A Fusion Model for Road Detection based on Deep Learning and Fully Connected CRF

Fei Yang[a, b], Zhong Jin[a, b*], Jian Yang[a, b], Huan Wang[b]
[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[b] Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science and Technology, Nanjing 210094, China
{yangfei92516}@163.com, {zhongjin, csjyang, wanghuanphd}@njust.edu.cn

***Abstract*—This paper presents a road detection model based on deep learning and fully connected CRF (conditional random field) to fuse image and point cloud data. Firstly, a convolutional neural network is trained to extract multi-scale features of the image. And a point-based deep neural network is trained to extract the multi-scale features of the point cloud. Secondly, the point cloud data is projected to the image plane. The probability maps of image and point cloud in the image plane are obtained by their corresponding multi-scale features, respectively. Thirdly, a Markov-based up-sampling method is used to get a dense height image from a sparse one which is from the point cloud data. A fully connected CRF model based on the outputs of the two networks and the height image is constructed on the image plane. Finally, the fusion model is effectively solved by the mean-field approximate algorithm. Experiments on KITTI Road dataset show that the proposed model can effectively fuse the image and the point cloud data. Furthermore, the fusion model can also exclude the shadows, road curbs and other interferences in complex scenes. Code is available at https://github.com/yangfei1223/FusionCRF.**

***Index Terms*—road detection, fusion, convolutional neural network, fully connected CRF, Markov-based up-sampling**

I. Introduction

Road detection has attracted munch attention in computer vision. Since the UGV (Unmanned Ground Vehicle) / UAV (Unmanned Aerial Vehicle) has developed rapidly in recent years, many different road detection systems have emerged. These systems use a variety of sensors to obtain environmental information, and then use this information to model for the purposes of road detection. Most of sensor systems for road detection are monocular vision systems, stereo vision systems, lidar systems, IMU (Inertial Measurement Unit), or the combination of these systems. Monocular vision based detection algorithm can not capture the spatial structure information of scenes. Although stereo vision can make up for the deficiencies of monocular vision, it has the disadvantages of high computational cost and serious noise. In recent years, due to the popularization of range sensor such as lidar, the range sensors based detection methods have been widely used. Point cloud data provided by lidar can effectively exclude the light interference, but it is not sensitive to the area where the height characteristic is not obvious. And because of the sparseness of point cloud data, it is hardly to get a precise result. Considering the deficiencies of single sensor, multi-sensor can effectively make up for the lack of single sensor. So multi-sensor fusion based road detection methods is a research trend. Most previous fusion models are based on the TextonBoost features [1] and the local pairwise CRF (conditional Random Field). Only the local features can be extracted during the feature extraction stage, and only the local neighborhood information is considered during the fusion stage in this way. Because on the one hand, the receptive filed of features is limited, which leads to the lack of discrimination of the features. On the other hand, the error correction capability of the CRF model has also been limited. It results in that the final label is mostly determined by the pixel classification, and that the label smoothing always happens in local area.

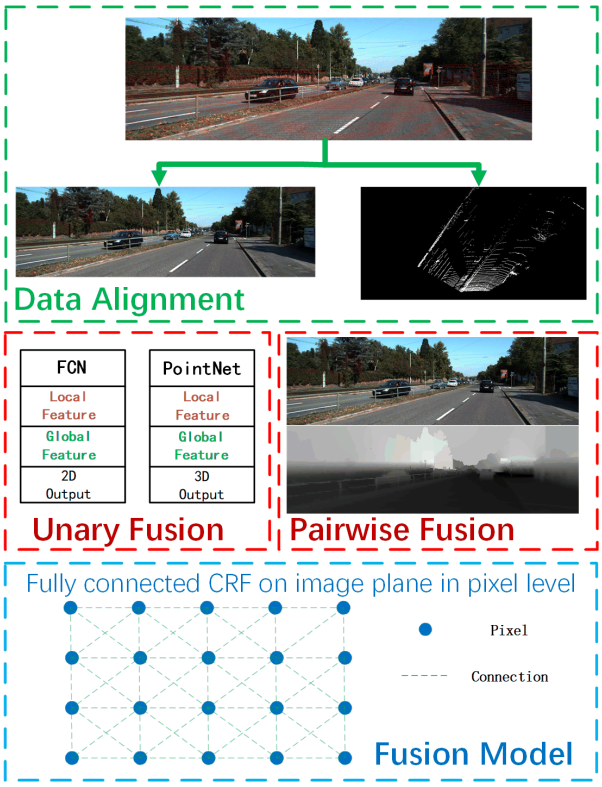


Fig. 1. Diagrammatic sketch of the proposed method.

In order to solve these problems, we present a fusion model based on deep neural network and fully connected random field. The model architecture is shown in Fig. 1. Firstly, the

image data and the point cloud data are aligned with cross calibration data by projecting the point cloud data to the image plane. Secondly, the multi-scale semantic features of images and point cloud are extracted by the deep learning architecture, which increases the discrimination of features of pixels and points. Finally, the pixel-level fully connected CRF which can extract correlations between the image data and the point cloud data is established on the image plane to fuse the outputs of the networks. So the proposed model possesses a bigger error correction capability during fusion stage. Experiments on KITTI show that the proposed model performs well in normal scenes. Moreover, it can distinguish complex regions which can not be dealt by traditional methods such as strong shadows, strong lights and road curbs efficiently and correctly.

This paper is organized as follows: Section 2 reviews the related work. Section 3 describes the deep neural network architectures for image and point cloud pre-segmentation in detail. In section 4, the fusion model based on fully connected CRF is introduced. Experiments on KITTI Road dataset and the comparison of results with some current popular road detection methods are analyzed in section 5. Finally, section 6 summarizes the work of this paper.

II. Related Work

At present, most of image-based road detection methods regard the road detection problem as an image semantic segmentation problem. And They adopt pixel or superpixel level classifiers for image segmentation. There are several shortcomings in these method. Firstly, the features for pattern classification are often artificially designed. These features belong to local features that contain only low level semantic information. Secondly, the classifier only considers the information of current pixel. In other words, the pixel are assumed to be independent with each other. As a result, the context information can not be utilized for segmentation. Thirdly, the superpixel-based methods rely heavily on the quality of unsupervised segmentation of the image. In view of these drawbacks, more advanced features are applied for road detection. For example, the TextonBoost features proposed by J. Shotton et al [1] are more capable of representation. The Spatial Ray feature proposed by J Fritsch et al [2] uses the spatial structure information which is suitable for road scenes. Since the classification model based on pixel or superpixel can not extract the context information, it is a popular trend that the CRF model is used for post-processing. Especially in recent years, CRF has a wide range of applications in semantic segmentation, scene reconstruction and object detection [1], [3].

Ladicky et al. [4] apply the features of [1] into the hierarchical CRF model. The features of different scales are extracted, then the Gradient Boosting based classifiers are trained to classify pixels and superpixels. Finally, a hierarchical CRF model is established to fuse the multi-scale information. This model can be optimized efficiently by the graph-cut based algorithm. In addition, with the development of CNN (convolutional neural network), some semantic networks such as FCN [5], SegNet [6] and DeepLab [7] achieve end-to-end semantic segmentation. These semantic networks use CNN to extract multi-scale complex image features and obtain better results. Although these methods take into account the context information and improve the receptive fields, they can not solve the confused areas such as complex shadows and road curbs effectively because of the lack of spatial structure information of the scenes.

Stereo system has the advantages of low price and convenient system setup. In order to utilize structure information, Fernandez et al. [8] use stereo camera to reconstruct the 3D scene, and then extracts the structure features such as the curvature and the normal vector from 3D points. However, the stereo system also has obvious shortcomings. On the one hand, the stereo system has a larger noise in the distance due to the perspective projection principle. So the reconstructed 3D model is often not accurate enough. On the other hand, the dense scene reconstruction has a relatively huge computation cost.

Except for traditional passive sensor (such as camera), active sensor (such as lidar) has also made great strides in road detection in recent years. Lidar-based methods extract the spatial structure information by analyzing the point cloud data of the scene , and then regard the plane area as road area. Moosmann et al. [9] propose a method to segment point cloud data by using local concavity and convexity. Chen et al. [10] use Gaussian process regression on grid maps in polar coordinates to segment the road area. Chen et al. [11] use a Lidar-histogram algorithm for road detection by analyzing the v-disparity map of the lidar-imagery they proposed. Compared with stereo vision system, lidar has the advantages of precise location and rapid computation. However, the density of point cloud data is usually low, even the density of 64-line lidar is still not comparable with image. Because of the lack of color and texture information, the distinction of road and non-road area in point cloud is mostly determined by the height information of the area.

Since single sensor is difficult to meet all needs, the fusion methods are a future development trend in environmental perception especially in road detection. Fusion algorithms can take the advantages of a variety of sensors to make up for the deficiencies of single sensor. Vitor et al. [12] adopt a method of combining 2D and 3D information for road detection. Shinzato et al. [13] map the lidar point cloud data to the image plane, then triangulate the lidar points to obtain the spatial relationship between points. Finally, they obtain the assessment of the road and the non-road by multiple free space detection. Hu et al. [14] use the lidar point cloud data to fit a plane. Then the points are projected to the image plane as the seed points of the Gaussian mixture model, and classified by using the Gaussian mixture model. Xiao et al. [15] map the point cloud to the image plane, then classify the point cloud data and the image data respectively. Finally they obtain the smooth road area by using a CRF for post-processing. However, all of these fusion algorithms only solve the problem of data information complementary. The problem

about how to solve the receptive field is not involved. Xiao et al. [16] propose a hybrid model named HybirdCRF, which is an improved fusion CRF model. However, this model merely encourages pixels and point cloud points to take the same label. It still can not solve the problem of image receptive field. Deep CNNs such as FCN, SegNet, and DeepLab use deep networks for feature extraction. These features contain multi-scale and contextual information. So the features extracted by deep networks have stronger representation ability than the artificially designed features. This paper combines the deep learning method with the fully connected CRF model for road detection. The proposed model benefits from the big receptive field of the deep networks and the big scope of the fully connected CRF. Compared with the locally connected CRF (4/8 neighborhoods) based fusion model, the proposed model has a stronger representation ability both in the feature extraction stage and the fusion stage.

III. Network Architecture

With the development of deep learning technology, it has made revolutionary achievements in computer vision, especially in image processing. Due to the powerful ability of representation learning, which can automatically extract the corresponding features by using convolution in the specific tasks. Compared with the artificially designed features. These features have stronger discrimination ability. Not only in the field of image processing, deep learning has also yielded many achievements in dealing with point cloud data. Compared with traditional segmentation methods, deep learning based segmentation methods have following advantages: end-to-end system; easy to acceleration by using GPU; automatic learning features based on data itself; easy to extend in multi-scale feature extraction. We use deep neural networks instead of traditional approaches to pre-segment both the image and the point cloud.

A. Image Segment Net Architecture

There are many image segmentation networks, such as SegNet [6], DeepLab [7] etc. These networks appear to be different, all of them benefit from the FCN [5] network, which firstly achieves end-to-end dense image segmentation efficiently. We adopt the FCN-8s network for the pre-segmentation of the image. The architecture of the network is shown in Fig. 2 in detail.

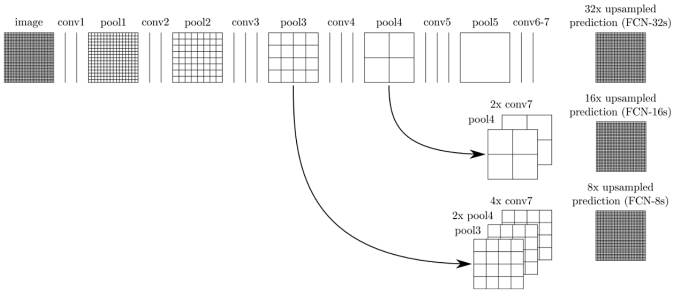


Fig. 2. FCN Architecture [5].

As shown in Fig. 2, the VGG architecture is adopted in convolution layers as conv1-conv5. The fully connected layers in original VGG are replaced with 1×1 convolution layers as conv6-conv7. The up-sampling layers are initialized with bilinear interpolation. Combining $4\times$ up-sampling of conv7 and $2\times$ up-sampling of pool4 with pool3 at stride 8 to produce a dense prediction as the same size as the original input image.

B. Point Cloud Segment Net Architecture

Because of the unordered characteristic of point cloud, the traditional image-based network is not suitable to migrate to point cloud processing directly. Although 3D-CNN such as [17] can handle point cloud, it is is very inefficient to treat the sparse data as 3D voxels. Charles et al. [18] propose a new network named PointNet to classify or segment unordered point cloud data directly. Since the first generation of PointNet does not consider the neighborhood information just as the CNN does, the extracted features of the point cloud can not accurately represent the context. Based on the work of PointNet, Charles et al. [19] expand the PointNet to PointNet++, in which a neighborhood information extraction unit is proposed to extract multi-scale features. In this paper, we use PointNet++ based architecture to pre-segment the point cloud. The network architecture is shown in Fig. 3 in detail.

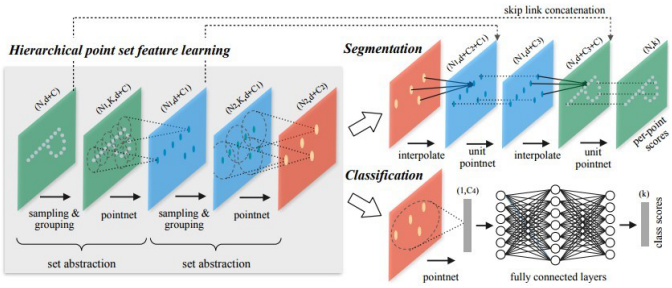


Fig. 3. PointNet++ Architecture [19].

The encoder of PointNet++ consists of three modules: The sampling layer which generates centroid points with most distance. The group layer which group local points as a subset with per centroid. The pointnet layer which can be described as a function:

$$f(x_1, x_2, ..., x_n) = g(\underset{x_i \in S}{MAX}\, h(x_i)), \tag{1}$$

where $f(\cdot)$ is a symmetric function, x_i represents a point in points set $S = \{x_1, x_2, ..., x_n\}$, $g(\cdot)$ and $h(\cdot)$ are general MLP (multi-layer perception) networks. MAX is a vector max operator that takes n vectors as input and returns a new vector of the element-wise maximum. The pointnet layer encodes a set of points to a feature vector. As same as FCN, interpolating and skip layer connection are integrated to up-sample the predictions. After up-sampling, the network predicts the score of every point for corresponding with input point cloud. It should be noted that the coordinate of the centroid is concatenated with the corresponding feature vector

in the network all the time to reserve the spatial position information of the feature.

IV. Fusion Model based on Fully Connected CRF

Road detection can be expressed as the pixel labeling problem of an image, that is, labeling each pixel with a road or non-road label. After obtaining the prior heat map of the image and point cloud by the segment networks, the labeling problem can be solved by the CRF model. The multi-sensor data can be fused by the CRF model, and the final labeling result can be obtained after optimization. In this section, we firstly introduce the traditional CRF approach in labeling problem. Then we introduce the CRF model used in this paper. Finally, we introduce the optimization of the proposed model.

A. CRF in Computer Vision

CRF is a probabilistic graph model which is widely used in labeling problems. Let $\boldsymbol{X} = \{x_1, x_2, ..., x_n\}$ denotes a sequence of states, let $\boldsymbol{Y} = \{y_1, y_2, ..., y_n\}$ denotes a sequence of observations. Given a observation sequence $\boldsymbol{Y}$, a random variable x_i in the state sequence $\boldsymbol{X}$ can obtain any label in the label set $\boldsymbol{L} = \{l_1, l_2, ..., l_k\}$. The purpose of labeling is to find the most possible state sequence $\boldsymbol{X}^*$ by solving the problem:

$$\boldsymbol{X}^* = argmax_{x \in \boldsymbol{L}} P(\boldsymbol{X}|\boldsymbol{Y}), \tag{2}$$

where $\boldsymbol{Y}$ is the observation sequence. $P(\boldsymbol{X}|\boldsymbol{Y})$ can be represented by an undirected graph $G = (V, E)$. V is the vertex set in the graph which represents the random variables of state sequence. E is the edge set in the graph which represents the correlations between random variables. According to the Hammersley-Clifford theorem, the posterior probability distribution $P(\boldsymbol{X}|\boldsymbol{Y})$ can be written as a Gibbs distribution defined on the largest cliques of G:

$$P(\boldsymbol{X}|\boldsymbol{Y}) = \frac{1}{Z(\boldsymbol{Y})} \exp(-\sum_{c \in C_G} \psi_c(\boldsymbol{x}_c, \boldsymbol{Y})), \tag{3}$$

where $\boldsymbol{x}_c$ is the largest clique of G, C_G is a set of all largest cliques of G. $\psi_c(\boldsymbol{x}_c|\boldsymbol{Y})$ is a potential function defined on the clique $\boldsymbol{x}_c$, and $Z(\boldsymbol{Y})$ is a normalized function to ensure a probability distribution. Maximizing the posterior probability $P(\boldsymbol{X}|\boldsymbol{Y})$ is equivalent to minimizing the Gibbs energy function:

$$E(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{c \in C_G} \psi_c(\boldsymbol{x}_c|\boldsymbol{Y}). \tag{4}$$

In computer vision and image processing, most CRF models are expressed as pairwise form. That is, only the unary potential and pairwise potential are considered. These models can be described as the following unified form:

$$E(\boldsymbol{X}) = \sum_{i \in V} \psi_u(x_i) + \sum_{i \in N_i} \psi_p(x_i, x_j), \tag{5}$$

where x_i is the label of vertex i. N_i represents all vertexes of G which is adjacent to vertex i. $\psi_u(\cdot)$ is the unary potential and $\psi_p(\cdot)$ is the pairwise potential. Note that the observation sequence $\boldsymbol{Y}$ is omitted for convenient writing.

B. Fully connected CRF for fusion

The traditional CRF models generally consider the output of the pixel classifier as the unary potential, and only take into account the connection of local 4 or 8 neighborhoods in the pairwise potential. This kind of assumption leads to the the CRF model is optimized locally and can not capture the global information. As a result, the pixels misclassified in the pixel classification stage are hardly to be corrected during the model optimization. Because of local smoothing, the edge of road is often not clear enough. Therefore, the unary potential of the CRF model adopted in this paper takes into account both the image priori and the lidar priori. The structural information enhances the discrimination of pixels. Furthermore, we adopt the full connection form in the pairwise potential. Each pixel is not only connected with its local neighborhood, but also connected with all of other pixels in the graph. Thus, the pairwise potential is actually defined on the whole graph. The CRF fusion model in this paper can be denoted as:

$$E(\boldsymbol{X}) = \sum_{i \in V} \psi_u^I(x_i) + \sum_{i \in V} \psi_u^L(x_i) + \sum_{\substack{i<j \\ i,j \in V}} \psi_p(x_i, x_j), \tag{6}$$

where ψ_u^I is the unary potential of the image, ψ_u^L is the unary potential of the point cloud, and ψ_p is the pairwise potential defined on all nodes of the graph.

a) Unary Potential: The unary potential is derived from the outputs of the image segmentation network and the point cloud segmentation network. The unary potential, also called data cost, only depends on the data itself and can be regarded as a priori prediction for the observation sequence. According to the definition of potential, the form of the unary potential of the image and the point cloud can be denoted as:

$$\psi_u^I(x_i) = -\log(H_p(x_i)), \tag{7}$$

$$\psi_u^L(x_i) = -\lambda \log(H_l(x_i)), \tag{8}$$

where $H_p(\cdot)$ and $H_l(\cdot)$ represent the output of the image network and point cloud network respectively. λ is a weight factor to control the two unary potentials. Because of the sparseness of point cloud, points and pixels can not be one-to-one correspondence in the image plane. In order to construct a fully connected CRF in the image plane, we adopt an up-sampling strategy to get a dense probability map of point cloud instead of simply setting a fixed probability at the position without a corresponding lidar point. The up-sampling strategy will be described in detail in the following sections.

b) Pairwise Potential: The definition of the pairwise potential is generally in the form of a weighted sum of Gaussian functions. It can model the interdependence of connected nodes in the graph. The pairwise potential, also called smooth cost, actually encourages local consistency of the labeling

result. It has a smoothing effect on the labeling result. The pairwise potential has the following general form:

$$\psi_p(x_i, x_j) = \delta(x_i, x_j) \sum_{m=1}^{M} \omega_m k_m(\boldsymbol{f}_i, \boldsymbol{f}_j), \tag{9}$$

where $\delta(\cdot)$ is a label compatibility function. $k_m(\cdot)$ is a Gaussian function. M is the number of functions, and ω_m is the weighting factor of each function.

Generally, $\delta(\cdot)$ is an indicator function:

$$\delta(x_i, x_j) = \begin{cases} 1, & if \quad x_i \neq x_j \\ 0, & otherwise. \end{cases} \tag{10}$$

And $k_m(\cdot)$ can be denoted as:

$$k_m(\boldsymbol{f}_i, \boldsymbol{f}_j) = \exp(-\frac{1}{2}(\boldsymbol{f}_i - \boldsymbol{f}_j)^{\mathrm{T}} \Lambda_m (\boldsymbol{f}_i - \boldsymbol{f}_j)), \tag{11}$$

where $\boldsymbol{f}_i$ is the feature vector of pixel i. Λ_m is a positive-definite precision matrix which defines the shape of kernel.

In image labeling problem, most of pairwise potentials are defined as the linear combination of a Gaussian spatial kernel and a bilateral filter kernel:

$$\begin{aligned} \psi_p(x_i, x_j) = \delta(x_i, x_j) \Big\{ & \omega_1 \exp(-\frac{\|p_i - p_j\|_2^2}{2\theta_\alpha^2} - \frac{\|I_i - I_j\|_2^2}{2\theta_\beta^2}) \\ & + \omega_2 \exp(-\frac{\|p_i - p_j\|_2^2}{2\theta_\gamma^2}) \Big\}, \end{aligned} \tag{12}$$

where p_i is the position vector of pixel i in the image coordinate, and I_i is the color vector of the pixel i in RGB space. θ_α, θ_β and θ_γ are the parameters of kernels. The former term is a bilateral filter kernel called *appearance kernel*, which encourages nearby pixels with similar color to take the same label. The latter term is a Gaussian space kernel called *smooth kernerl*, which can remove small isolated regions. However, this model only considers pixel position and RGB color space. It leads that the model may be easily misled by the shadows and the road curbs.

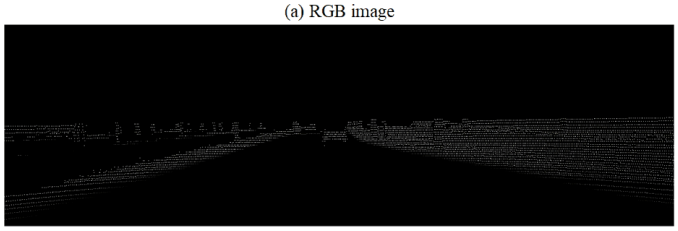
Height information is very discriminative to detect the ground plane and big obstacles such as vehicles and buildings. It is based on the assumption that the road area is the largest flat area in front of the car, that is the road area is highly consistent in height at least. In view of this, we propose a new pairwise potential for road detection, which takes into account not only the location and color space of pixel, but also the height information obtained from the point cloud. In this way, the color information and the spatial structure information are thought about simultaneously. A sparse height image can be obtained by mapping the lidar points to the image plane. Thanks to the work of Diebel et al. [20], we can obtain a dense height image from the sparse one according to the Markov-based up-sampling method they proposed. Similar to the image labeling problem, the loss function as follow will be minimized in the Markov random field.

$$E(h) = \sum_{i \in L} k(h_i - h_i')^2 + \sum_{i \in V} \sum_{j \in N_i} w(h_i - h_j)^2, \tag{13}$$

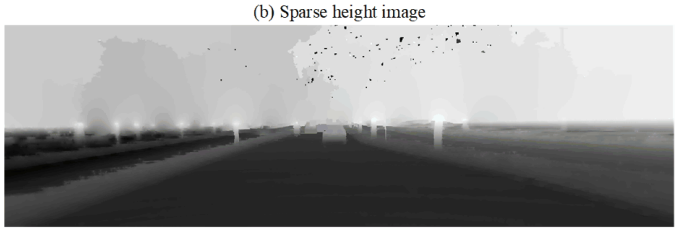
$$w = \exp(-c\|x_i - x_j\|_2^2), \tag{14}$$

where h is the dense height value to be up-sampled, and h' is the sparse height value pixel before up-sampling. x is the gray value of the pixel. Both k and c are the weight factors of the model. L is the set of vertexes with 3D points. The objective function seems to be NP-hard to solve in the problem of image labeling. But fortunately, the objective function is a continuous and derivable function in the problem of height up-sampling. So the objective function can be simply solved by the gradient-based methods. Specific deduction can be found in [20]. Besides, we can get both a dense height image and a dense probability map by using this method. The up-sampled result of the sparse height image is shown in Fig. 4.

(a) RGB image



(b) Sparse height image



(c) Dense height image (upsampled)

Fig. 4. Up-sampled height image by the Markov-based method.

After getting the dense height image, we define the pairwise potential in this paper as the following form:

$$\begin{aligned} \psi_p(x_i, x_j) = \delta(x_i, x_j) \Big\{ & \omega_1 \exp(-\frac{\|p_i - p_j\|_2^2}{2\theta_\alpha^2} - \frac{\|I_i - I_j\|_2^2}{2\theta_\beta^2}) \\ & + \omega_2 \exp(-\frac{\|p_i - p_j\|_2^2}{2\theta_\gamma^2} - \frac{\|H_i - H_j\|_2^2}{2\theta_\delta^2}) \\ & + \omega_3 \exp(-\frac{\|p_i - p_j\|_2^2}{2\theta_\epsilon^2}) \Big\}, \end{aligned} \tag{15}$$

where H_i is the height value of pixel i in the lidar coordinate. θ_α, θ_β, θ_γ, θ_δ, θ_ϵ are the parameters of kernels. The first term

is a *color bilateral kernel*, which encourages nearby pixels with similar color to take the same label. In the same way, the second term is a *height bilateral kernel*, which encourages nearby pixels with similar height to take the same label and is robust to the shadows and the road curbs. And third term is a *spatial kernel*, which can remove small isolated regions as previous models do.

C. Optimization

The CRF model of image labeling problem is NP-hard to solve. The common approximation methods are graph-cut based algorithm such as $\alpha-expansion$, $\alpha\beta-swap$, etc. [21]. These methods minimizes the energy function by exchanging labels and applying max-flow algorithm to compute the energy value after exchanging iteratively until convergence or the max iteration is met. The graph-cut algorithm has a wide range of applications for the local neighborhood CRF [4]. However, the CRF used in this paper is a full graph, in which the number of edges is gigantic. So the graph-cut based algorithms become very inefficient in this case. Krhenbhl et al. [22] propose a method based on *Mean Field Approximation* to solve the fully connected CRF efficiently. This method states that if the pairwise potential in the CRF model is a combination of Gaussian functions, then the message passing step in mean field approximation algorithm can be expressed as a convolution with Gaussian kernel. The computational complexity can be reduced in this way and the proposed model is solved by this strategy.

V. Experiment

A. Dataset

The proposed model is evaluated on KITTI Road Benchmark [23], whose data is acquired by an UGV equipped with sensors such as stereo cameras, lidar and IMU. It is a common dataset established for environmental perception algorithm exclusively. KITTI dataset contains gray-scale and color stereo images, Velodyne-HDL-64E lidar point cloud data, inertial navigation information, and the corresponding calibration parameters. It is convenient to convert the sensor data freely between the coordinate systems of each sensor. KITTI Road dataset contains 289 frames training data and 290 frames testing data under different scenes, including UM (urban marked), UMM (urban multiple marked), and UU (urban unmarked). We consider only the road detection task, so the lane detection task is ignored.

Because the groundtruth of KITTI Road test set is not provided, the 289 frames training data are averaged to a training set and a validation set. The evaluation indexes used in this paper are consistent with KITTI Road Benchmark, including MaxF (maximum F1-measure), REC (recall), PRE (precision), AP (average precision), FPR (false positive rate), FNR (false negative rate).

B. Parameter Setting

About network settings, the FCN-8s is used for image pre-segmentation. The VGG-16 is used for feature encoding. Three skip layer connections which can integrate features of three scale are used for up-sampling. We use the pre-training weights on ImageNet to initialize the encoder layers and the bilinear interpolation up-sampling to initialize the decode layers. The training epoch is set to 200. The learning rate is set to 1e-4 for the VGG pre-trained parameters and 1e-2 for other parameters. The PointNet++ based network architecture is used for point cloud pre-segmentation. We only use about 20000 points per frame which is in the field of camera view. Four abstraction layers are adopted to encode point features. And four propagation layers are adopted to up-sample the decoded point features. The label of the point cloud data can be easily obtained from the groundtruth image by cross calibration. In point cloud training, the training epoch is set to 200, and the learning rate is initially set to 1e-2 and divided by 10 every 100 epochs.

The parameters of the model include λ, ω_1, ω_2, ω_3, and θ_α, θ_β, θ_γ, θ_δ, θ_ϵ are empirical value. In order to obtain the best model parameters, the cross-validation method is used to obtain the optimal model parameters from the validation data. In our experiment, λ is set to 1.0, ω_1, ω_2, ω_3 are set to 100, 60, 30, respectively. θ_α, θ_β, θ_γ, θ_δ, θ_ϵ are the covariance matrix of each Gaussian actually, they are set to $diag(9,3)$, $diag(30,10,10)$, $diag(9,3)$, 5, $diag(9,3)$, respectively.

C. Evaluation

In this section, we conduct a scientific assessment for the proposed model. Since the groundtruth of the test set is not provided in KITTI Road dataset, the validation experiments are conducted on the training set only. The final results submitted to KITTI website are from the experiments on the complete training set and testing set of KITTI Road dataset.

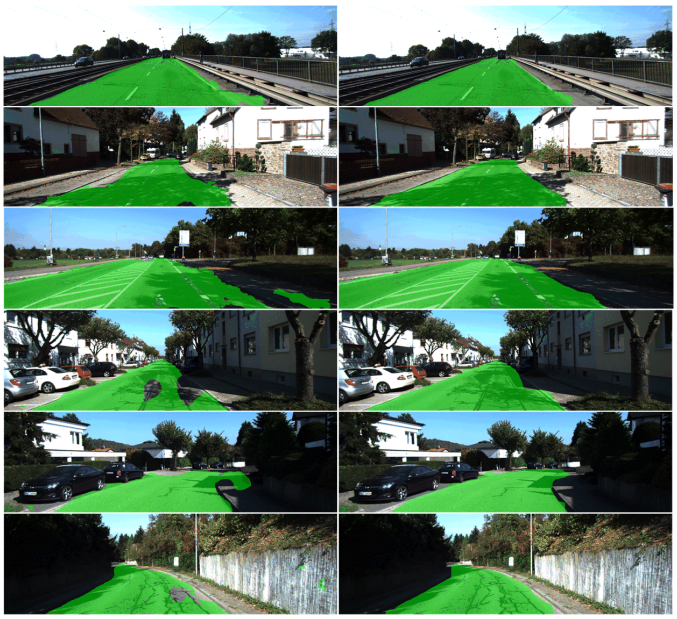


Fig. 5. Comparison results of the proposed method between before fusion and after fusion.

To demonstrate the effectiveness of the proposed fusion method, we show some comparison visualizations before/after fusion in Fig. 5. The left and right columns respectively denote the results before and after fusion. According to the visual comparison results, the proposed model can perform well in complex scenes. Specifically, the proposed model can deal with the edge area correctly.

Furthermore, we also compare the proposed method with some other CRF fusion models, such as the traditional local connected pairwise CRF model, the high order CRF model, and the fusion hierarchical CRF model. The comparison of results are shown in TABLE I in term of the indexes mentioned above. It can be seen that the proposed model has a significant improvement over some traditional CRF models. Specially, all of six indexes are improved based on the experiment results.

In order to enhance the persuasion of the experiments, we also conduct the experiment on the standard test set of KITTI Road dataset and submit the experimental results to KITTI website. Because the results must be evaluated in bird view in KITTI Benchmark, while our model is based on perspective view. So the experimental results in perspective are converted to bird view through the camera calibration parameters before submitting. The proposed model performs well on KITTI Road test set. The statistical results are shown in TABLE II in detail. The main index MaxF reaches 95.40% on the UMM_ROAD dataset and the average value of MaxF reaches 93.38% on the whole test set.

Fig. 6. Performance on KITTI of the proposed method (Perspective).

Some visualization results are given in Fig. 6 and Fig. 7. Fig. 6 shows the results of perspective on KITTI Road test set. The green indicates true positive, the blue indicates false negative and the red indicates false positive, respectively. Fig. 7 is the results of bird view on KITTI Road test set. The green indicates true positive, the blue indicates false negative and the red indicates false positive. The size of the bird view grid is about 20m wide and 40m high, which represents a range of 20m from the left of the car to the right of the car and 40m in the front of the car. Each row of Fig. 7 is corresponding to each column of Fig. 6.

We also compare the proposed model with some current popular methods, most of which are based on CRF or fusion of image and lidar. These methods include MixedCRF, HybirdCRF, LidarHisto, FusedCRF, SPARY, ProbBoost, GRES3D+VELO, RES3D-Velo, and geo+gpr+crf, FHM, etc.

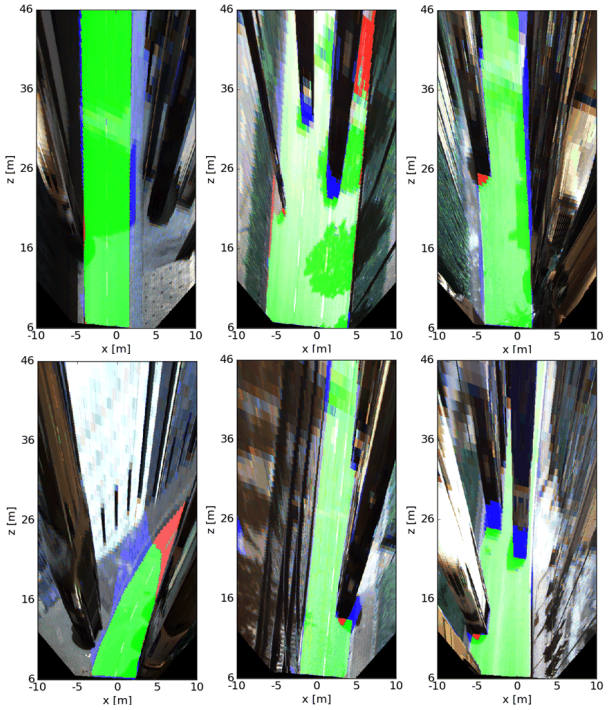


Fig. 7. Performance on KITTI of the proposed method (Bird View).

The statistical results are shown in TABLE III. Compared with some current popular methods, the proposed method outperforms all other methods on the main metric of MaxF. According to the results of TABLE III, the proposed model has certain advantages than some current popular methods.

VI. Conclusion

In this paper, we propose a fusion model based on deep learning and fully connected CRF for road detection. It can not only fuse image and point cloud data in unary potential, but also integrate structure information in pairwise potential. The proposed model has three advantages:

- Local and global features of image and point cloud are captured by deep network.
- Height image used in pairwise potential makes the proposed model more robust to shadows and road curbs.
- Fully connected CRF increases the error correction capability of the fusion model.

Experiments on KITTI Road dataset demonstrate the effectiveness of the proposed method. It can be seen that the proposed model performs better in general scenes. A promising performance can be alos achieved even in some complex scenes.

Acknowledgment

This work is partially supported by *National Natural Science Foundation of China* under Grant Nos. 61872188, 61703209, U1713208, 61972204, 61672287, 61861136011, and 61773215.

TABLE I
COMPARISON RESULTS BETWEEN THE PROPOSED MODEL AND SOME TRADITIONAL MODELS.

Methods	**Datasets**	MaxF(%)	AP(%)	PRE(%)	REC(%)	FPR(%)	FNR(%)
Pairwise CRF [15]	UM_ROAD	93.01	82.93	89.60	96.69	2.19	3.31
	UMM_ROAD	95.14	87.27	93.62	96.72	2.05	3.28
	UU_ROAD	90.99	80.46	87.15	95.18	2.20	4.82
High Order CRF [4]	UM_ROAD	93.71	83.41	90.12	97.60	2.09	2.40
	UMM_ROAD	95.75	87.94	94.36	97.19	1.81	2.81
	UU_ROAD	91.54	80.45	87.14	96.41	2.23	3.59
FHM [24]	UM_ROAD	94.32	84.45	91.27	97.60	1.82	2.40
	UMM_ROAD	96.08	88.46	94.94	97.26	1.62	2.74
	UU_ROAD	92.83	82.53	89.42	**96.50**	1.79	**3.50**
Ours	UM_ROAD	**95.51**	**85.97**	**92.89**	**98.28**	**1.52**	**1.72**
	UMM_ROAD	**96.26**	**88.60**	**95.24**	**97.29**	**1.39**	**2.71**
	UU_ROAD	**94.29**	**85.32**	**92.44**	96.22	**1.30**	3.78

TABLE II
PERFORMANCE OF THE PROPOSED METHOD ON KITTI TEST SET.

Datasets	MaxF(%)	AP(%)	PRE(%)	REC(%)	FPR(%)	FNR(%)
UM_ROAD	92.54	82.80	87.95	97.64	6.09	2.36
UMM_ROAD	95.40	89.30	92.99	97.93	8.11	2.07
UU_ROAD	90.68	80.81	86.44	95.37	4.88	4.63
URBAN_ROAD	93.38	84.90	89.83	97.22	6.06	2.78

TABLE III
COMPARISON OF RESULTS BETWEEN THE PROPOSED MODEL AND SEVERAL CURRENT POPULAR METHODS.

Methods	MaxF(%)	AP(%)	PRE(%)	REC(%)	FPR(%)	FNR(%)
MixedCRF [25]	90.59	84.24	89.11	92.13	6.20	7.87
HybirdCRF [16]	90.81	86.01	91.05	90.57	4.90	9.43
LidarHisto [11]	90.67	84.79	**93.06**	88.41	**3.63**	11.59
FusedCRF [15]	88.25	79.24	83.62	93.44	10.08	6.56
ProbBoost [26]	87.78	77.30	86.59	89.01	7.60	10.99
SPARAY [27]	87.09	**91.12**	87.10	87.08	7.10	12.92
GRES3D+VELO [28]	86.07	84.34	82.16	90.38	10.81	9.62
RES3D-Velo [13]	86.58	78.34	82.63	90.92	10.53	9.08
geo+gpr+crf [29]	85.56	74.21	82.81	88.50	10.12	11.50
FHM [24]	90.88	83.10	87.86	94.12	7.16	5.88
Ours	**93.38**	84.90	89.83	**97.22**	6.06	**2.78**

REFERENCES

[1] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost : joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *European Conference on Computer Vision*, 2006, pp. 1–15. I, II

[2] J. Fritsch, T. Khnl, and F. Kummert, "Monocular road terrain detection by combining visual and spatial information," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 4, pp. 1586–1596, 2014. II

[3] S. Swarup, S. Swarup, S. Swarup, S. Swarup, and S. Swarup, "Crf based method for curb detection using semantic cues and stereo depth," in *Tenth Indian Conference on Computer Vision, Graphics and Image Processing*, 2016, p. 41. II

[4] L. Ladickỳ, C. Russell, P. Kohli, and P. H. Torr, "Associative hierarchical crfs for object class image segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2009, pp. 739–746. II, IV-C, I

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2014. II, III-A, 2

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for scene segmentation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. PP, no. 99, pp. 2481–2495, 2017. II, III-A

[7] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *Computer Science*, no. 4, pp. 357–361, 2014. II, III-A

[8] D. F. L. C Fernandez, R Izquierdo, "A comparative analysis of decision trees based classifiers for road detection in urban environments," in *IEEE International Conference on Intelligent Transportation Systems*, 2015, pp. 719–724. II

[9] F. Moosmann, O. Pink, and C. Stiller, "Segmentation of 3d lidar data in non-flat urban environments using a local convexity criterion," in *Intelligent Vehicles Symposium*, 2009, pp. 215–220. II

[10] T. Chen, B. Dai, R. Wang, and D. Liu, "Gaussian-process-based real-

time ground segmentation for autonomous land vehicles," *Journal of Intelligent & Robotic Systems*, vol. 76, no. 3-4, pp. 563–582, 2014. II

[11] L. Chen, J. Yang, and H. Kong, "Lidar-histogram for fast road and obstacle detection," in *IEEE International Conference on Robotics and Automation*, 2017, pp. 1343–1348. II, III

[12] G. B. Vitor, D. A. Lima, A. C. Victorino, and J. V. Ferreira, "A 2d/3d vision based approach applied to road detection in urban environments," in *Intelligent Vehicles Symposium*, 2013, pp. 952–957. II

[13] P. Y. Shinzato, D. F. Wolf, and C. Stiller, "Road terrain detection: Avoiding common obstacle detection assumptions using sensor fusion," in *Intelligent Vehicles Symposium Proceedings*, 2014, pp. 687–692. II, III

[14] X. Hu, F. S. A. Rodriguez, and A. Gepperth, "A multi-modal system for road detection and segmentation," in *Intelligent Vehicles Symposium Proceedings*, 2014, pp. 1365–1370. II

[15] L. Xiao, B. Dai, D. Liu, T. Hu, and T. Wu, "Crf based road detection with multi-sensor fusion," in *Intelligent Vehicles Symposium*, 2015, pp. 192–198. II, I, III

[16] L. Xiao, R. Wang, B. Dai, Y. Fang, D. Liu, and T. Wu, "Hybrid conditional random field based camera-lidar fusion for road detection," *Information Sciences*, 2017. II, III

[17] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *Ieee/rsj International Conference on Intelligent Robots and Systems*, 2015, pp. 922–928. III-B

[18] R. Q. Charles, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," 2016. III-B

[19] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," 2017. III-B, 3

[20] J. Diebel and S. Thrun, "An application of markov random fields to range sensing," *Advances in Neural Information Processing Systems*, pp. 291–298, 2005. IV-B0b, IV-B0b

[21] C. Russell, L. Ladicky, P. Kohli, and P. H. S. Torr, "Exact and approximate inference in associative hierarchical networks using graph cuts," in *UAI 2010, Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, Catalina Island, Ca, Usa, July*, 2010, pp. 501–508. IV-C

[22] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in Neural Information Processing Systems 24*. Curran Associates, Inc., 2011, pp. 109–117. IV-C

[23] J. Fritsch, T. Kuehnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *International Conference on Intelligent Transportation Systems (ITSC)*, 2013. V-A

[24] Y. Fei, W. Huan, and J. Zhong, "Fusion hierarchical condition random fields based road segmentation model," *Robot*, no. 6, p. 14, 2018. I, III

[25] X. Han, H. Wang, J. Lu, and C. Zhao, "Road detection based on the fusion of lidar and image data," *International Journal of Advanced Robotic Systems*, vol. 14, no. 6, p. 1729881417738102, 2017. III

[26] G. B. Vitor, A. C. Victorino, and J. V. Ferreira, "A probabilistic distribution approach for the classification of urban roads in complex environments," in *Proceedings of the Workshop on Modelling, Estimation, Perception and Control of All Terrain Mobile Robots on IEEE International Conference on Robotics and Automation*, 2014. III

[27] T. Kühnl, F. Kummert, and J. Fritsch, "Spatial ray features for real-time ego-lane extraction," in *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*. IEEE, 2012, pp. 288–293. III

[28] P. Y. Shinzato, "Estimation of obstacles and road area with sparse 3d points," *Institute of Mathematics and Computer Science/University of Sao Paulo*, 2015. III

[29] Z. Xiao, B. Dai, H. Li, T. Wu, X. Xu, Y. Zeng, and T. Chen, "Gaussian process regression-based robust free space detection for autonomous vehicle by 3-d point cloud and 2-d appearance information fusion," *International Journal of Advanced Robotic Systems*, vol. 14, no. 4, p. 1729881417717058, 2017. III