional Random Field with Deep Neural Network For Road Detection

Xiongxuan Huang[1,a] and Xueyun Chen[1,b]

College of electrical engineering, Guangxi University, Nanning, Guangxi Province, China
a.549317863@qq.com, b. cxy177@163.com

Keywords: Automatic pilot, deep learning, conditional random field, convolutional neural network.

Abstract: Deep learning methods have achieved excellent detection results in various road databases. However, due to some harsh road conditions, these network models still suffer from some abnormal phenomena such as hollowness, unsmooth edges and so on. To address this problem, we improve the road detection network via a novel convolutional-kernel-based conditional random field (CK-CRF), which designs some special convolutional kernels for hollowness scanning. Experiments on two open road datasets show that the proposed method outperforms the-state-of-the-art models by an obvious margin, including the famous deeplab network with a traditional conditional random fields based on fully-connected layers.

1. Introduction

Road detection has a significant meaning to auto-driving technology. Due to the influences from various illumination, shadows and waterlogging, it remains far away from being solved.

Current detection methods can be divided into three kinds: model-based, features-based and learning-based. In the first-kind methods: Wang et al. [1] utilized B-spline curve to build the road model, which can quickly distinguish different road shapes and build matching models. Alvarez et al. [2] proposed an adaptive road geometry model, which can adjust the model by the analysis on the current and previously-collected road scenes. To alleviate the influence of illumination, various features were proposed by the second-kind methods: Alvarez et al. [3] used color features, extracted light-source-invariant features function to classify road pixels. Tsai et al. [4] applied three road clue types (road smoothness, color and lane line), and conditional random field (CRF) to integrate all features for the urban road detection. In the third-kind frameworks, Lee et al. [5] proposed a multi-task network for road detection, which can detect and identify road and traffic signs simultaneously in extreme weathers by using the information of vanishing points. Inspired by semantic segmentation and instance segmentation, Neven et al. [6] proposed a multi-task network model with branch structure, which converted the lane detection problem into an instance segmentation problem, and treated each lane as a separate instance.

The fully-connected conditional random field (FC-CRF) is the most popular technique to optimize the detection results, which uses a fully convolution neural network to minimize the operation of conditional random field in the neural network model, such as deeplabv1 [7]. However, it has three defects: 1. FC-CRF cannot give an end-to-end result; 2.The computing of FC-CRF must be iterated many times. 3. FC-CRF cannot be used to improve the training of the back-bone network.

To address these problems, we propose a convolutional-kernel-based conditional random field (CK-CRF), which realizes the operation of conditional random field based by an extra convolutional neural layer connected to the back-bone network. The main contributes are listed as:

1) The energy function of the CK-CRF is combined into the loss of the back-bone network, making their training be a united one.

2) The special design of the fixed convolutional-kernels make it suitable for the reducing of a large variety of road defects.

3) The CK-CRF is only used and trained in the training stage, and will not affect the testing speed of any practical applications.

2. Related Work

2.1.FCRN

The fully convolution regression neural network (FCRN) is proposed to improve the fully convolution neural network (FCN). The output of FCRN is the prediction of the target spacial density distribution on the image. The density distribution is usually the same as the shape of the target. The value of each pixel is between 0 and1. A value close to 1 means that the pixel belongs most likely to the target region, otherwise, it is outside of the target.

In this paper, we use the U-net network as the back-bone of FCRN. The U-net model adopts the traditional encoder decoder structure, which is characterized by some skip-connection structures. The loss function uses the mean square error as equation (1):

$$Loss_{pix}(x_{label}, x_{out}) = \|x_{label} - x_{out}\|^2 \tag{1}$$

Where x_{label} is the sample in the calibration dataset S_l, and x_{out} is the output of FCRN.

2.2.Conditional Random Field (CRF)

FCNs have excellent classification and precise positioning capabilities, but they cannot have both. For example, the deep neural networks with the max-pooling layers have achieved excellent classification accuracy in the classification tasks, but when faced with the problems of pixel-level classification, such as semantic segmentation, the boundaries between different classification areas will become confused and unable to accurately locate. To tackle this problem, deeplabv1 combines the FCN and probability graph model to develop the FC-CRF, which can enhance the boundary location effect of the recognition area.

In deeplabv1, the energy function of the CRF used can be expressed as:

$$E(y_j, y_i, x) = \sum_{i,l} \theta_i(y_i, x)\,) + \sum_{i,k} \theta_{ij}(y_i, y_j, x) \tag{2}$$

$$\theta_i(y_i, x) = \|y_i - z_i\|^2 \tag{3}$$

$$\Theta_{ij}(y_i, y_j, x) = \mu(y_i, y_j) \sum_{m=1}^{k} \omega_m\, k^m(f_i, f_j)\text{，}\ \mu(y_i, y_j) = \begin{cases} 1\text{，}\ y_i \neq y_j \\ 0, y_i = y_j \end{cases} \tag{4}$$

Where $z_i \in Z$， $Z = (z_1,z_2,...,z_n)$, Z represents the calibration picture, ω_m is the weight of each pixel, $k^m(f_i,f_j)$ uses Gaussian kernel function to describe the relationship between pixel i and its surrounding pixel j, and f_i refers to the characteristic function of pixel i. In theory, the binary potential Θ_{ij} exists in any two pixels of Y. It means that the CRF is fully connected in Y, so it's named fully-connected conditional random field (FC-CRF).

The unitary potential in FC-CRF is used for training the fully convolution neural network. Binary potential energy judges the prediction probability or prediction category of pixel i by the prediction probability or prediction category of pixel j around pixel i. The binary potential energy optimizes the prediction results to the certain extent, but the use of fully-connected conditional random field requires two stages of operation. First, we need to complete the training of the neural network part, and then debug the fully connected condition random field.

Therefore, FC-CRF breaks the advantages of the end-to-end use of fully convolution neural network. Not only that, FC-CRF optimizes based on the output of the network, which is limited and cannot fundamentally solve the problems of hollowness and edge anomalies.

3. Proposed Method

We believe that when the target spacial density distribution of adjacent pixels changes greatly, there will be holes in the middle of the road recognition area, spots in the no Road area and strange edge shape, that is shown in output without CK-CRF of Figure 1. In order to address these problems, we design a convolutional-kernel-based conditional random field to smooth the probability distribution of FCRN output.

The energy function of CK-CRF is as equation (5):

$$P_e(x_{out}) = Relu\left(Conv\left(x_{out},k\left(x_i,x_j\right)\right)\right) \qquad (5)$$

Where $Conv()$ means to use convolution kernel $k(x_i,x_j)$ to convolute the output x_{out} of the FCRN, and $Relu()$ means to perform relu operation.

$$Loss_{E-CRF}(x_{out}) = \frac{1}{n}Sum(P_e(x_{out})) \qquad (6)$$

Where n is the number of positive pixels in P_e, and $Sum()$ is the sum function.

$$k(x_i,x_j) = \begin{cases} exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right),i \neq j \\ exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right) - 1,i = j \end{cases} \qquad (7)$$

Where $x_i \in Y$, $x_j \in \partial$, ∂ is a neighborhood of x_i in the open interval $(i-r,i+r)$. The convolution kernel can be instantiated in scale 3*3 as follows:

$$\begin{bmatrix} 0.0625 & 0.125 & 0.0625 \\ 0.125 & -0.75 & 0.125 \\ 0.0625 & 0.125 & 0.0625 \end{bmatrix}$$

The energy function is designed to smooth the target spacial density distribution. It can be seen that there are quite different between output with CK-CRF and output without CK-CRF in Figure 1.

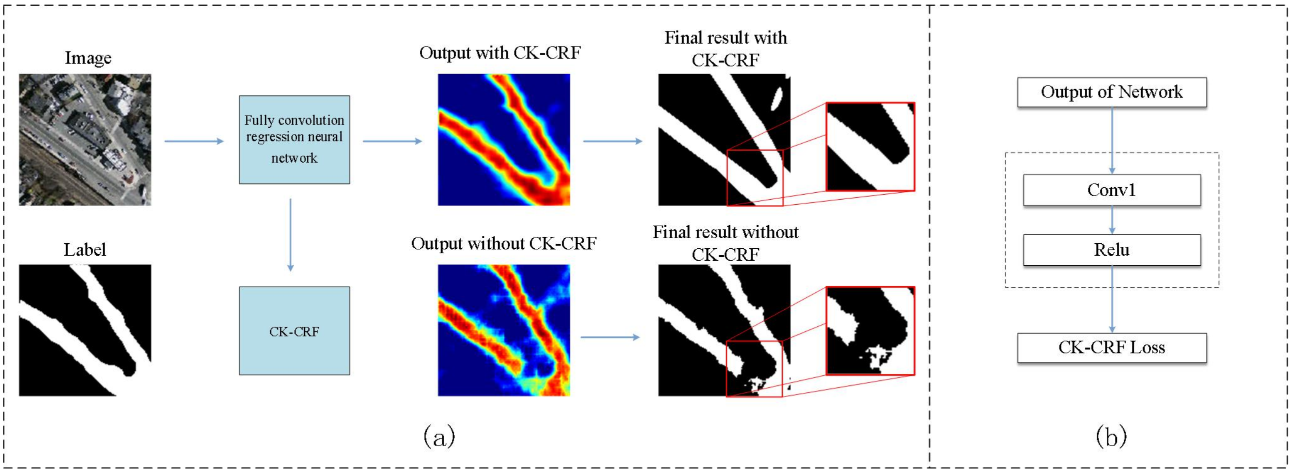

Figure 1: Model illustration. (a) System structure; (b) Structure of CK-CRF.

Then the total loss of the network is:

$$Loss_{all}(x_{label}, x_{out}) = \sum_{x_{label} \in S_l} \begin{pmatrix} \omega_1 Loss_{E-CRF}(x_{out}) + \\ \omega_2 Loss_{pix}(x_{label}, x_{out}) \end{pmatrix} \tag{8}$$

Where x_{label} is the label in the calibration dataset S_l, and x_{out} is the output of network with input x_{in}.

At this point, we add the optimization function of conditional random field to the network, so that the probability distribution of the network output is smooth and not discrete. In Figure 1, we can see that CK-CRF is composed of a convolution layer and a Relu layer, which is relatively independent of the semantic segmentation recognition network and does not increase the scale of network parameters. This feature enables CK-CRF could be placed at the output of various semantic segmentation networks. CK-CRF is only used in the training phase, that makes it will not affect the running speed of the network in the test stage or the practical stage.

Algorithm 1:

```
Input: Real image dataset $\boldsymbol{S_R}$, label image dataset $\boldsymbol{S_L}$; Parameters to be optimized ω; Loss functions used include $\boldsymbol{Loss_{E-CRF}}$, $\boldsymbol{Loss_{pix}}$; Maximum times $\boldsymbol{t_1}$in first stage; Maximum times $\boldsymbol{t_{max}}$ in total; Input images $\boldsymbol{x_R}$ obtained by random transformation in $\boldsymbol{S_R}$; Label images $\boldsymbol{x_L}$ obtained by the same random in $\boldsymbol{S_L}$.
1. For $t < \mathbf{t_{max}}$ do
2. From the real picture dataset $\mathbf{S_R}$, the input image $\mathbf{x_R}$ is obtained
   by randomly selecting iamges randomly, randomly cutting,
   randomly rotating and randomly translating
3. Take $\mathbf{x_R}$in as input and get output $\mathbf{x_{out}}$ through network
4. If $t < \mathbf{t_1}$ do
   Calculate $Loss_{pix}(x_{label}, x_{out})$
   Else do
   Calculate $Loss_{all}(x_{label}, x_{out})$
5. Update ω, return to step 1
6. End for
```

4. Experiment

In order to prove the performance of the proposed method, we did the experiment in two dataset: driving recorder dataset (DRD) and remote sensing road dataset (RSRD).

In our experiment, we use mini-batch SGD and apply the Adam solver, with a learning rate of 0.001 for RSRD and 0.0001 for DRD.

4.1.Dataset

4.1.1. RSRD

The RSRD showed in Figure 2 contains 26 remote sensing images of urban roads with the size of 1500*1500. The maximum width of the road in the image is only 30 pixels, and there is a lot of occlusion on the road area, such as trees, shadows. Therefore, the characteristics of RSRD is small target and low resolution. We take a 128*128 image from the random position of the image as the input of the network.

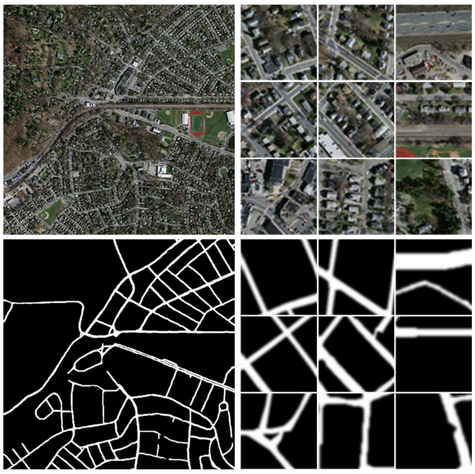

Figure 2: Remote sensing road dataset. This dataset is characterized by low resolution of small targets. In the image, the maximum road width is 30 pixels, and the minimum is less than 5 pixels. The difficulty of this data set lies in the limited information provided by the image and the occlusion of objects.

4.1.2. DRD

This dataset showed in Figure 3 contains 2000 urban roads images with the size of 720p or 1080p from driving recorder. The road scenes of image in various lighting conditions include various road locations, such as crossroads in the city, T-junctions in front of railway stations. DRD is a large target and high resolution dataset. We adjust the size of the picture to 1024 *512 as the input of the network.

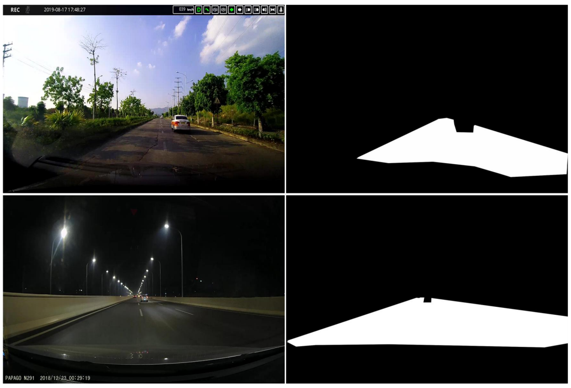

Figure 3: Driving recorder dataset. The characteristic of database is large target with high resolution, that is to say, the image is very clear, but the difficulty lies in the interference of different intensity light.

4.2.Visualization Comparison

The test results on RSRD are shown in Figure 4. There are some empties and unsmooth edges on the outputs from FCN and U-net, which is due to the shelter of vehicles and the shadow of buildings on the road. For example, in the first line, the reason for the disconnection of the road identification area on the U-net prediction result is that there is a shelter of buildings on the road. It can be seen that, even after the optimization with FC-CRF of deeplab, the result is still not ideal which shown in the fifth column. After using the CK-CRF, the output of the network is smooth, and the edge of final threshold segmentation result is smooth and the empties are completely eliminated.

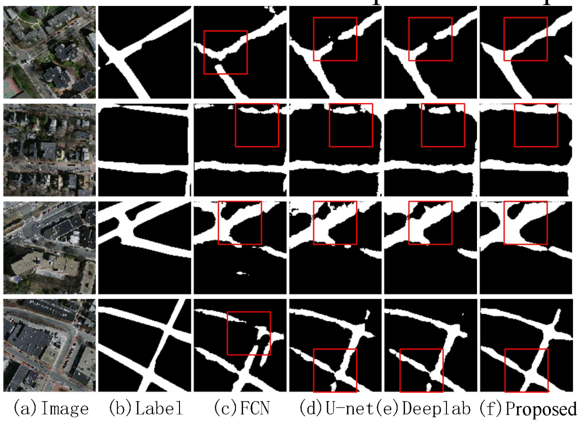

Figure 4: Visualization comparison in RSRD. There are holes, spots and unsmooth edges in the final result of FCN and U-net. The FC-CRF of deeplab can slightly improve the prediction results. CK-CRF provides much better prediction results.

The test results on the DRD are shown in Figure 5. The outputs of FCN and U-net are not smooth and discrete, that there are some empties and unsmooth edges in the prediction results, especially in the case of strong light interference. For example, in the second line, due to the lack of light, there are spots and uneven edges on the right of the recognition area. And in the case of strong backlight in the third line of Figure 5, empties appear in the middle of the recognition area. The optimization result of deeplab is not good enough that there is progress in small details, but the overall accuracy has not been greatly improved. Our proposed method can fundamentally optimize the training effect. Then, we can see that the CK-CRF with FCRN have filled empties and smoothing edges, and greatly improves the prediction accuracy of road identification area.

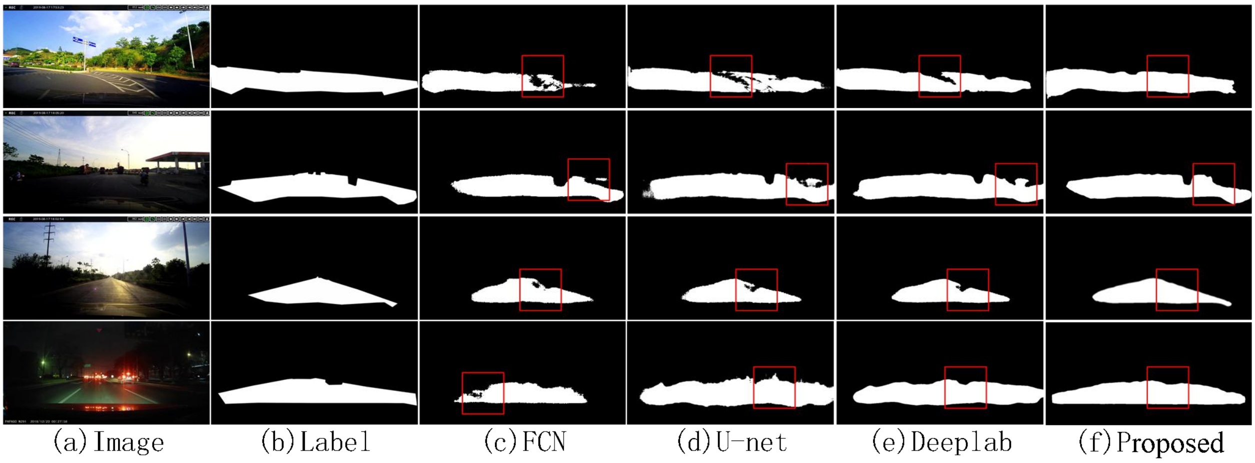

Figure 5: Visualization comparison in DRD. In the environment of strong light, backlight or dark road, there are some hollowness and unsmooth edges in the red box of FCN and U-net's prediction results. It has limited improvement for prediction results of deeplab. In contrast, the prediction results of CK-CRF are better.

4.3.Quantitative Comparison

We measure this methods from two objective evaluation indicators: Intersection over Union (IoU) and pixel accuracy (PA).

Table 1: Test results in the RSRD.

	IoU/%	**PA/%**
FCN	58.04%	91.31%
U-net	65.56%	93.14%
deeplab	63.37%	92.83%
Proposed	67.91%	93.47%

From Table 1, we can see that the algorithm we proposed has the highest IoU value and the highest PA value. On the original prediction accuracy (58.04%) of the FCN, CK-CRF increases it by 9%, which is 4% higher than the FC-CRF. This shows that FCRN with CK-CRF has better performance than deeplab with FC-CRF in the RSRD.

In the experimental results in Table 2, it can be seen that the algorithm we proposed is the best in the DRD. On the original prediction accuracy of the FCN (69.93%), FCRN with CK-CRF increases it by 15%, which is 7% higher than that of the deeplab. This shows that FCRN with CK-CRF has better performance than deeplab with FC-CRF in the DRD.

Table 2: Test results in the RSRD.

	IoU/%	**PA/%**
FCN	69.93%	94.87%
U-net	72.34%	95.82%
deeplab	78.43%	96.53%
Proposed	84.27%	97.66%

5. Conclusions

In this paper, a convolutional-kernel-based conditional random field (CK-CRF) is proposed. Experiments show that the CK-CRF with FCRN not only improves the prediction results of the network, but also provides the optimization performance beyond the deeplab with FC-CRF. It does not increase the scale of the network model and affect the running speed of the network model. In the future work, the loss function of the CK-CRF should be further improved to obtain better performance.

Our models and code are publicly available at https://github.com/fd851583/CK-CRF.

References

[1] Wang Y, Teoh E K, Shen D. Lane detection and tracking using B-Snake[J]. Image & Vision Computing, 2004, 22(4): 269–280.

[2] Alvarez J M, Gevers T, Diego F, Lopez A M. Road Geometry Classification by Adaptive Shape Models[J]. Intelligent Transportation Systems IEEE Transactions on, 2013, 14(1): 459-468.

[3] Alvarez J M A, x, opez A M. Road Detection Based on Illuminant Invariance[J]. Intelligent Transportation Systems, IEEE Transactions on, 2011, 12(1): 184-193.

[4] Tsai J-F, Huang S-S, Chan Y-M, Huang C-Y, Fu L-C, Hsiao P-Y. Road detection and classification in urban environments using conditional random field models[C]. in Intelligent Transportation Systems Conference, 2006, Toronto, Canada, 2006: 963-967.

[5] Seokju Lee, Junsik Kim, Jae Shin Yoon, Seunghak Shin, Oleksandr Bailo, Namil Kim, Tae-Hee Lee, Hyun Seok Hong, Seung-Hoon Han, VPGNet: Vanishing Point Guided Network for Lane and Road Marking Detection and Recognition International Conference on Computer Vision (ICCV 2017).

[6] Davy Neven, Bert De Brabandere, Stamatios Georgoulis, Marc Proesmans, Luc Van Gool, Towards End-to-End Lane Detection: an Instance Segmentation Approach. arXiv:1802.05591 [cs.CV].

[7] Liang-Chieh Chen Univ. of California, Los Angeles, DeepLabv1: Semantic image segmentation with deep convolutional nets and fully connected CRFs ICLR 2015 (International Conference on Learning Representations).

[8] Wang B, Fremont V. Fast road detection from color images[C]. in Intelligent Vehicles Symposium (IV), 2013 IEEE, 2013: 1209-1214.

[9] Rasmussen C. Grouping dominant orientations for ill-structured road following[C]. in CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, 2004, 1: 470-477.

[10] Kuhnl T, Kummert F, Fritsch J. Monocular road segmentation using slow feature analysis[C]. in Intelligent Vehicles Symposium (IV), 2011 IEEE, 2011: 800-806.

[11] Kong H, Audibert J-Y, Ponce J. Vanishing point detection for road detection[J]. Computer Vision and Pattern Recognition, 2009: 96-103.

[12] Kong H, Audibert J-Y, Ponce J. General Road Detection From a Single Image[J]. Image Processing, IEEE Transactions on, 2010, 19(8): 2211-2220.

[13] Moghadam P, Starzyk J A, Wijesoma W S. Fast Vanishing-Point Detection in Unstructured Environments[J]. Image Processing, IEEE Transactions on, 2012, 21(1): 425-430.

[14] Chang C-K, Siagian C, Itti L. Mobile robot monocular vision navigation based on road region and boundary estimation, 10.1109/IROS.2012.6385703.

[15] boundary estimation[J]. IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012, 7198(6): 1043-1050.

[16] Gao Q, Luo Q, Sun M. Rough Set based Unstructured Road Detection through Feature Learning[C]. in Automation and Logistics, 2007 IEEE International Conference on, 2007: 101-106.

[17] Foedisch M. Adaptive real-time road detection using neural networks[J]. Proc.int.conf.on Intelligent Transportation Systems Washington D.c, 2004: 167-172.

[18] Foedisch M. Adaptive real-time road detection using neural networks[J]. Proc.int.conf.on Intelligent Transportation Systems Washington D.c, 2004: 167-172.

[19] Passani M, Yebes J J, Bergasa L M. CRF-based semantic labeling in miniaturized road scenes[C]. in Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on, 2014: 1902-1903.

[20] Everingham M, Eslami S M A, Gool L V, Williams C K I, Winn J, Zisserman A. The Pascal Visual Object Classes Challenge: A Retrospective[J]. International Journal of Computer Vision, 2015, 111(1): 98-136.

[21] Scharwachter T, Franke U. Low-level fusion of color, texture and depth for robust road scene understanding[C]. in Intelligent Vehicles pixel accuracy Symposium (IV), 2015 IEEE, 2015: 599-604.

[22] *Gupta A, Efros A A, Hebert M. Blocks World Revisited: Image Understanding Using Qualitative Geometry and Mechanics[J]. Lecture Notes in Computer Science, 2010: 482-496.*

[23] *Floros G, Rematas K, Leibe B. Multi-Class Image Labeling with Top-Down Segmentation and Generalized Robust P^N Potentials[J]. In: BMVC. (2011), 2011: 1-8.*