Depth Meets CNN: A Fusion Based Approach for Semantic Road Segmentation

Deepak Singh [*1] **Abhinav Atrishi** [*1] **Sarthak Gupta** [*1] **Raghav Marwaha** [*1]

Abstract

This work presents a stereo vision based road segmentation algorithm. To tackle the problem of uneven textures faced by traditional 2D based approaches and poor point cloud resolution in purely 3D methods, a 2D-3D fusion based approach has been proposed in this paper. 3D cues from v disparity map are fused with 2D cues obtained from a standard CNN model using conditional random fields, thus improving accuracy on varying road textures.

1. Introduction

Precise road detection in cluttered environments is currently one of the most active research topics. Complexity level of the problem increases many-folds when going from highways to urban environment. This happens due to uneven texture of roads, presence of potholes, varying illumination and shadows on the roads.

The methods for road detection are divided into three classes of 3D, 2D and hybrid methods. 3D methods are dependent on scene geometry for road detection and make use of sensors like LIDAR, stereo camera etc (Fernandes et al., 2014). These methods are independent of textures and lighting conditions but suffer from poor resolution to differentiate between road and pavement coordinates. On the other hand, 2D based approaches make use of low-level cues like colors or a combination of colors and textures (Kim et al., 2011). Use of deep learning techniques in this domain has proven to be very successful (Badrinarayanan et al., 2017; Long et al., 2015). 2D based approaches are prone to errors in case of varying road textures and environment illumination.

Hybrid techniques make use of 3D and the 2D cues. These methods usually project the 3D points onto the 2D image and fuse it with the result obtained from the purely 2D method (Caltagirone et al., 2019; Patra et al., 2018). Such techniques leverage the capabilities of both the approaches.

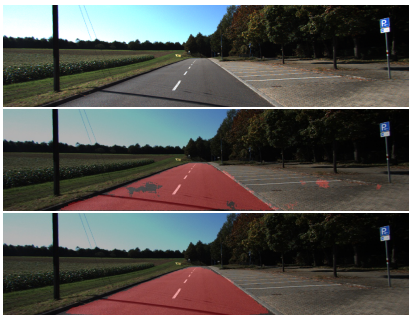


Figure 1. Road detection using proposed method. (Top to Bottom) Input image; Road mask from ResNet; Improved road mask using the proposed algorithm

This paper presents a fusion based algorithm to take advantage of complementary strengths of 2D and 3D methods using stereo vision. The proposed method initially uses v disparity (Labayrade et al., 2002) to estimate the ground plane from the input stereo images. To effectively filter out inconsistencies in the v disparity map and detect the line corresponding to the ground plane, a shallow convolutional neural network is used in the prior. This helps generalize the method better by eliminating extra lines from the v disparity space without putting constraints on the line parameters like angle or slope. The probabilities of the road are then estimated from the ground plane detected using longitudinal and lateral road profiling and depth accuracy constraints. Finally a fusion is performed between these probabilities and the probabilities given by any state of the art road detection network like ResNet (He et al., 2016) or VGG-16 (Simonyan & Zisserman, 2014) using a fully connected Conditional Random Fields (CRF) framework (Krähenbühl & Koltun, 2011). The proposed algorithm is explained step wise in the following section, the flowchart is represented in Fig. 2 and Fig. 1 demonstrates the improvement in road detection using proposed method.

2. Proposed Algorithm

2.1. Depth Prior Generation

The depth prior for the road is generated using v disparity algorithm proposed in (Labayrade et al., 2002). The dispar-

[*]Equal contribution [1]Division of Electronics & Communication Engineering, Netaji Subhas Institute of Technology, New Delhi, India. Correspondence to: Deepak Singh <deepaks.ec.16@nsit.net.in>.

Workshop on AI for Autonomous Driving (AIAD), the 37^{th} International Conference on Machine Learning, Vienna, Austria, 2020.

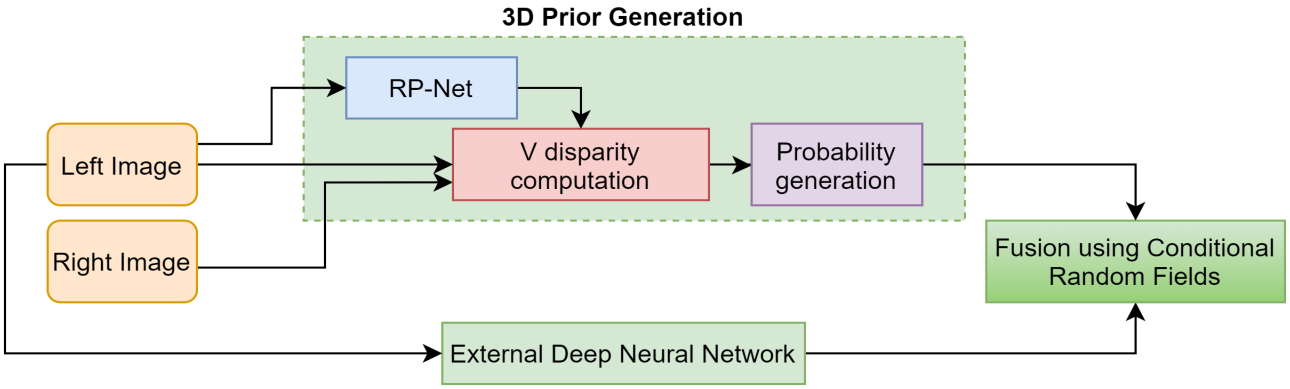


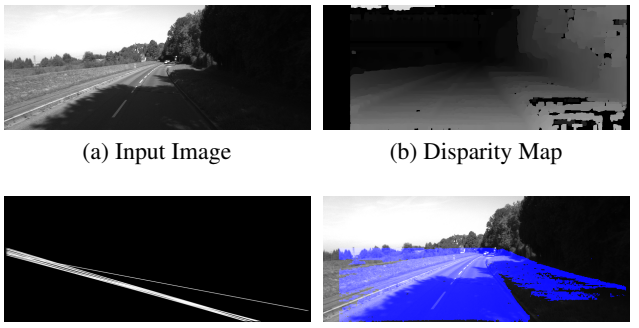
Figure 2. Flowchart of the proposed algorithm

ity map is computed using Semi Global Block Matching Technique (SGBM) (Hirschmuller, 2008) from which the v disparity map is generated by accumulating same disparity pixels in each row of the disparity map. The v disparity algorithm generates a straight line corresponding to the ground plane in euclidean space (Labayrade et al., 2002). Fig. 3 represents the road mask estimation using v disparity algorithm. Here Fig. 3a is the input image and the corresponding disparity is shown in Fig. 3b. Fig. 3c shows the final line in v disparity space after applying Hough transform which generates the road mask depicted in Fig. 3d. The stereo setup geometry is shown in appendix A.

Line detection or fitting in v disparity space is very important as poor line detection can lead to obstacles appearing in the road mask as evident from Fig. 3d. Direct application of Hough transform on the v disparity map can generate multiple lines and hence geometrical constraints depending on camera alignment with respect to the road (pitch angle and height) are required to detect the optimum line. (Zhao et al., 2007) tackled the problem of varying camera pitch angles but assumed the height of the camera to be constant. In order to make line fitting or detection more adaptive, a shallow CNN road detection model is used as a prior along with v disparity in this paper. The pixels marked as road by this network are detected in the disparity map and the corresponding v disparity map. These pixels give a hint for the optimum line and hence there is no dependency on the camera alignment with respect to the road. Finally, Hough transform is used on these pixels to get the final line in v disparity space.

2.2. RPNet - Road Prior Network

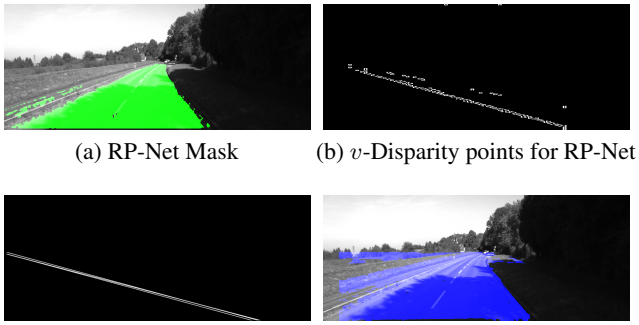
The proposed network is an encoder-decoder network. Inspired from ResNet (He et al., 2016), a skip connection with element-wise addition is utilized between layers 1 & 3 to achieve better accuracy. A custom shallow CNN was chosen to keep the prior independent from the external models and to keep the computational cost low.



(a) Input Image (b) Disparity Map

(c) Line in V-Disparity space (d) Road Mask on Original Image

Figure 3. v-Disparity based Road Detection



(a) RP-Net Mask (b) v-Disparity points for RP-Net

(c) Line: RP-Net & v-Disparity (d) Road Mask on Original Image

Figure 4. v-Disparity & RP Net based Road Detection

The architecture of the proposed network is presented in Fig. 5. The network is essentially composed of three blocks, each block has a batch normalization (Ioffe & Szegedy, 2015) layer to overcome the internal covariate shift problem followed by a Conv2D layer with (3×3) kernel, finally PReLU (He et al., 2015) is used as the activation function. As shown in (Maas et al., 2013) Leaky ReLU converges faster than ReLU, PReLU extends upon LReLU by learning the negative slope of non-linearities. Due to this PReLU accelerates the learning of the network, while also giving higher accuracy's as shown in (He et al., 2015). ELU (Clev-

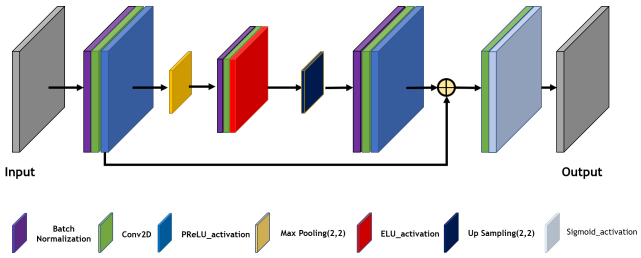


Figure 5. Architecture of the proposed RP-Net

ert et al., 2015) activation is chosen for the second layer as it helps in reducing the number of parameters and reducing the computational complexity of the network, no significant loss in the accuracy was observed by changing the second layer from PReLU to ELU. The final layer of the network is a (1×1) convolution with n-channels, where n is the number of classes to be predicted ($n = 2$, in this case). To get the output probabilities of the network sigmoid activation function is used (softmax for $n > 2$) which are then further used in the CRF Fusion as described in the section 2.5 below.

RP-Net was trained on 289 images of the KITTI Road Dataset (Fritsch et al., 2013). The network was designed to provide a hint of road to the V-Disparity prior and its performance is shown in Fig. 4. Here, Fig. 4a shows the mask as predicted by RP-Net and Fig. 4b shows the corresponding points detected by RP-Net in the v disparity space. Finally, Fig. 4c represents the line detected applying Hough transform on the points shown in Fig. 4b and the corresponding road mask obtained using this is depicted in Fig. 4d.

2.3. Probability generation for the prior

In order to generate probabilities for the road mask obtained from the depth prior, road profiling has been done along rows and columns thereby generating row and column probabilities. In a stereo vision system, the accuracy of depth estimation decreases quadratically with distance from the camera (Gallup et al., 2008). Therefore, the row probability distribution is obtained using a row decay model which assigns higher probability of accuracy to road pixels having higher row value (closer to the camera). For column probabilities, firstly mean of each row in the prior road mask is calculated. In each row, higher probability of accuracy is assigned to pixels that lie closer to the calculated mean. Detailed probability calculations for an image are attached as supplementary material and can be found in appendix B.

2.4. Road Mask from External DNN Model

The above proposed prior when fused with any state of the art deep neural network, improves the result of the same. To demonstrate this capability, a U-Net (Ronneberger et al., 2015) was created, which is a state of the art algorithm for image segmentation. Further two popular architectures namely ResNet-18 (He et al., 2016) and VGG-16 (Simonyan & Zisserman, 2014) are chosen. Two U-Net model's were then created using ResNet-18 & VGG-16 as encoders. Transfer learning is then utilized to speed up the process wherein ResNet & VGG Net's initial weights are taken from the pre-trained model on ImageNet (Deng et al., 2009). The two models namely UNet using ResNet encoders (URNET) and UNet using VGG Net encoders (UVGG) are then trained on 289 training images for just 5 epochs on the KITTI Road Dataset (Fritsch et al., 2013). The network was purposely trained using few epochs to better demonstrate the ability of the proposed algorithm to improve the results even when using lightly trained networks. The predictions from these two networks are then fused with the proposed prior using conditional random fields fusion.

2.5. Conditional Random Fields Fusion

A fully connected CRF framework has been used to perform fusion between the road masks obtained from external neural network and from the depth based prior.

For an image of size $H \times W$ where H and W are height (number of rows) and width (number of columns), the energy function $E(x)$ where x denotes the label vector is given as

$$E(x) = \sum_i \phi_i(x_i) + \sum_{i,j} \phi_{i,j}(x_i, x_j) \qquad (1)$$

where $\phi_i(x_i)$ is the unary potential for i^{th} pixel and $\phi_{i,j}(x_i, x_j)$ is the pairwise potential between the i^{th} and j^{th} pixels as defined in (Krähenbühl & Koltun, 2011) and x_i is the label assigned to the i^{th} pixel.

Total unary potential can be calculated as shown in Eq. 2,

$$\phi(x_i) = -\lambda_1 \log(P_d(x_i)) - \lambda_2 \log(P_p(x_i)) \qquad (2)$$

where $P_d(x_i)$ and $P_p(x_i)$ are the probabilities obtained from the external neural network model and the depth prior respectively. λ_1 , λ_2 are weights of individual po-

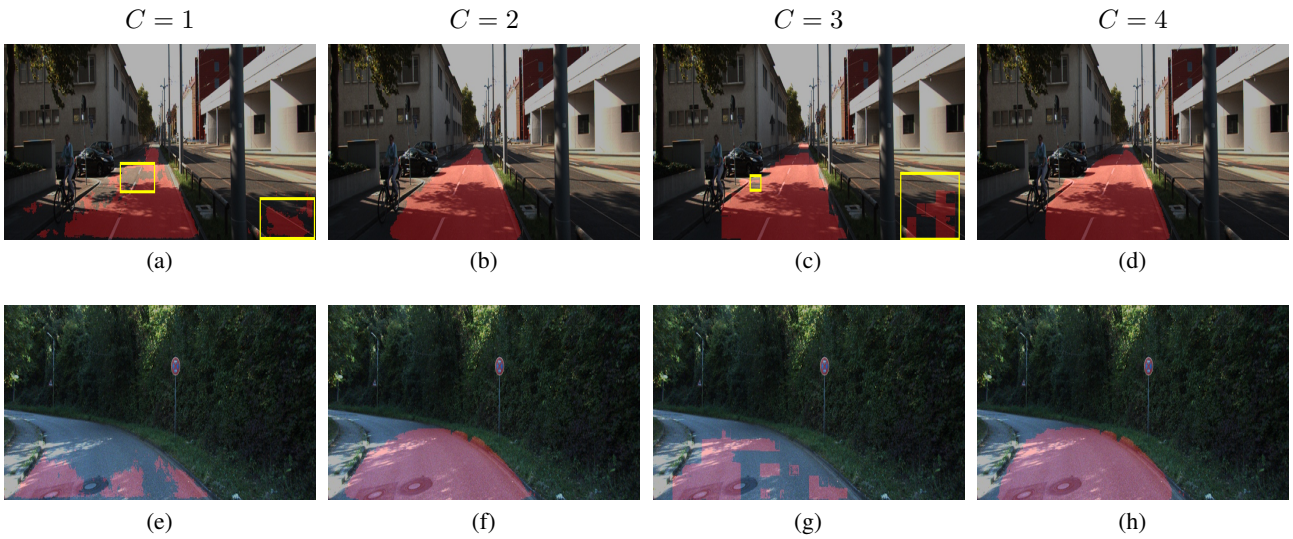


Figure 6. Comparison of Results; C indicates the column number. $C1$ represents the predictions of URNET, $C2$ shows the output of CRF fusion between URNET and the prior generated by the proposed algorithm. $C3$ displays the output of UVGG network, finally $C4$ represents the output of CRF fusion between UVGG and the prior generated by the proposed algorithm. Best viewed in color.

tentials. Then, the pair-wise potential is calculated as done in (Krähenbühl & Koltun, 2011).

3. Preliminary Results

Since the accuracy of results obtained using the proposed technique depends on the quality of the disparity map, preliminary results have been obtained on a subset of KITTI road dataset (Fritsch et al., 2013) which gave accurate disparity maps on applying the SGBM technique (Hirschmuller, 2008). Noisy disparity maps led to insignificant improvement using the proposed algorithm and these cases will be dealt with in future.

The accuracy of the proposed algorithm has been compared with only URNET and UVGG outputs. The parameters used for comparison are Precision, Recall, F1 Score and Intersection Over Union (IOU). The parameters were computed for a subset of images some of which are shown in Fig. 6. The average scores of the comparison parameters computed across the comparison dataset are shown in Table 1. Where, F-UVGG and F-URNET represent the output of UVGG and URNET respectively when fused with the proposed prior.

As can be seen from Fig. 6, the proposed algorithm generates high precision road masks in case of shadows on road. Mask generated from URNET in Fig. 6a misses a lot of road pixels and generates noisy estimates as well (highlighted in the image), which are detected correctly using the proposed method as shown in Fig. 6b. Similar is the case with mask generated from UVGG in Fig. 6c and its improved output in Fig. 6d. From Table 1, significant improvement can be seen in the F1 score and IOU values along with higher precision and recall values.

Table 1. Evaluation on a subset of the KITTI Road Dataset

MODEL	PRECISION	RECALL	F1 SCORE	IOU
UVGG	0.9187	0.7756	0.8192	0.7121
F-UVGG	**0.9522**	**0.9034**	**0.9264**	**0.8634**
URNET	0.9132	0.7675	0.8144	0.7062
F-URNET	**0.9548**	**0.9161**	**0.9341**	**0.8778**

Additional results are attached as supplementary material and can be referred to in appendix C.

4. Future Work and Conclusion

Possible fail cases occur when both v disparity based prior and the external deep neural network model give poor results as can be seen in Fig. 6e and Fig. 6g. The outputs generated in Fig. 6f and Fig. 6h shows some improvement but does not cover all road pixels . Since, v-Disparity based prior depends on the accuracy of disparity map, future work involves computing highly accurate disparity maps as proposed in (Chang & Chen, 2018) and improving the performance of RP-Net as well, making the overall algorithm more time efficient and robust.

This paper presented an approach to improve the performance of state of the art road segmentation algorithms in case of uneven textures. The prior road mask obtained from the shallow CNN model along with v disparity method enabled the whole system to perform better as can be seen from Table 1.

References

Badrinarayanan, V., Kendall, A., and Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

Caltagirone, L., Bellone, M., Svensson, L., and Wahde, M. Lidar – camera fusion for road detection using fully convolutional neural networks. *Robotics and Autonomous Systems*, 111:125 – 131, 2019. ISSN 0921-8890. doi: https://doi.org/10.1016/j.robot.2018.11.002.

Chang, J.-R. and Chen, Y.-S. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5410–5418, 2018.

Clevert, D.-A., Unterthiner, T., and Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). *CoRR*, abs/1511.07289, 2015.

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

Fernandes, R., Premebida, C., Peixoto, P., Wolf, D., and Nunes, U. Road detection using high resolution lidar. In *2014 IEEE Vehicle Power and Propulsion Conference (VPPC)*, pp. 1–6, Oct 2014. doi: 10.1109/VPPC.2014.7007125.

Fritsch, J., Kuehnl, T., and Geiger, A. A new performance measure and evaluation benchmark for road detection algorithms. In *International Conference on Intelligent Transportation Systems (ITSC)*, 2013.

Gallup, D., Frahm, J.-M., Mordohai, P., and Pollefeys, M. Variable baseline/resolution stereo. 06 2008. doi: 10.1109/CVPR.2008.4587671.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Hirschmuller, H. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, Feb 2008. ISSN 0162-8828. doi: 10.1109/TPAMI.2007.1166.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 448–456. JMLR.org, 2015.

Kim, B., Son, J., and Sohn, K. Illumination invariant road detection based on learning method. In *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pp. 1009–1014, Oct 2011. doi: 10.1109/ITSC.2011.6082917.

Krähenbühl, P. and Koltun, V. Efficient inference in fully connected crfs with gaussian edge potentials. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, pp. 109–117, Red Hook, NY, USA, 2011. Curran Associates Inc. ISBN 9781618395993.

Labayrade, R., Aubert, D., and Tarel, J. P. Real time obstacle detection in stereovision on non flat road geometry through "v-disparity" representation. In *Intelligent Vehicle Symposium, 2002. IEEE*, volume 2, pp. 646–651 vol.2, June 2002.

Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, June 2015. doi: 10.1109/CVPR.2015.7298965.

Maas, A. L., Hannun, A. Y., and Ng, A. Y. Rectifier nonlinearities improve neural network acoustic models. In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

Patra, S., Maheshwari, P., Yadav, S., Banerjee, S., and Arora, C. A joint 3d-2d based method for free space detection on roads. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv e-prints*, arXiv:1409.1556, 2014.

Zhao, J., Katupitiya, J., and Ward, J. Global correlation based ground plane estimation using v-disparity image. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pp. 529–534, 2007.

SUPPLEMENTARY MATERIAL

A. Stereo Setup Geometry

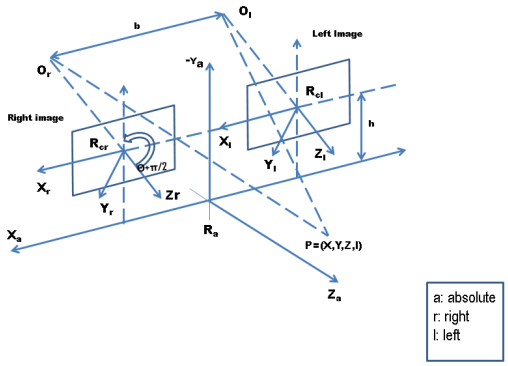


Figure A1. Stereo Setup Geometry

Fig. A1 shows the stereo camera setup geometry as in (Labayrade et al., 2002).

In Fig. A1, R_{cr} is the frame of reference of the right camera and accordingly R_{cl} is that of the left camera. R_a is the absolute plane or the road plane. θ is the orientation of the camera optical axis with the ground plane, b is the baseline and h is the height of the camera from the ground plane.

From (Labayrade et al., 2002), the ground plane in R_a represented by the Eq. A.1.

$$Z = aY + d; \tag{A.1}$$

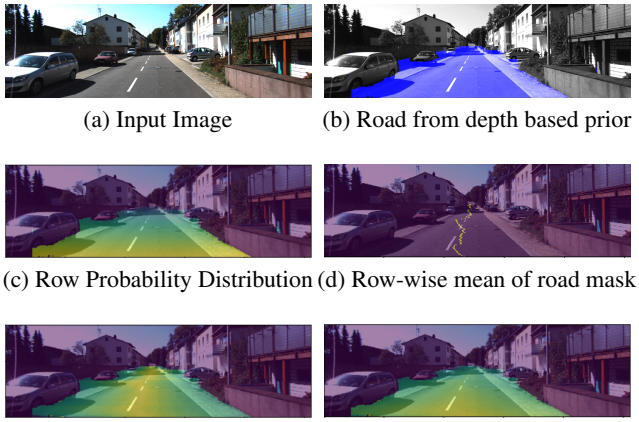
Any point P in the world with coordinates $(X, Y, Z, 1)$ with respect to R_a frame of reference has vertical coordinates in left and right frames as v_l and v_r which are equal to v as the stereo cameras are perfectly aligned represented in Fig. A1. From (Labayrade et al., 2002), the vertical coordinate or ordinate can be represented as shown in Eq. A.2.

$$v = \frac{[v_o \sin\theta + \alpha\cos\theta](Y+h) + [v_o\cos\theta - \alpha\sin\theta]Z}{[Y + h\sin\theta] + Z\cos\theta} \tag{A.2}$$

Here v_0 is the ordinate of the optical center in the image frame. α is $\frac{f}{t_c}$ where f is the camera focal length and t_c is the size of the pixel. Now from (Labayrade et al., 2002), the equation of line in the V-disparity space that corresponds to the road plane in R_a is

$$\Delta_M = \frac{b(v-v_o)[a\cos\theta + \sin\theta]}{ah-d} + \frac{b[\alpha(a\sin\theta - \cos\theta)]}{ah-d} \tag{A.3}$$

Here Δ_M is the disparity value corresponding to the ordinate v. Now by putting the values of the camera height (h), baseline (b) and the orientation angle (θ), the line corresponding to the road plane and hence the road is determined. As can be seen, there is dependence on camera orientation with respect to ground plane.



(a) Input Image (b) Road from depth based prior

(c) Row Probability Distribution (d) Row-wise mean of road mask

(e) Column Probability Distribution (f) Net Probability Distribution

Figure B2. Prior Probability Generation.

B. Probability generation for the prior

This section gives detailed explanation for the probability generation model for the depth based prior.

Since in a typical stereo vision system the error in depth estimation using stereo vision system increases quadratically with distance, a row decay model has been designed to get the row based probabilities from the mask obtained from the depth based prior. Column based probabilities are obtained using a similar column decay model. For this, a weighted mean is calculated for each row and the probability of accurate classification of a pixel as road decreases going farther from the mean.

Prior road mask is taken to be an image of size $H \times W$ where H and W are height (number of rows) and width (number of columns). In each row (k) of this image, a

weighted mean μ_k is determined using the Eq. B.4.

$$\mu_k = \frac{\sum_{j=0}^{cols-1} j * Vdisp[k][j]}{\sum_{j=0}^{cols-1} Vdisp[k][j]} \tag{B.4}$$

where $Vdisp$ is the prior road mask, $cols$ refers to total number of columns in image (1024 in current test cases) and j is the current column number.

For column probability distribution upper (u) and lower (l) limits of the columns are obtained for each row from first and last non zero column indices in the prior road mask. A variable val is defined in a way such that it's value is inversely proportional to the distance (along a row) from the weighted mean to the current pixel in the image.

$$val[k][j] = \begin{cases} \left(1 - \left(\frac{\beta*(\mu_k - j)}{\mu_k - l}\right)\right), & \text{if } l <= j <= \mu_k \\ \left(1 - \left(\frac{\beta*(j - \mu_k)}{u - \mu_k}\right)\right), & \text{if } \mu_k < j <= u \end{cases} \tag{B.5}$$

where, j represents the column index for k^{th} row. β represents the magnitude of linear decrease of the road probability over the range from the mean to the extremes(lower or upper limit). Its value varies from 0.3 on highways to 0.6 for urban cluttered environment. The probability distribution along the columns is given by Eq. B.6.

$$ColProb[k][j] = val[k][j] * \frac{Vdisp[k][j]}{255} \tag{B.6}$$

where $ColProb$ is the matrix containing the column probabilities of the road.

The row probability distribution, valid for $i > vRow$ is as shown in Eq. B.7.

$$RowProb[k][j] = \left(\frac{k - vRow}{Rows - vRow}\right)^{\alpha} * \frac{Vdisp[k][j]}{255} \tag{B.7}$$

$RowProb$ is the matrix containing the row probabilities. $vRow$ is the row number before which no road exists, $Rows$ is the total number of rows in the image (256 in this case) and α determines the probability decay rate. The value of α is set to 0.5 keeping in consideration that the depth error increases quadratically with the distance from the baseline. The net probability distribution (denoted by $RoadPrior$) of the depth prior is obtained from Eq. B.8

$$RoadPrior[k][j] = \frac{ColProb[k][j] + RowProb[k][j]}{2} \tag{B.8}$$

Fig. B2 demonstrates the probability generation model for the mask from depth based prior. Fig. B2a is the input image and the corresponding prior road mask is shown in Fig. B2b. Fig. B2c shows the generated row probability distribution. Weighted means calculated for each row are shown in Fig. B2d and the corresponding column probability distribution in Fig. B2e. The net probability distribution obtained by combining row & column distributions is depicted in Fig. B2f.

C. Additional Results

Fig. C1 shows some additional results on the KITTI dataset.

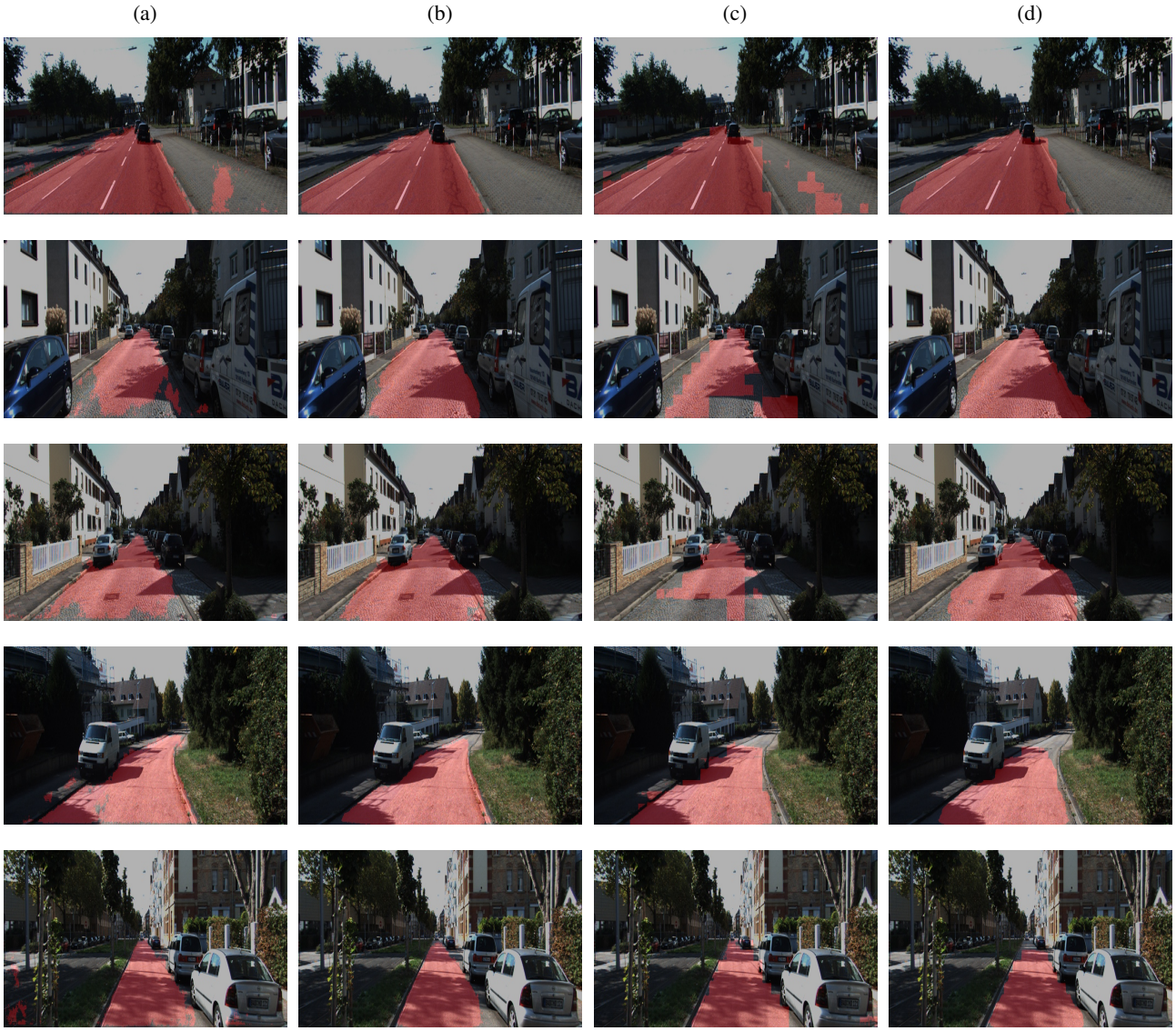


Figure C1. Comparison of Results. Column (a) represents the input images with the URNET predictions super-imposed, Column (b) shows the output of CRF fusion between URNET and the prior generated by the proposed algorithm. Column (c) displays the output of the UVGG network, finally column (d) represents the output of CRF fusion between UVGG and the prior generated by the proposed algorithm. Best viewed in color.