Lidar-Camera Fusion For Road Detection Using Recurrent Conditional Random Field Model

Lele Wang

University of Shanghai for Science & Technology

Yingping Huang (✉ huangyingping@usst.edu.cn)

University of Shanghai for Science & Technology

Research Article

Keywords: sensor fusion, semantic segmentation, road detection, autonomous vehicles, condition random field

Posted Date: March 8th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1392000/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

LiDAR-Camera Fusion for Road Detection Using Recurrent Conditional Random Field Model

Lele Wang[†], Yingping Huang^{*†}

- School of Optical-Electrical and Computer Engineering, University of Shanghai for Science & Technology, Shanghai 200093, China. E-mail: 1491658722@qq.com, E-mail: huangyingping@usst.edu.cn

^{*} Correspondence author. [†] Co-first authors.

Abstract: Reliable road detection is an essential task in autonomous driving systems. Two categories of sensors are commonly used: camera and LiDAR, each of which can provide corresponding supplemental. Nevertheless, existing sensor fusion methods do not fully utilize multimodal data. Most of them are dominated by image or taking point cloud as a supplement rather than making the best of them, the correlation between modalities is ignored. This paper proposes **Recurrent Conditional Random Field Model (R-CRF)** for road detection by fusing image and point cloud segmentation. The R-CRF integrates results (information) from modalities in a probabilistic way. Each modality is independently processed with its own semantic segmentation network. The probability scores obtained are considered as unary term for individual pixel nodes in random field while RGB image and the densified LiDAR-images are used as pairwise terms. The energy function is then iteratively optimized by mean-field variational inference so that we can refine the labeling results by exploiting fully-connected graphs of the RGB image and LiDAR-images. Extensive experiments are conducted on KITTI-Road dataset, and these demonstrate the method achieves competitive performance.

Key words—sensor fusion, semantic segmentation, road detection, autonomous vehicles, condition random field

1. INTRODUCTION

Road detection is a prerequisite for autonomous driving. Autonomous vehicles must be capable of distinguishing road from obstacles in various environments. Autonomous vehicles are often equipped with multiple sensors for environment perception, among which Light Detection and Ranging (LiDAR) and camera are the most informative and commonly used. Road detection has been studied for decades and can be categorized into three types of methods including camera-based, LiDAR-based, and fusion-based methods. Camera-based methods can be done using images through RGB data [1-3]. LiDAR-based methods segment road area from 3D point cloud as top view images [7], spherical view images [8], and the front view images [4,9].

Multi-modality sensor fusion [10-26] has been a way to improve the perception robustness in autonomous vehicles. Camera provides texture and colors, but the nature of passive sensor makes it susceptible to variations in environment lighting. Compared to the camera, LiDAR is not affected by season and illuminate conditions and offers 3D geometry information to complement visual data shortcomings. The fusion-based methods have been intended to overwhelm the weakness of each-sensor case and exhibited promising performance. Semantic segmentation is an efficient style to analyze the sensor inputs for autonomous driving, the image and point cloud are classified into semantic classes. Conditional Random Field (CRF) is an effective tool to integrate results obtained from each sensor and therefore refine the segmentation result. CRF [19] is a discriminative probability model. Since labels of pixels can be regarded as random variable, CRF can be used to model labeling problem. Normally, the CRF is defined as an undirected graph with pixels as nodes. It can be solved by an approximate graph inference algorithm through minimizing

the energy function. The function contains unary and pairwise potential. The unary potential is only concerned with the node itself and determines probability of node being labeled. The pairwise potential describes interactions between neighboring nodes and is defined as similarity between them.

Existing works for road detection using CRF [20-26] have the following issues: 1) These works do not fully make use of results (information) from two sensors. As to energy function, method [20] only uses the result generated by RGB images as input of the unary term. Other works [21-22,25] make use of both results to define the unary term but result generated by point cloud takes effect only as a supplement because point cloud data is sparse. In the pairwise term of the energy, most works [21,23,25] only make use of interactions between neighboring nodes of image and do not take the point cloud information into consideration at all. In summary, the correlation between two modalities of data is ignored. 2) Existing works use the graph cut-based algorithm to conduct graph inference. However, the graph cut-based inference is only applicable in a locally-connected graph which only considers the local interaction. Ideally, the undirected graph should be a fully-connected graph that considers the local and globe interaction of the RGB image or LiDAR image.

To cover the issues mentioned above, *Recurrent Conditional Random Field Model (R-CRF)* is proposed which employs mean-field variational inference to conduct graph inference rather than graph cut-based algorithm. Formulated as a recurrent model, mean-field variational inference performs iterative optimization through a series of message-passing steps, each step updates one variable by aggregating information from all other variables. Considering that pairwise potential can be seen as a compounding of linear combinations of Gaussian kernels,

message passing step in mean-field variational inference can be stated as a convolution. R-CRF using mean-field variational inference dramatically reduces the computational complexity, therefore, enables us to conduct graph inference in a form of a fully-connected graph. On the other hand, the proposed R-CRF fully makes use of results (information) of two sensors. It takes probability scores generated by two modalities of data as the unary potential term, both the RGB image and the densified LiDAR-images are utilized as pairwise potential terms to encode the contextual consistency. Followed by such a fusion process, the proposed model possesses a big error correction capability.

Compared to the existing literature, the major contributions can be concluded as follows:

- 1) The R-CRF model is proposed to fully integrate results (information) of multi-sensor data (images and point clouds) in a probabilistic way. Specifically, the densified LiDAR-image and RGB image are added to the pairwise input reasonably to encode the contextual consistency.

- 2) The mean-field variational inference is utilized to solve the graph inference problem rather than the graph cut-based inference so that we can refine the labeling results through a fully-connected graph that makes use of the local and globe interaction of the RGB image or LiDAR image. Specifically, the message passing step in the inference is reformulated to a convolution with truncated Gaussian kernel.

- 3) We conduct extensive experiments on KITTI road benchmark, results indicate approach in this paper is robust to environment, and it achieves a promising detection performance.

The paper is formed: Section 2 surveys related work in the field of road detection. Section 3 introduces the data pre-processing, semantic segmentation network, and the details of method. In section 4, experiment results on the KITTI dataset are displayed and analyzation of current popular methods are listed. Section 5 is summarization.

2. RELATED WORK

Various approaches have been developed and can be divided into two groups in terms of the use of sensors, including one-sensor-based, and multiple sensors fusion-based methods.

2.1 One -sensor-Based Road Detection

Departing from Fully Convolutional Network (FCN), diversely structures have been proposed to provide accurately pixel-wise prediction result for the task of road detection. MultiNet [1] was proposed through a unified architecture for multi-tasks. An encoder and decoder scheme, RBNet [2], a DCNN was applied to recollect features at different scales. Additional driving scene images were generated by Fan [3]. However, quality of image is heavily impacted by weather conditions, reducing the accuracy.

Other related approaches are focusing on using point cloud, which utilize the geometric properties measured from sparse range data. Compared with the diversities in images, the geometric characteristics in LiDAR are relatively simple and easier to learn. Fernandes et al. [4] obtained an accurate road estimation result through the sliding window technique, and utilized morphological processing to classify the road from point cloud. Projection-based method [5-6] projected point clouds into BEV view or spherical front view. These representation are adequate to realtime systems. LoDNN [7]

transformed unstructured LiDAR data into top-view representation by basic statistics, such as point number, mean, average, standard deviation, minimum and maximum, and then those maps were employed as input for CNN to achieve the desired result. Lyu et al. [8] arranged the points into specific views as input, then the proposed FCN was implemented on an FPGA. Gu et al. [9] obtained inverse map and acquired the approximate road regions by extracting the vertical and horizontal histograms.

2.2 Multiple-sensor fusion Based Road Detection

For robust environment perception in autonomous vehicles and to get rid of inherent disadvantages and absorb essence of various sensors, data-fusion approaches for road detection can be classified into following three levels:

- 1) Early level fusion: Different types of sensor data are combined to produce a new kind of data through data alignment, preserving all information. Wulff et al. [10] proposed the UGrid-Fused approach, a multi-dimensional occupation grid representation based on BEV, which could be imported into the FCN. Each cell in UGrid-Fused contains 15 statistics, including binary map, count map, obstacle map, six height measurement maps and six reflectivity intensity maps. Byeonjun Yu et al. [11] transformed both modalities bird's eye view to facilitate fast data transfuser.[10] Lee and Park [12] focused on the idea of contracting the size of the inputs and expanding perceptual field of the network. The two modalities are transformed into spherical coordinates, and height data of point cloud and three R, G, B channels are superimposed on channels and subsequently fed into modified SegNet network.

- 2) Middle level fusion: Features from multiple sensor data are used to accompaniment scenario. Chen et al. [13] solved feature space mismatch by performing altitude difference on LiDAR data, and then cascaded fusion structure was implemented based on DCNN. Caltagirone et al. [14] [13]used RGB image data and the interpolated 2D LiDAR-image data by [15] into a modified CNN. These fusion strategies in deep network are essentially an Addition /Concatenation.

- 3) Late level fusion: The data has been operated individually by each of their networks, then integrates the 2D and 3D results based on mutual relationship, or probabilistic modeling. RES3D-Velo [16] projected point clouds to image space and constructed a graph by Delaunay Triangulation, and applied spatial relationship to discriminate obstacle. Since it only uses cross-calibration parameter to get points, the color information is not utilized at all. After projecting points to the image, Hu et al. [17] employed plane estimation to identify points on the ground plane. The Gaussian model was used to learn image features, and pixels were also classified through this model. However, this segmentation process is implemented only on image, which has a large limitation. Jihun Park et al. [18] proposed a drivable region identification for dirt roads by fusing semantic segmentaions of modalities. The two segmentation results are integrated in BEV grid.

Current popular CRF-based [19] methods are proposed for road detection. The fusion with CRF [20] is done at the unary stage, CRF is only used as a post-processing for superpixel labeling. FusedCRF [21] proposes boosting classifiers for two modalities, result of LiDAR classifier is

only available as an additional observation. Pairwise term only considers the image difference between adjacent pixels. A hybrid model [22], an advanced CRF fusion model, further considers the interactions between 3D points, image pixels and one between them. The results are optimized with sub-CRFs. The features of each sensor are extracted traditionally. Due to the sparsity of LiDAR data, the imbalance still exists. Gu et al. [23] put forward modified convolutional network (IDA-FCN) on RGB image and a line-scanning strategy on point cloud, late fusion was performed, the LiDAR result still worked as supplementary as in [21]. The depth images generated by joint bilateral filters [24], and features of both modalities are extracted and also given into the Adaboost classifier for coarse result. The fine results are obtained by CRF operation. Gu et al. also [25] applied a Fast height-difference-based approach to generate dense results in a spherical view to blend the outputs of the two modalities in a balanced way. The energy contained 2D unary potential, 3D potential, and 2D-3D pairwise potential. Based on conference [25], this author further considered the distribution of projection points and proposed an improved

Delaunay triangular upsampling strategy [26].

3. METHOD

This paper illustrates architecture in Fig. 1. Both modalities are aligned through cross-calibration, also corresponding depth and height image are generated. The generated LiDAR maps are integrated into pairwise potential in the R-CRF model as described in the next. The RGB image is input into Deeplab V3+ semantic segmentation network while the 3D point clouds are input into PointNet segmentation network. The segmentation results generated from the two networks are probability scores for pixels. The proposed Recurrent Conditional Random Field Model is then followed to integrate results (information) of two modalities of data. Specifically, the R-CRF model takes segmentation results as unary term. Meanwhile, it adds the RGB image, densified LiDAR depth and height images as pairwise term to make the proposed approach more robust. At last, the proposed method is iteratively optimized by mean-field variational inference.

Fig. 1 Method framework in this paper. The output of the image space can be seen as the terminal result for the test.

3.1 Data preprocessing

LiDAR scans the surrounding environment and obtains a large number of point clouds. In order to extract meaningful point cloud that corresponds to pixels, it is necessary to pre-process the data and remove the redundant points.

A. Data Alignment

Both point cloud and RGB image are organized by the different data structure, also have different coordinate systems. LiDAR data consists of numerous points in real-world, and each LiDAR point is identified by 3D coordinate vector. The RGB image consists of pixels and each pixel is described by an RGB value. In this section, the alignment is introduced. Points $P_{lidar} = (x_l, y_l, z_l, 1)^T$ in 3D LiDAR coordinate system are transformed into 3D points $P_{cam} = (x_c, y_c, z_c, 1)^T$ in camera coordinate. 3D points in camera coordinate system P_{cam} with $z_c > 0$ (front view of

camera) are turned into $p_{cam} = (u_c, v_c, 1)^T$ in image coordinate. The transformation equation is as follows:

$$P_{cam} = R_{rect}^0 \cdot T_{velo}^{cam} \cdot P_{lidar} \quad (1)$$

$$p_{cam} = T_{proj} \cdot P_{cam} \quad (2)$$

$$T_{velo}^{cam} = \begin{bmatrix} R_{velo}^{cam} & t_{velo}^{cam} \\ 0 & 1 \end{bmatrix} \quad (3)$$

where R_{rect}^0 is rotation matrix, T_{velo}^{cam} is transformation matrix, and T_{proj} is the projection matrix.

The above transformation is applied to each point. Note that points with positive Z-value are remained, and projected inside the field of view are dropped. Fig. 2 illustrates an illustration of data alignment, includes image (top left), data alignment (top right), a point cloud by LiDAR scanner in 3D real world colored by height (bottom left), and LiDAR (FOV

of image, bottom right).

Fig 2. Illustration of image and the corresponding point clouds alignment. The top left is RGB image of camera with 1242x375 pixels. The bottom left is point cloud of LiDAR with 64 channels in 3D real world. On the bottom right is point cloud (only the data overlaps the front view is displayed). The top right is the demo of point cloud and RGB image alignment, color indicates the distance: Red dots are near and blue dots are far.

B. Dense LiDAR-image Maps Generation

After transformation, three-channel tensors with the same dimension are received. Each channel encodes 3D spatial coordinates. Due to the sparse nature of LiDAR data, projected points with corresponding plane is much sparser than associated image, thus the sparse LiDAR-image representation is then processed to generate the dense one. As shown in Figure 3, we utilize the strategy [15] to get

dense depth image as shown in Fig. 3(e) and the height transformation operation [13] to obtain height difference image as illustrated in Fig. 3(f) that could better preserve characteristics. In Fig. 3(e), pixel values become larger or brighter as distance increases. While road and non-road areas can be similar, height map is very helpful in distinguishing road areas as road is usually lower in height than on-road.

Fig 3. Dense LiDAR-image generation. a. RGB image, b. Projective image (RGB image with superimposed sparse point cloud), c. Generated sparse LiDAR depth image, d. Generated sparse LiDAR height image, e. Generated dense LiDAR depth image, f. Generated dense LiDAR height image.

C. LiDAR samples labeling

Labeling point cloud is extremely labor-intensive. Considering modalities are already aligned, label of corresponding point clouds can be easily obtained from ground truth image. The equation presents as follows:

$$Label_{LiDAR}^i = \begin{cases} 1, & \text{if } Label_{Image}^{T_{LiDARtoImage} \times LiDAR^i} = \\ 0, & \text{otherwise} \end{cases}$$

where $Label_{LiDAR}^i$ indicates label of i th point cloud.

In addition, $Label_{Image}^{T_{LiDARtoImage} \times LiDAR^i} =$ road area means that semantic label ($Label_{Image}$) of projected image pixel of i th point cloud ($T_{LiDARtoImage} \times$

$LiDAR^i$) is road. Fig. 4 illustrates labeling results.

Fig.4 Illustration of labeling .

3.2 RGB image and point-cloud detection

To make better use of both modalities, before the R-CRF, each modality should be trained through types of semantic segmentation algorithms independently to identify road area. PointNet [28] is a pioneer in consuming 3D point clouds. It straightly models disordered point sets and captures local and global point features via MLP layers. After data alignment, point clouds within the image's field are extracted, the processed point clouds are analyzed through 3D segmentation network PointNet [28], it distinguishes point clouds into two classes, road points with other points.

DeepLabV3+ [29] is an existing competitive work of image semantic segmentation. It takes advantages of encoder-decoder structure to connect different-level features at different scales. It classifies all pixels into semantic classes. To speed up inference procession, a lightweight network MobileNet-V2 is used as backbone. Through two types of semantic segmentation networks, probabilistic scores could be obtained.

3.3 Recurrent Conditional Random Field

A. The general CRF-based labeling in Computer Vision

Conditional random field (CRF) is a probabilistic graphical model, it models a probability distribution of pixel labels, conditioned on global observation. Consider a random field $X = \{X_1, X_2, \dots, X_N\}$ defined as the random variables to be inferred from RGB image Y . Every random variable X_i takes a label from $\mathcal{L} = \{l_1, l_2, \dots, l_k\}$, k is the semantic label. Any possible assignment of all random variables is called labeling which can take values from $\mathcal{L}$.

The general CRF-based labeling model is defined over an undirect graph $G = (V, \xi)$, where V contains all pixels, $V = X_1, X_2, \dots, X_N$, N is the size of RGB image, ξ defines connectivity between random variables. For each pixel, the neighborhood system is usually adopted 4 or 8 connected. The general energy function as follows:

$$P(X = x|Y) = \frac{1}{Z(Y)} \exp(-E(X|Y)),$$

$$\text{and } Z(Y) = \sum_x \exp(-E(X|Y)) \quad (5)$$

where Z is partition function. The Gibbs energy function could be written as:

$$E(X|Y) = \sum_{i \in V} \psi_u(x_i) + \sum_{i,j \in \xi} \psi_p(x_i, x_j) \quad (6)$$

For notational convenience, it is wise choice to omit the conditioning on Y , and $\psi_u(\cdot)$ is unary potential, cost of assigning label x_i to pixel i . $\psi_p(\cdot)$ is pairwise potential, cost of assigning labels (x_i and x_j) to pixels (i and j), only penalizes the different labels.

B. The R-CRF Model

The traditional CRF generally considers the result of RGB image as unary term, and it only requires to connect local 4 or 8 neighbors in pairwise potential. And the graph inference is based on the Graph-cut. This approach leads to inefficient optimization of the CRF model locally and cannot capture global features. Therefore, the proposed R-CRF fully makes use of results (information) of two sensors. Each modality is independently processed with its own semantic segmentation network. It takes probability scores generated

by two modalities of data as the unary potential term, both the RGB image and the densified LiDAR-images are utilized as pairwise potential terms to encode the contextual consistency. Then the energy function is iteratively optimized by mean-field variational inference so that we can refine the labeling results through a fully-connected graph that makes use of the local and globe interaction of the RGB image and LiDAR image. R-CRF model can be formulated by minimizing the energy function defined as follows: x denotes labels assigned to pixels.

$$\min_x E(X = x) = \sum_{i \in V} \psi_u(x_i) + \sum_{\substack{i < j \\ i, j \in V}} \psi_p(x_i, x_j) \quad (7)$$

1) **Unary potential:** $\psi_u(x_i)$ can be regarded as the prior distribution. It takes negative log-likelihood of variable X predicted by the outputs of segmentation network.

$$\psi_u(x_i) = \psi_u^l(x_i) + \psi_u^r(x_i) \quad (8)$$

$$\psi_u^l(x_i) = -\log(H_p^l(x_i)), \quad \psi_u^r(x_i) = -\lambda \log(H_p^r(x_i)) \quad (9)$$

The $\psi_u^l(x_i)$ and $\psi_u^r(x_i)$ represent the potentials of the point cloud and image data separately. $H_p^l(x_i)$ and $H_p^r(x_i)$ are the results of each modality segmentation network. The λ is utilized to balance the tradeoff between terms in (9). In order to fuse equally, λ is set to 1 in the experiment.

2) **Pairwise potential:** $\psi_p(x_i, x_j)$ consists of a weight sum of Gaussian functions, it is only related to the difference between pixels i and j . It encourages neighboring pixels to have the same labels, also having a smoothing effect on the labeling result.

$$\psi_p(x_i, x_j) = u(x_i, x_j) \sum_{m=1}^M w^m k^{(m)}(f_i, f_j)$$

$$=$$

$$u(x_i, x_j) g(f_i, f_j) \quad (10)$$

where $k^{(m)}$ for $m=1 \dots, M$, the Gaussian kernel applied on feature vectors f and w^m is corresponding coefficient.

Label compatibility function $u(x_i, x_j) = 1$ if $x_i \neq x_j$ and 0 otherwise, it catches compatibility between different label pairs. Traditional methods only utilize features extracted from RGB modality, whereas in this paper, several Gaussian kernels take point cloud along with RGB image into consideration.

The first Gaussian kernel in formula (11) is observed by RGB image, former is called gaussian *appearance kernel*, it boosts nearby pixels with similar color may belong to same class. The latter is Gaussian *spatial kernel* called smoothing kernel, it removes minor obscure regions as previous models do.

$$g^{(1)}(f_i, f_j) = w^{(1)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\right)$$

$$+ w^{(2)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right) \quad (11)$$

where p_i and p_j are the positions in image coordinate, I_i and I_j are color values. $\theta_\alpha, \theta_\beta, \theta_\gamma$ are parameters of kernels.

The second and third Gaussian kernels are observed by point cloud, the height and depth maps are obtained from the aligned point cloud. The second Gaussian kernel is a *height bilateral kernel*:

$$g^{(2)}(f_i, f_j) = w^{(3)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\epsilon^2} - \frac{|H_i - H_j|^2}{2\theta_\eta^2}\right) \quad (12)$$

where H_i and H_j are height values. The third is *distance bilateral kernel*, it assumes nearby pixels with close distance are likely to be the same semantic:

$$g^{(3)}(f_i, f_j) = w^{(4)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\epsilon^2} - \frac{|D_i - D_j|^2}{2\theta_\omega^2}\right) \quad (13)$$

where D_i and D_j are the value of distance in the LiDAR coordinate.

C. Mean Field Iteration in the Recurrent-CRF model

Minimizing the above function (6) yields the most likely label assignment for given data. Extract minimization of equation is intractable, then mean-field variable inference algorithm [29] is proposed to solve the fully connected graph approximately and efficiently. Aspired by the work of ConvCrf [32], we bring the conditional independence assumption to the full-connected CRF, and message passing step is reformulated to a convolution with truncated Gaussian kernel. Following [31-32], we approximate the Gibbs distribution $P(X)$ with mean-field distribution $Q(X)$ to minimize KL divergence between $P(X)$ and $Q(X)$. The form of $Q(X)$ is as follows:

$$Q(X) = \prod_{i=1}^N Q_i(X_i) \quad (14)$$

Mean-field variational inference is usually implemented by continuously updating the distribution $Q(X)$ iteratively, and finally the optimal solution is obtained, which is expressed as follow:

$$\begin{aligned} KL(Q(X)||P(X|Y)) &= \\ \sum_X Q(X) \log \frac{Q(X)}{P(X|Y)} &= \sum_X Q(X) \log \frac{Q(X)Z(Y)}{\exp(-E(X, Y))} \\ &= \\ \sum_X Q(X)E(X, Y) + \sum_X Q(X) \log Q(X) + \log Z(Y) \end{aligned} \quad (15)$$

Iterative update equation is as follows:

$$\begin{aligned} Q_i(x_i = l) &= \frac{1}{Z_i} \\ \exp\left\{-\psi_u(x_i) - \sum_{l' \in L} u(l, l') \sum_{m=1}^k w^{(m)} \sum_{j \neq i}^k k^{(m)}(f_i, f_j) Q_j(l')\right\} \end{aligned} \quad (16)$$

A brief description of how to break update equation down into simpler steps in algorithm 1 is provided. It is composed of six steps:

Algorithm 1. Mean-field Iteration in the Recurrent-CRF Model

1. Initialization:

$$Q_i(l) \leftarrow \frac{1}{\sum_l \exp(U_i(l))} \exp(U_i(l)), i = 1, 2, \dots, N$$

While not converged **do**

2. Message Passing

$$\tilde{Q}_i(l) \leftarrow \sum_{j \neq i}^k k^{(m)}(f_i, f_j) Q_j(l)$$

3. Weighting Filter Outputs

$$\hat{Q}_i(l) \leftarrow \sum_m \omega^{(m)} \tilde{Q}_i(l)$$

4. Compatibility Transform

$$\check{Q}_i(l) \leftarrow \sum_{l' \in L} u(l, l') \hat{Q}_i(l')$$

5. Adding Unary Potentials

$$\bar{Q}_i(l) \leftarrow U_i(l) - \check{Q}_i(l)$$

6. Normalizing

$$Q_i \leftarrow \frac{1}{\sum_l \exp(U_i(l))} \exp(\bar{Q}_i(l))$$

end while

Step1: Initialization. The probability scores obtained from segmentation algorithms are utilized to initialize.

Step2: Message-passing. The message passing step involves filtering the approximated marginal. Gaussian kernels based on the images are processed to obtain differences in function (11) to (13). As the distance between two pixels increases, the value of Gaussian kernels mentioned above decreases very fast. Therefore, assuming that label distribution of two data i, j are conditionally independent, for all pixels whose Manhattan distance $d(i, j) > k$, k is hyperparameter, then pairwise potential is zero, it reduces the complexity of the pairwise potential greatly. It reflects the correlation between a pixel and others.

Step3: Weighting Filter Outputs. We apply Gaussian kernels to filter the probability map in step 2, this step can be seen as 1×1 convolution.

Step4: Compatibility Transform. This step is utilized to determine the extent to which it changes the distribution. This step can be seen as convolution with 1×1 kernel.

Step5: Adding Unary potentials. We update it by adding the unary potential received from step 1 to result of step 4.

Step6: Normalization. SoftMax is used to normalize it.

The output of this module is a refined probability map that can be further refined by iterative applications.

Generally, one iteration can be modeled as a bunch of ordinary CNN layers, by processing multiple iterations, the output of one time iteration becomes the input for the next iteration, as illustrated in Fig.5.

Fig.5 The update equation for mean-field inference of [36] is decomposed into smaller steps, one time iteration can be seen as neural network.

4. EXPERIMENTS

4.1 Dataset and metrics

R-CRF is evaluated on broadly utilized KITTI ROAD benchmark [27]. ROAD dataset includes corresponding calibration parameters, ground-truth images, RGB images, point cloud, and scripts for evaluation. It consists of 289

labeled frames for training set, and 290 frames for testing set. Terminal results are evaluated on KITTI's online server. For road detection, the KITTI dataset presents four scenarios: Urban Unmarked road (UU), Urban Marked road (UM), and Urban Multiple Marked-lanes (UMM) and all three urban subsets (URBAN). In addition, a category called URBAN is calculated, which supplies an overall score. In this case, only road area is taken into consideration, the lane detection task is ignored.

Following the benchmark evaluation, the KITTI-ROAD dataset provides the maximum F-measure at the pixel-level in Bird's-eye view (BEV) space. Pprincipal metric matrix values are used to evaluate the accuracy, including Max F (Maximum F1), PRE (precision), REC (recall), AP (average precision), FPR (false positive rate) and FNR (false negative rate). The definition of the matrix:

$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN} \quad (17)$$

$$AP = \frac{TP + TN}{TP + FP + TN + FN} \quad (18)$$

$$FPR = \frac{FP}{FP + TN}, FNR = \frac{FN}{TP + FN} \quad (19)$$

$$\text{MaxF} = \max\left(\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right) = \frac{2TP}{2TP + FP + FN} \quad (20)$$

where TP , FP , TN , and FN represent the number of samples. Precision and Recall bring different insights into method's performance: low precision implies that many background pixels are sorted as roads, while low recall indicates the road surfaces are not detected. The KITTI benchmark ranks all methods according to maxF .

4.2 Implementation Details

About the implementation details, a modified Deeplabv3+ is used for 2D segmentation network, also PointNet for 3D segmentation network. For a more focused view, the input of RGB camera is resized into a size of 1242×375 , and learning rate for image training is set to 0.001. Input of LiDAR point clouds is rectified, for one image, approximately 20,000 points are used in camera field of view, the learning rate is set to 0.001 for point cloud segmentation. Furthermore, the value of epoch and batch size are set to 400 and 4. The parameters of pairwise potential are learnable

parameters of the model. The proposed framework is implemented with PyTorch and tensorflow on Ubuntu 18.04 operating system, and the environment is carried out in an NVIDIA 1080 TI GPU.

4.3 Experimental Results

A. Ablation Study

We compare the results obtained from image-only, point cloud-only and the proposed fusion-based. The experiments are conducted on validation. Results are illustrated in Table 1, image-only means only image-segmentation algorithm is employed, and point cloud-only means only the point cloud segmentation algorithm is employed. The last one is the whole framework described in this paper, with the input of two multi-modality data. The fusion model achieves best performance with a MaxF score of 94.64%, it improves by 2.76% over image-based, and 1.26% improvement over point cloud-based results. Fusion model has a significant improvement through a combination of geometric properties and color information.

TABLE 1. Comparison results under different modalities on validation dataset (/%)

Modality	MaxF	AP	PRE	REC
Image-only	91.88	88.12	92.45	91.21
Point cloud-only	93.38	92.71	95.06	92.73
Fusion	94.64	93.23	94.32	94.22

In addition, we fetch some examples of the road segmentation results on validation dataset in Fig.6, respectively. Each line presents an image from UM, UMM and UU dataset separately. Obviously, in the case of image-only, when there are many shadows on the road (second row in Fig.6), it will be unmanageable to observe road accurately. Point cloud modality, on the other hand, is less affected by illumination, hence it gives better output than RGB image modality, as illustrated in the second row of Fig.6. However, in the case of point cloud-only, it is not easy to detect road accurately if height difference from roadside is small. Last row in Fig.6, some misclassified regions in both separate modalities are corrected after fusion, the performance can be enhanced with multi-modality data fusion. Fusion method can effectively aggravate complementary features from image and point cloud to achieve performance improvement for the single modal.

Fig. 6. Visual results. The first row to the last row is the original images, results of the image-based, point cloud-based, and proposed fusion results, respectively.

B. Evaluation on KITTI Benchmark Test Dataset.

As the KITTI road benchmark evaluates bird's-eye view results, results in perspective are mapped to a 400×800 bird's eye view. Mapped images represent the accessibility

of the region of 40 meters ahead (from 6 to 46 meters) and 10 meters on each side (or so), then submitted to website for evaluation. Fig.7 shows some evaluation results.

(a) Visual results of the perspective view version.

(b) Visual results of the bird view version.

Fig. 7 Visual results on KITTI-ROAD test dataset.

Fig. 7 illustrates some instances of road results by proposed method, in perspective view of image in Fig.7 (a) and a bird's eye view in Fig. 7 (b). Each row of Fig. 7 (a) matches the row in Fig.7 (b). Before evaluation, results of perspective version will be converted to bird view version. Since pixel resolution of perspective decreases with distance from camera, distant pixels are more important when converting to the bird view. As can be seen in some areas, the edges of road and shadows located by cars are slightly uneventful. Note that red indicates incorrect drivable region (false negatives), blue areas is the missing drivable region (false positives), and green represents correct drivable region (true positives). It demonstrates proposed method has comparable performance.

TABLE 2. Performance of the proposed model on KITTI (BEV) (%)

Dataset	MaxF	AP	PRE	REC	FPR	FNR
UM_ROAD	94.54	93.23	94.57	94.52	2.47	5.48

UMM_ROAD	95.41	95.39	95.42	95.41	5.04	4.59
UU_ROAD	92.00	92.19	92.01	91.98	2.60	8.02
URBAN_ROAD	94.27	93.63	94.22	94.32	3.19	5.68

Table 2 shows the statistical test results of 4 scenarios, obtained directly from the evaluation server, on UMM_ROAD dataset, main indicator MaxF reaches 95.41% and average MaxF on the entire test set attains 94.27%. The UU scenario has the lowest performance compared to other scenarios because of its multiple complex environment, and is more irregular than that both datasets. Further, Fig. 8 shows Precision-Recall results on testing set for each urban scenario.

Fig. 8. Precision-Recall curves from evaluation server.

C. Comparison with Other Fusion-based Methods

To prove effectiveness, we compare this method with several high-ranking methods leveraging multi-modality data on KITTI testing dataset, including deep learning-based algorithms (PLARD [13], LidCamNet [14]), feature-based algorithm (RES3D-Velo [16]), and CRF fusion-based methods (FusedCRF[21], HybridCRF[22] and MixedCRF [24][24]). The statistical performance comparison results on UM, UMM, UU subsets, and average results on all sets (URBAN_ROAD) are illustrated in Tables 3-6. Note that the Red, Blue and Green fonts represent the First, the Second, and the Third place in corresponding column.

TABLE 3. Comparison with Several Popular Fusion-based Methods on UM_ROAD (BEV) (%)

Method	MaxF	AP	PRE	REC	FPR	FNR
PLARD	97.05	93.53	97.18	96.92	1.28	3.08
LidCamNet	95.62	93.54	95.77	95.48	1.92	4.52
RES3D-Velo	83.81	73.95	78.56	89.80	11.16	10.20
FusedCRF	89.55	80.00	84.87	94.78	7.70	5.22
HybridCRF	90.99	85.26	90.65	91.33	4.29	8.67
MixedCRF	91.57	84.68	90.02	93.19	4.71	6.81
Ours	94.54	93.23	94.57	94.52	2.47	5.48

TABLE 4. Comparison with Several Popular Fusion-based Methods on UMM_ROAD (BEV) (%)

Method	MaxF	AP	PRE	REC	FPR	FNR
PLARD	97.77	95.64	97.75	97.79	2.48	2.21
LidCamNet	97.08	95.51	97.28	96.88	2.98	3.12
RES3D-Velo	90.60	85.38	85.96	95.78	17.20	4.22
FusedCRF	89.51	83.53	86.64	92.58	15.69	7.42
HybridCRF	91.95	86.44	94.01	89.98	6.30	10.02
MixedCRF	92.75	90.24	94.03	91.50	6.39	8.50
Ours	95.41	95.39	95.42	95.41	5.04	4.59

TABLE 5. Comparison with Several Popular Fusion-based Methods on UU_ROAD (BEV) (%)

Method	MaxF	AP	PRE	REC	FPR	FNR
PLARD	95.95	95.25	96.25	95.65	1.21	4.35
LidCamNet	94.54	92.74	94.64	94.45	1.74	5.55
RES3D-Velo	83.63	72.58	77.38	90.97	8.67	9.03
FusedCRF	84.49	72.35	77.13	93.40	9.02	6.60
HybridCRF	88.53	80.79	86.41	90.76	4.65	9.24
MixedCRF	85.69	75.12	80.17	92.02	7.42	7.98
Ours	92.00	92.19	92.01	91.98	2.60	8.02

TABLE 6. Comparison with Several Popular Fusion-based Methods on URBAN_ROAD (BEV) (%)

Method	MaxF	AP	PRE	REC	FPR	FNR
PLARD	97.03	94.03	97.19	96.88	1.54	3.12
LidCamNet	96.03	93.93	96.23	95.83	2.07	4.17
RES3D-Velo	86.58	78.34	82.63	90.92	10.53	9.08
FusedCRF	88.25	79.24	83.62	93.44	10.08	6.56
HybridCRF	90.81	86.01	91.05	90.57	4.90	9.43
MixedCRF	90.59	84.24	89.11	92.13	6.20	7.87
Ours	94.27	93.63	94.22	94.32	3.19	5.68

As table 3-5 illustrate, our method acquires good results in four scenarios, which demonstrates for different situations, this method is not only accurate but also robust. Further, it is obvious that the method is competitive (Third place). Particularly, compared with deep learning-based approaches (PLARD [13] and LidCamNet [14]), the PLARD [13] performs best, LidCamNet [14] is in the second place, and the MaxF values of our method get the third place. The reason behind its slightly lower performance than PLARD [13] and LidCamNet [14] is that the height map features are fused multiple times in the deep learning network at middle level fusion.

Compared with these hand-crafted CRF fusion approaches, our approach excels by all criteria, and it performs best on the main index MAF in general, achieving 6.02%, 3.46%, 3.68% improvement over FusedCRF [21], HybridCRF [22], and MixedCRF [24] on the URBAN_ROAD dataset, respectively. In general, our method has certain advantages: it can not only add the results (information) in unary potential, but also integrate the RGB image, the densified height and depth images in pairwise potentials which could increase the data density; the energy function is iteratively optimized by mean-field variational inference; followed by such a probabilistic fusion process, the proposed model possesses a big error correction capability. All results are calculated on the KITTI's online evaluation server, results from other studies are based on the values accounted on KITTI's website.

Some distinctive results of the urban scenarios are also illustrated in Fig. 9. In Fig. 9, first column is RGB image, then starting from second column, road detection results from methods mentioned in table 6 are displayed. This model performs better than hand-crafted CRF fusion-based methods in complex scenes.

Fig. 9 Visual results comparison on KITTI-ROAD testing set.

5. CONCLUSION

This paper proposes a camera-LiDAR sensor fusion method for road detection. It employs a novel R-CRF model to combine the results generated from the two sensors as the unary term. Densified LiDAR-image and RGB image are treated as pairwise terms in which edges are fully connected. The road detection is formulated as a two-class pixel labeling problem, and iteratively optimized by mean-field variational inference. After fusion process, the proposed model can possess great error correction ability. Extensive experiments are carried out on the KITTI dataset, and result demonstrates that it performs better than single-modality-based methods. Compared to existing up-to-date models, our method is competitive in detection accuracy.

Author Contributions: Conceptualization, L.W. and Y.H.; methodology, L.W. and Y.H.; software, L.W.; validation, L.W. and Y.H.; formal analysis, L.W. and Y.H.; investigation, L.W.; writing—original draft preparation, L.W. and Y.H.; writing—review and editing, L.W., Y.H.; project administration, Y.H.; funding acquisition, Y.H. All authors have read and agreed to the published version of the manuscript.

Data availability:

The datasets generated during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflict of interest

Reference

- [1]. Teichmann, M.; Weber, M.; Zöllner, M.; Cipolla, R.; Urtasun, R. MultiNet: Real-time joint semantic reasoning for autonomous driving. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 26–30 June 2018; pp. 1013–1020.
- [2]. Z. Chen, Z. Chen, RBNet: a deep neural network for unified road and road boundary detection, in: Proceedings of the International Conference on Neural Information Processing, 2017, pp. 677–687, doi:10.1007/978-3-319-70087-870.
- [3]. Fan, R.; Wang, H.; Cai, P.; Wu, J.; Bocus, J.; Qiao, L.; Liu, M. Learning Collision-Free Space Detection from Stereo Images: Homography Matrix Brings Better Data Augmentation. IEEE ASME Trans. Mechatron 2021.
- [4]. R. Fernandes, C. Premebida, P. Peixoto, and D. Wolf, “Road detection using high resolution LiDAR,” in IEEE Vehicle Power and Propulsion Conference, 2014, pp. 1–6. Xu, F., Chen, L., Lou, J., Ren, M., & Guo, Y. (2019). A real-time road detection method based on reorganized lidar data. PLoS ONE, 14(4).
- [5]. L. Wang and Y. Huang, “A Survey of 3D Point Cloud and Deep Learning-Based Approaches for Scene Understanding in Autonomous Driving,” in IEEE Intelligent Transportation Systems Magazine, doi: 10.1109/MITS.2021.3109041.
- [6]. Df A., As A., Rn A., et al. Point-Cloud based 3D Object Detection and Classification Methods for Self-Driving Applications: A Survey and Taxonomy[J]. Information Fusion, 2020.
- [7]. Caltagirone, L.; Scheidegger, S.; Svensson, L.; Wahde, M. Fast LIDAR-based road detection using fully convolutional neural networks. In Proceedings of the 2017 IEEE Intelligent Vehicles Symposium (IV), Los Angeles, CA, USA, 11–14 June 2017; pp. 1019–1024.
- [8]. Lyu, Y.; Bai, L.; Huang, X. ChipNet: Real-Time LiDAR Processing for Drivable Region Segmentation on an FPGA. IEEE Trans. Circuits Syst. I Regul. Pap. 2019, 66, 1769–1779.
- [9]. S. Gu, Y. Zhang, X. Yuan, et al., Histograms of the normalized inverse depth and line scanning for urban road detection, IEEE Trans. Intell. Transp. Syst. 20 (8) (2019) 3070–3080, doi: 10.1109/TITS.2018.2871945.
- [10]. F. Wulff, B. Schauffele, O. Sawade, D. Becker, B. Henke, and I. Radusch, “Early fusion of camera and Lidar for robust road detection based on U-Net FCN,” in Proc. IEEE Intell. Vehicles Symp. (IV), Jun. 2018, pp. 1426–1431.
- [11]. Yu, B.; Lee, D.; Lee, J.-S.; Kee, S.-C. Free Space Detection Using Camera-LiDAR Fusion in a Bird’s Eye View Plane. Sensors 2021, 21, 7623. Lee, J.-S.
- [12]. Park, T.-H. Fast Road Detection by CNN-Based Camera-Lidar Fusion and Spherical Coordinate Transformation. IEEE Trans. Intell. Transp. Syst. 2021, 22, 5802–5810.
- [13]. Chen, Z.; Zhang, J.; Tao, D. Progressive LiDAR adaptation for road detection. IEEE CAA J. Autom. Sin. 2019, 6, 693–702.
- [14]. Caltagirone, L.; Bellone, M.; Svensson, L.; Wahde, M. Lidar-camera fusion for road detection using fully convolutional neural networks. Robot. Auton. Syst. 2019, 111, 125–131.
- [15]. C. Premebida, J. Carreira, J. Batista, and U. Nunes, “Pedestrian detection combining rgb and dense LiDAR data,” in Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on. IEEE, 2014, pp. 4112–4117.
- [16]. P. Y. Shinzato, D. F. Wolf, and C. Stiller, “Road terrain detection: Avoiding common obstacle detection assumptions using sensor fusion,” in 2014 IEEE Intelligent Vehicles Symposium Proceedings. IEEE, 2014, pp. 687–692.
- [17]. Xiao Hu, Rodriguez F. Gepperth A. A Multi-Modal System for Road Detection and Segmentation [C]// Intelligent Vehicles Symposium. IEEE, 2014.1365-1370
- [18]. Ji-Heon P., et al. “Drivable Dirt Road Region Identification Using Image and Point Cloud Semantic Segmentation Fusion.” IEEE Transactions on Intelligent Transportation Systems (2021).
- [19]. Arnab, A., et al. “Conditional Random Fields Meet Deep Neural Networks for Semantic Segmentation: Combining Probabilistic Graphical Models with Deep Learning for Structured Prediction.” IEEE Signal Processing Magazine 35.1(2018):37-52.
- [20]. R. Zhang, S.A. Candra, K. Vetter, A. Zakhor, Sensor fusion for semantic segmentation of urban scenes, in: Robotics and Automation (ICRA), 2015 IEEE International Conference on, IEEE, 2015, pp. 1850–1857.
- [21]. L. Xiao, B. Dai, D. Liu, T. Hu, and T. Wu, “Crf based road detection

- with multi-sensor fusion,” in 2015 IEEE Intelligent Vehicles Symposium (IV), June 2015, pp. 192–198.
- [22]. L. Xiao, R. Wang, B. Dai, Y. Fang, D. Liu, and T. Wu, “Hybrid conditional random field based camera-LiDAR fusion for road detection,” *Inf. Sci.* 432, 543–558 (2018).
- [23]. S. Gu, T. Lu, Y. Zhang, J. Alvarez, J. Yang, and H. Kong, “3D LiDAR + monocular camera: an inverse-depth induced fusion framework for urban road detection,” *IEEE Transactions on Intelligent Vehicles*, vol. PP, no. 99, pp. 1–1, 2018.
- [24]. Han X, Wang H, Lu J, et al. Road detection based on the fusion of LiDAR and image data[J]. *International Journal of Advanced Robotic Systems*, 2017, 14(6): 1729881417738102.
- [25]. S. Gu et al., “Road detection through CRF based LiDAR-camera fusion,” in *Int. Conf. Rob. Autom. (ICRA)*, IEEE, pp. 3832–3838 (2019)
- [26]. Gu, S.; Zhang, Y.; Tang, J.; Yang, J.; Alvarez, J.M.; Kong, H. Integrating Dense LiDAR-Camera Road Detection Maps by a Multi-Modal CRF Model. *IEEE Trans. Veh. Technol.* 2019, 68, 11635–11645..
- [27]. Fritsch, J.; Kühnl, T.; Geiger, A. A new performance measure and evaluation benchmark for road detection algorithms. In *Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, Auckland, New Zealand, 27–30 October 2019; pp. 1693–1700.
- [28]. R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation,” in *Proc. CVPR*, Dec. 2017, pp. 77–85.
- [29]. L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *European Conference on Computer Vision (ECCV)* (Munich, 2018), pp. 833–851.
- [30]. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* 2018, 40, 834–848.
- [31]. P. Krahenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, pages 109–117, 2012.
- [32]. Teichmann, M., and R. Cipolla. “Convolutional CRFs for Semantic Segmentation.” (2018).
- [33]. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected CRFs,” in *Proc. Int. Conf. Learning Representations*, 2015.
- [34]. S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr, “Conditional random fields as recurrent neural networks,” in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 1529–1537
- [35]. B. Wu, A. Wan, X. Yue, and K. Keutzer, “SqueezeSeg: Convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3D LiDAR point cloud,” in *Proc. IEEE Int. Conf. Robot. Autom.*, 2018, pp. 1887–1893. doi: 10.1109/ICRA.2018.8462926
- [36]. Shotton, J. Winn, C. Rother, and A. Criminisi, “Textonboost for image understanding: Multiclass object recognition and segmentation by jointly modeling texture, layout, and context,” *Int. J. Comput. Vision*, vol. 81, no. 1, pp. 2–23, 2009.

Acknowledgment

This work was sponsored by the Shanghai Nature Science Foundation of Shanghai Science and Technology Commission, China (Grant NO. 20ZR14379007), and National Nature Science Foundation of China (Grant NO. 61374197).