2018 IEEE Intelligent Vehicles Symposium (IV)
Changshu, Suzhou, China, June 26-30, 2018

Sensor Fusion of Intensity and Depth Cues using the ChiNet for Semantic Segmentation of Road Scenes

V. John[1], M. K. Nithilan[1], S. Mita[1], H. Tehrani[2], M. Konishi[2], K. Ishimaru[3], and T. Oishi[2]

***Abstract*— Vision-based environment perception is an important research topic for autonomous driving and advanced driver assistance systems. Vision sensors, such as the monocular camera and stereo camera, are widely used for environment perception. The monocular camera provides the appearance information like intensity, and the stereo camera provides the depth information. The appearance and depth information are complementary, and their effective fusion would result in robust environment perception. Consequently, in this paper, we propose a novel deep learning framework, termed as the ChiNet, for the effective sensor fusion of the appearance and depth information for free space and road object estimation. The ChiNet has two input branches and two output branches. The ChiNet input branches contains separate branches for the intensity and depth information. For the output branches, the ChiNet contains separate branches for the free space and road object semantic segmentation. A comparative of the proposed framework with state-of-the-art baseline algorithms is performed using an acquired dataset. Moreover, a detailed parameter analysis is performed to validate the ChiNet architecture as well as the advantages of sensor fusion. The experimental results show that the ChiNet is better than baseline algorithms. We also show that the proposed ChiNet architecture is better than other variations of the ChiNet architecture.**

I. Introduction

An increase in the need for driver safety as well as driver comfort has led to the rapid progress in the autonomous driving and advanced driver assistance systems (ADAS's) research community [1], [2]. In ADAS's and autonomous driving system, environment perception is an important task which is performed using an array of sensors such as LIDAR, monocular camera, stereo camera etc. Among these different sensors, the vision-based sensors are widely used as they are cost effective, while providing rich descriptive information of the environment. For example, the monocular camera provides the appearance information, and the stereo camera provides the depth information [3] (Fig 1). Since these information are complementary, researchers have sought to effectively fuse them to enhance the perception accuracy.

In this paper, we present a novel deep learning semantic segmentation framework, termed as the ChiNet, for the sensor fusion of intensity and depth information for environment perception in autonomous driving. The ChiNet is based on the encoder-decoder segmentation frameworks with the skip connections [4], [5]. The ChiNet has two separate input branches for the intensity and depth information, and two separate output branches for the free space and road object semantic segmentation. For effective semantic segmentation, the intensity and depth feature maps at the different encoding levels are transferred to the corresponding decoder levels using the skip connections. The sensor fusion and sharing of the intensity and depth features are performed at the free space and road object segmentation branches. In the proposed ChiNet network, by using separate input branches, the learnable convolutional filters individually extract the intensity and depth channel features in a more detailed manner. Additionally, the separate input branches are also shared by the separate output branches, increasing the segmentation accuracy, especially for small objects, objects at distance and pixels at inter-class boundaries (Fig 4).

The ChiNet is validated on an acquired dataset and a comparative analysis is performed with state-of-the-art baseline algorithms. Additionally, a detailed parameter analysis is performed using variations of the ChiNet. More specifically, we validate the multiple input branches, multiple output branches and intensity-depth sensor fusion. Based on the experimental results, we observe that the ChiNet is better than the baseline algorithms. The parametric analysis shows that the ChiNet with multiple input and multiple output branches are better than the ChiNet variations. To the best of our knowledge, the contribution to the literature are as follows:

- A novel deep learning-based semantic segementation framework is proposed for the sensor fusion of intensity and depth information for free space and road object estimation.
- Comparative analysis of different sensor fusion models with input branch variations of the ChiNet.
- Comparative analysis of multiclass and multilabel segmentation using output branch variations of the ChiNet.

We structure the paper as follows, we firstly review the literature in Section II. Next, we present the proposed deep learning framework along with the different variations in Section III. The experimental evaluation is presented in Section IV. The paper is finally concluded in Section V.

II. Related Work

Recently, deep learning-based semantic segmentation has been used for vision-based environment perception reporting state-of-the-art detection accuracy [4]–[9]. Long et al. [6], adapted the convolutional neural network (CNN) framework to semantic segmentation using the fully convolutional layers

[1] V. John, M. K. Nithilan and Seiichi Mita are with the Toyota Technological Institute, Japan {vijayjohn, nithilan,smita}@toyota-ti.ac.jp.

[2] H. Tehrani, M. Konishi and Tomoyuki Oishi are with DENSO CORPORATION, Japan {hossein_tehrani,masataka_konishi,tomoyuki_i_ooishi}@denso.co.jp

[3] K. Ishimaru is with Nippon Soken, Japan kazuhisa_ishimaru@soken1.denso.co.jp

978-1-5386-4452-2/18/$31.00 ©2018 IEEE

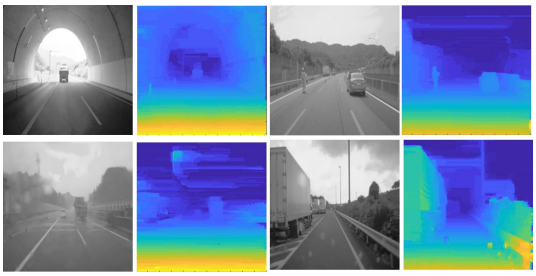

Fig. 1. Sample images from the acquired dataset with intensity and depth images.

and proposed the fully convolutional network (FCN). The accuracy of the FCN was improved by Noh et al. [7] using the encoder-decoder architecture in their deconvolutional network (Deconv). The encoding layers extract and subsample the feature maps from the input image using the convolutional and pooling layers. These sub-sampled feature maps, are then upsampled in the decoding layers using deconvolutional filters and max-pooling indices obtained from similar sized encoding layers. The Deconv was further improved in the Segnet [4] by using connections between the encoder and decoder layers. The U-Net [5] proposed by Ronneberger for the biomedical research community, also utilizes the encoder-decoder architecture to perform semantic segmentation. However, unlike the Segnet [4] and Deconvnet [7], in the U-Net [5], entire feature maps are transferred from the encoding layers to the decoding layers using skip connections, resulting in more precise output [5]. Since the aforementioned architectures, typically, perform environment perception using the monocular camera, their accuracy can be further improved using complementary sensor information [10], [11].

In literature, researchers have previously investigated deep learning for learning multimodal feature representation [12]–[15] which have been shown to be effective for pedestrian detection [16], [17]. In the work by Danut et al. [16], the authors use independent CNN branches to extract the intensity, depth and flow information, which are then combined using the multi-layer perceptron to perform pedestrian recognition. The authors show that their late-fusion scheme with multiple branches are better than an early-fusion scheme, where the multi-modal information is trivially combined and fed to an unique CNN. A similar observation is reported in the works of Wagner et al. [17], where visible and thermal images combined using the late-fusion report better pedestrian detection accuracy than an early-fusion model.

Apart from pedestrian detection, researchers have also sought to incorporate multimodal information for sensor fusion [18]–[20]. Gupta et al. [19] proposed a new representation, the HHA, as a three channel input to the CNN. By using the HHA, the authors integrated the color and depth information at the input in a pre-processing step. In the work by Li et al. [20], a novel LSTM and CNN model is proposed that fuses and combines the contextual color and depth information. Finally, in the work of Hazirbas2017 et al. [18], the depth and color information are fused within the encoder-decoder architecture termed as the FuseNet. The depth and color information are fused in the encoder region using element-wise summation, and used to perform semantic segmentation of indoor scenes.

Compared to the literature, in the proposed work, we propose a novel deep learning-based multimodal architecture for semantic segmentation. Multiple input and output branches are proposed in the novel architecture for effective semantic segmentation.

III. Algorithm

In this paper, a novel deep learning-based semantic segmentation architecture, termed as the ChiNet, is proposed for effective sensor fusion of intensity and depth information for semantic segmentation. The proposed architecture is based on encoder-decoder framework with skip connections [5]. Multiple input and output branches are formulated in the proposed architecture. Two separate input branches are proposed for the intensity and depth information in the encoder layers. Similarly, two separate output branches are proposed for the free space and road object semantic segmentation in the decoder layers. To enhance the segmentation accuracy, the feature maps in encoder layers are shared with the corresponding decoder layers in the separate output branches using the skip connections. An overview of the skip connections in the ChiNet architecture is presented in Fig 2. We next present the details of the proposed ChiNet architecture, Fig 3.

A. Proposed ChiNet Architecture

In semantic segmentation research, deep learning frameworks with the encoder-decoder architecture and connections have reported state-of-the-art detection accuracy [4], [5], [7] for monocular camera. In our work, we formulate a novel encoder-decoder architecture with skip connections, which performs semantic segmentation using both intensity and depth information. The proposed network, termed as the ChiNet, contains two input branches in the encoding region and two output branches in the decoding region. The proposed network is termed as the ChiNet owing to the resemblance to the greek alphabet χ.

In the ChiNet, the two input and output branches function as the feature extraction and semantic segmentation branches, respectively. In the feature extraction branches, the intensity and depth information are represented in two separate input branches. In the semantic segmentation branches, the free space and road object semantic segmentation are performed individually using the two separate output branches.

Each input branch contains multiple encoding-convolutional layers with learnable weights and pooling layers. Since we represent the intensity and depth information in separate input branches, the learnable weights are trained to extract intensity and depth specific feature maps in each individual branch. These specific features maps are then shared with the two output branches using skip connections. In the skip connections, the intensity

TABLE I

DIFFERENT VARIATIONS OF THE CHINET AND THE BASELINE ALGORITHMS.

Network	Input Branch	Output Branch	Connections
ChiNet SIMO	One branch with intensity and depth combined as a two channel input	Two branches with individual free space and object semantic segmentation	Skip connections between corresponding encoder layers in single input branch and decoder layers in both output branches
ChiNet MISO	Two branches with intensity and depth separate	One multiclass segmentation branch for free space and object semantic segmentation	Skip connections between corresponding encoder layers in both input branches and decoder layers in single output branch
ChiNet IIMO	One branch with intensity input	Two branches with individual free space and object semantic segmentation	Skip connections between corresponding encoder layers in single input branch and decoder layers in both output branches
U-Net [5]	One branch with intensity and depth combined as a two channel input	One multiclass segmentation branch for free space and object semantic segmentation	Skip connection between corresponding encoder layers in single input branch and decoder layers in single output branch
FuseNet [18]	Two branches with intensity and depth separate	One multiclass segmentation branch for free space and object semantic segmentation	Skip connection between intensity encoder layers and depth encoder layers

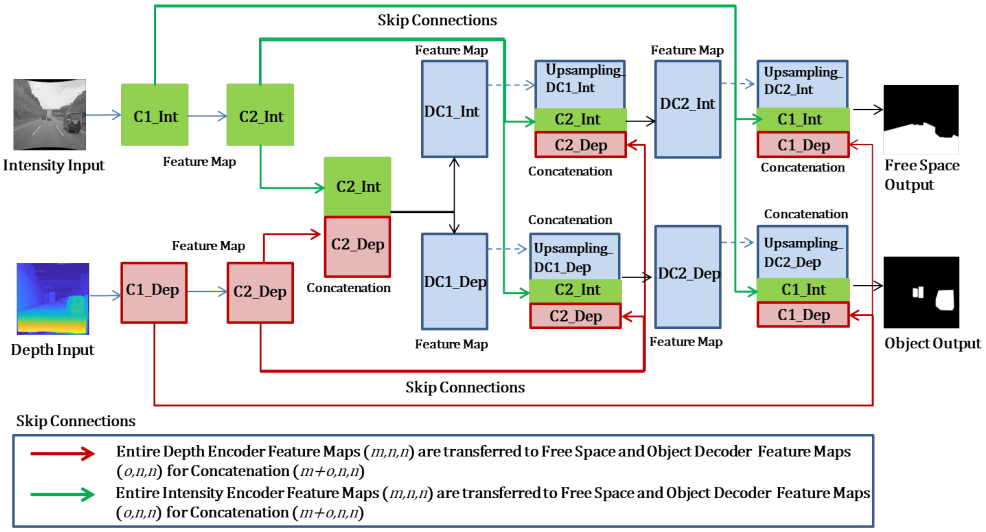

Fig. 2. An illustration of the skip connections in the proposed ChiNet

and depth features maps at the different encoding levels are transferred separately to the corresponding decoder levels in each output branch. Apart from being shared, the intensity and depth feature maps at each encoding layer are also fused at the corresponding decoding layer using the concatenation.

Each output branch contains multiple decoding-convolutional layers and upsampling layers. For a given output branch, each decoding-convolutional layer and upsampling layer receives the transferred intensity and depth feature map from the corresponding encoding-convolutional layer. The transferred encoding feature maps are fused with the decoding upsampled feature map using concatenation. Using the shared and fused feature maps, the individual output branches perform the semantic segmentation using the sigmoid activation function in the final layer. The binary cross-entropy error function is used individually by the output branches to perform the free space and road object semantic segmentation.

In the proposed network, due to the sharing of the input feature maps at the separate output branches, precise semantic segmentation of the free space and road objects are obtained. Compared to the state-of-the-art, the ChiNet obtains good segmentation results for small objects, objects at distance and pixels at the inter-class boundaries (Section IV). A detailed overview of the parameters in the ChiNet architecture is presented in Fig 3. The ChiNet was trained with a Adam optimizer with learning rate of 0.01, β_1 of 0.9, β_2 of 0.999 with no decay. The details of the batch size and training iterations are presented in the experimental section (Section IV).

B. Variations in the Network Architecture

We propose 3 different variations of the proposed ChiNet with varying number of input and output branches (Table I). In the ChiNet models with single output branch a softmax activation function in the output layer and a categorical cross entropy error function are used.

IV. EXPERIMENTAL RESULTS

The proposed algorithm is validated on acquired dataset with 9000 training and 3000 testing samples. The free space

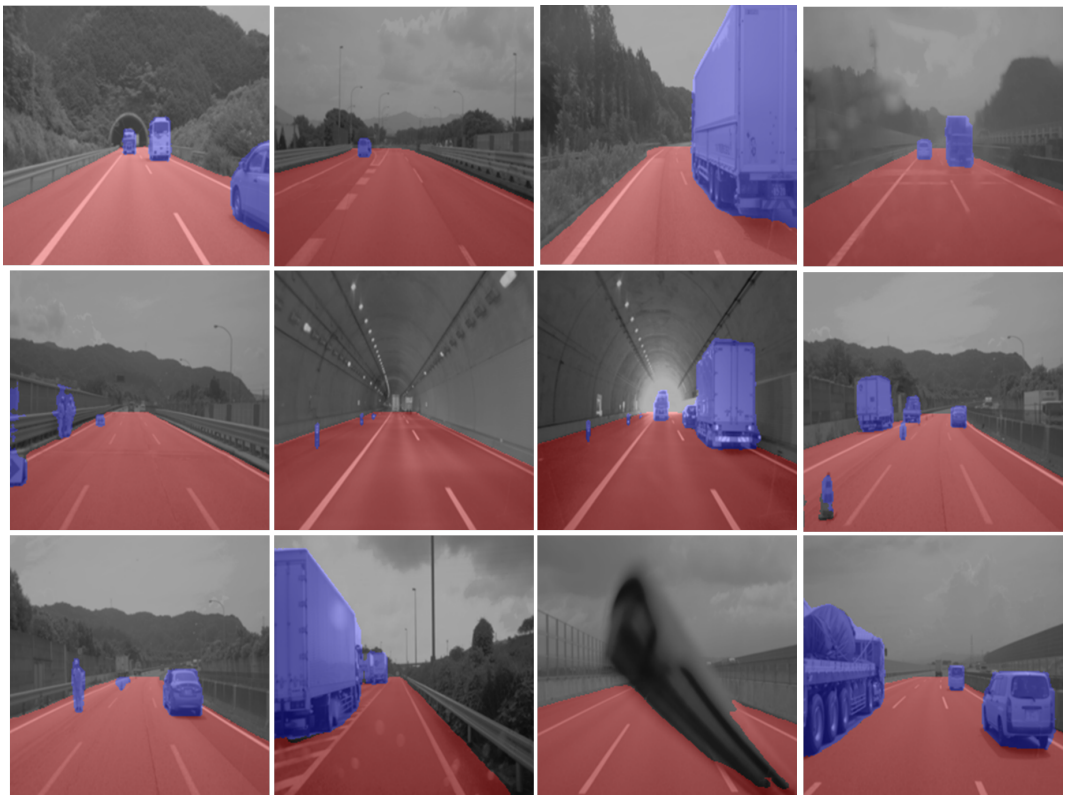

Fig. 4. Semantic segmentation results of the ChiNet for challenging road scenes containing rain, construction workers, construction cones, image flares etc. Red pixels denote the free space and blue pixels denote the road objects.

and road objects such as vehicles, construction workers, construction cones were manually annotated on the 12000 samples. The dataset was acquired with a stereo camera, which provided us with the left and right intensity images. The disparity images were generated using the MPV algorithm proposed by Long et al. [3]. The ChiNet and the variations were trained with the left intensity image and the disparity image with batch size 5 and 10000 iterations, and Adam optimizer. The networks were implemented on a Nvidia Titan X Ubuntu 16.04 machine using Keras-Theano backend. To validate the proposed framework, we perform a comparative analysis with baseline algorithms and the variations of the ChiNet.

A. *Comparative Analysis with Baseline*

For the comparative analysis, the performance of the ChiNet is compared with the performance of the U-Net [5] and the FuseNet [18]. In case of the U-Net the intensity and disparity maps are represented as a two channel input, while the free space and the road objects are estimated using the multiclass segmentation framework. Note that the U-Net architecture is similar to the ChiNet model with single input and single output branch. An overview of the baseline is presented in Table I.

In case of the FuseNet architecture, the intensity and depth are represented in two separate input branches and concatenated at the different encoding layers, while the free space and road objects are estimated using multiclass segmentation. The batch size and iterations were kept at 5 and 10000 for both the networks.

The segmentation accuracy of the different networks, tabulated in Table II, show that the ChiNet demonstrates better segmentation accuracy than the baseline networks, with computational complexity of 190ms. The results obtained by the ChiNet are shown in Fig 4.

TABLE II
COMPARATIVE ANALYSIS OF THE CHINET WITH BASELINE ALGORITHMS.

Algo.	Det. Rate %	Comp Time (ms)
ChiNet	97.35	192
U-Net. [5].	94.2	82
FuseNet [18]	95.2	125

B. *Analysis with ChiNet Variations*

The proposed ChiNet architecture was validated by modifying the input and output branches, and obtaining the segmentation results on the experimental dataset.

1) Input Channel Variations: In this experiment we validate the input branch variations, by comparing the performance of the proposed ChiNet framework with the ChiNet with single input branch (SIMO). The advantages of representing the intensity and depth in separate input branches are shown in Table III and Fig 5, where the proposed framework reports higher segmentation accuracy than the single input

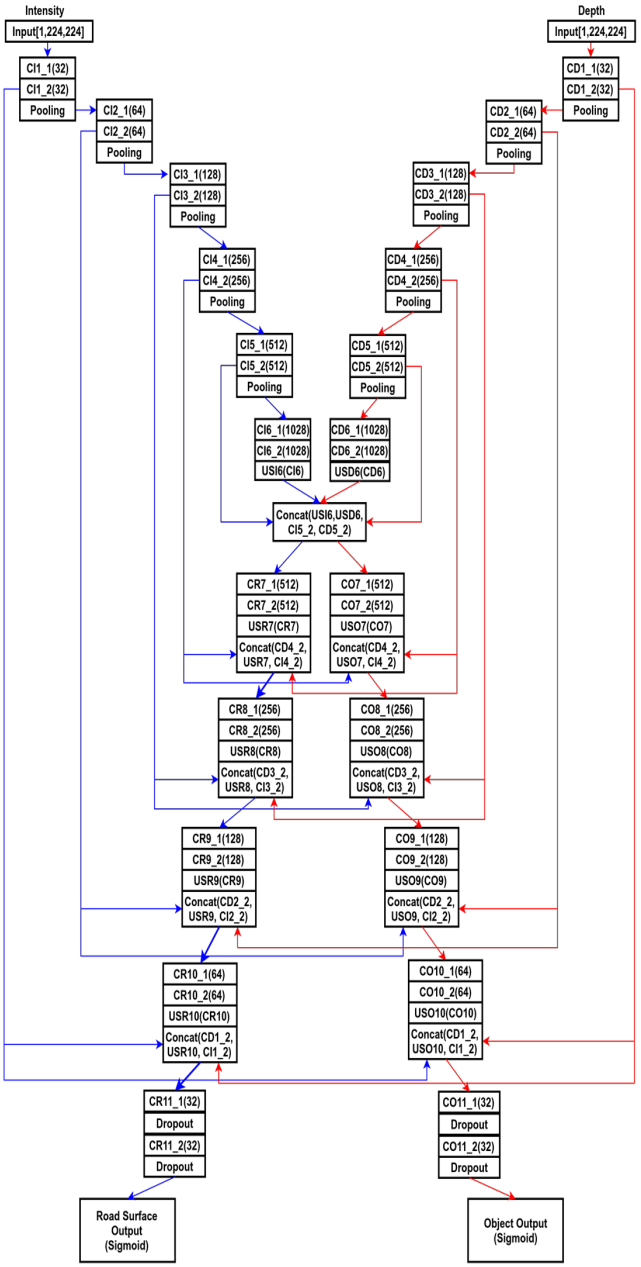

Fig. 3. A detailed overview of the proposed ChiNet

branch model. Moreover, the inter-class boundary pixels and small objects are better segmented in the proposed model. By representing the intensity and depth in separate input

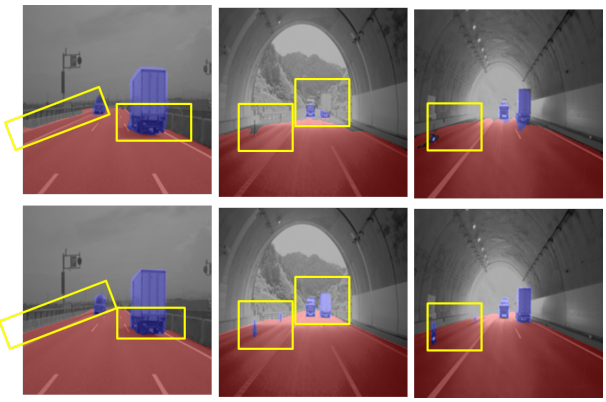

Fig. 5. Semantic segmentation results of the ChiNet SIMO in the top row and proposed ChiNet in the bottom row.

channels, the convolutional filters extract precise features relevant to the intensity and depth separately, enhancing the segmentation accuracy. However, as expected, the computational time for the proposed ChiNet is higher than the single input branch, owing to the increased number of input layers.

TABLE III
COMPARATIVE ANALYSIS OF THE CHINET VARIATIONS IN THE INPUT BRANCH.

Algo.	Det. Rate %	Comp Time (ms)
ChiNet (Proposed)	97.35	192
ChiNet SIMO	94.5	123

2) Output Channel Variations: In the second experiment, we validate the output branch variations, by comparing the performance of the proposed ChiNet framework with the ChiNet model with single output branch (ChiNet MISO), which performs the multiclass segmentation framework. The results tabulated in Table IV and Fig 6 show that the multilabel segmentation branch is better than the multiclass segmentation framework. This can be attributed to the precise free space and road object maps obtained in the separate output branches, resulting in fewer misclassified pixels and better segmentation of small objects.

TABLE IV
COMPARATIVE ANALYSIS OF THE CHINET VARIATIONS IN THE OUTPUT BRANCH

Algo.	Det. Rate %	Comp Time (ms)
ChiNet (Proposed)	97.35	192
ChiNet MISO	93.8	126

3) Sensor Fusion: In the final experiment, we validate the advantages of sensor fusion, by comparing the performance of the proposed ChiNet framework with the ChiNet model with intensity alone input branch (ChiNet IIMO). The results tabulated in Table V and Fig 7 demonstrate the advantage of integrating depth information into the semantic segmentation framework. By integrating depth information, the proposed

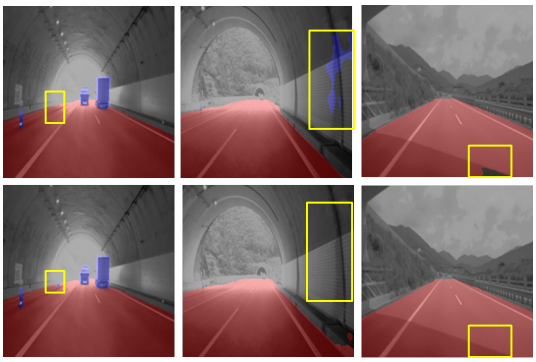

Fig. 6. Semantic segmentation results of the ChiNet MISO in the top row and proposed ChiNet in the bottom row.

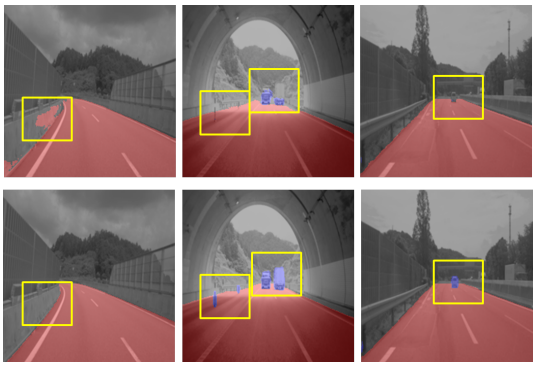

Fig. 7. Semantic segmentation results of the ChiNet IIMO in the top row and proposed ChiNet in the bottom row.

model improves the segmentation accuracy for small objects, objects at distance, and road boundaries.

TABLE V
COMPARATIVE ANALYSIS OF THE CHINET WITH INTENSITY ALONE-MODEL.

Algo.	Det. Rate %	Comp Time (ms)
ChiNet (Proposed)	97.35	192
ChiNet IIMO	95.8	123

C. Discussion

Based on the experimental results, we can observe that the ChiNet is better than the baseline and the ChiNet variations. This can be attributed to the:

- Individual feature extraction and semantic segmentation branches.
- Sharing and fusion of the encoder feature maps at the different decoder levels.
- Integration of the depth information to the intensity information.

V. CONCLUSION

In this paper, we presented a deep learning semantic segmentation framework, termed as the ChiNet, for the simultaneous sensor fusion of intensity and depth information and estimation of the free space and visible road objects. The ChiNet is based on the encoder-decoder architecture and contains multiple input and output branches. The intensity and depth information are represented in the input branches, while the free space and road object segmentation are represented in the output branches. The feature maps in the input channels are shared and fused at the different output branches. The proposed network is compared with baseline algorithms as well as the variations of the ChiNet. The experimental results show that the proposed algorithm is better than the state-of-the-art. In our future work, we will reduce the computational efficiency of the network.

REFERENCES

[1] V. John, Z. Liu, C. Guo, S. Mita, and K. Kidono, "Real-time lane estimation using deep features and extra trees regression," in *PSIVT*, 2015.

[2] V. John, C. Guo, S. Mita, K. Kidono, C. Guo, and K. Ishimaru, "Fast road scene segmentation using deep learning and scene-based models," in *ICPR*, 2016.

[3] Q. Long, Q. Xie, S. Mita, H. Tehrani, K. Ishimaru, and C. Guo, "Real-time dense disparity estimation based on multi-path viterbi for intelligent vehicle applications," in *Proceedings of the British Machine Vision Conference*, 2014.

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," in *CVPR.*, 2015.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *in CVPR*, Nov. 2015.

[7] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation." *CoRR*, vol. abs/1505.04366, 2015.

[8] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *CoRR*, vol. abs/1606.02147, 2016.

[9] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *CoRR*, vol. abs/1612.01105, 2016.

[10] A. Eitel, J. T. Springerberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust rgb-d object recognition." in *Proc. of The International Conference on Intelligent Robots and Systems (IROS)*, 2015.

[11] V. John, Y. Xu, S. Mita, Q. Long, and Z. Liu, "Registration of gps and stereo vision for point cloud localization in intelligent vehicles using particle swarm optimization," in *ICSI*, 2017.

[12] A. Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *NIPS*, 2014.

[13] D. Xu, Y. Yan, E. Ricci, and N. Sebe., "Detecting anomalous events in videos by learning deep representations of appearance and motion." *CVIU*, pp. 117–127, 2016.

[14] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe, "Learning deep representations of appearance and motion for anomalous event detection," in *BMVC*, 2015.

[15] J. Hoffman, S. Gupta, and T. Darrell, "Learning with side information through modality hallucination," in *CVPR*, 2016.

[16] O. Danut, A. Rogozan, F. Nashashibi, and A. Bensrhair, "Fusion of stereo vision for pedestrian recognition using convolutional neural networks," in *ESANN*, 2017.

[17] J. Wagner, V. Fischer, M. Herman, and S. Behnke, "Multispectral pedestrian detection using deep fusion convolutional neural networks," in *ESANN*, 2016.

[18] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, *FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture*, 2017.

[19] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik1, "Learning rich features from rgb-d images for object detection and segmentation," in *ECCV*, 2014.

[20] Z. Li, Y. Gan, X. Liang, Y. Yu, H. Cheng, and L. Lin, "Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling," in *ECCV*, 2016.