A Hybrid Spatial-Temporal Deep Learning Architecture for Lane Detection

Dhruv Patel [1] **Fenil Desai** [1] **Prajwal Poojari** [1] **Samarth Shah** [1]

Abstract

This project performs lane detection on continuous driving scenes and approaches it as a segmentation task. The network implements a novel hybrid spatial-temporal sequence-to-one deep learning architecture that integrates the following aspects: (a) the single image feature extraction module equipped with the spatial convolutional neural network (SCNN); (b) the spatial-temporal feature integration module constructed by spatial-temporal recurrent neural network (ST-RNN); (c) the encoder-decoder structure, which makes this image segmentation problem work in an end-to-end supervised learning format. Several experiments were performed to measure the accuracy, precision, recall and F1 measure of various networks formed by a combination of different variants of ST-RNN and Encoder-Decoder modules along with SCNN module.

1. Introduction

Accurate and reliable lane detection is vital for the safe performance of Lane Keeping Assistance and Lane Departure Warning systems. However, under certain challenging circumstances, it is difficult to achieve satisfactory performance in accurate lane detection from a single image as mostly done in previous literature. Since lane markings are continuous lines, the lanes that are difficult to be accurately detected in the current single image can potentially be better deduced if information from previous frames is incorporated. Also, the available methods do not take full advantage of the essential properties of the lane being long continuous solid or dashed line structures.

This project attempts to make the most of spatial-temporal information together with correlation and dependencies in continuous driving frames. The project is an exact implementation of (Dong et al., 2022). In this paper lane detection is treated as a segmentation task, in which a novel hybrid spatial-temporal sequence-to-one deep learning architecture is developed for lane detection through a continuous sequence of images in an end-to-end approach. To cope with challenging driving situations, the hybrid network takes multiple continuous frames of an image sequence as inputs, and integrates the single image feature extraction module, the spatial-temporal feature integration module, together with the encoder-decoder structure to make full use of the spatial-temporal information in the image sequence.

1.1. Research contributions

(Dong et al., 2022) introduces two major components in addition to the traditional Deep Learning segmentation modules like U-Net (Ronneberger et al., 2015) or SegNet (Badrinarayanan et al., 2017) for tackling the problem of lane detection. The single image feature extraction module utilizes modified common backbone networks with embedded spatial convolutional neural network (SCNN) (Pan et al., 2017a) layers to extract the features in every single image throughout the continuous driving scene. Next, the extracted features are fed into spatial-temporal recurrent neural network (ST-RNN) layers to capture the spatial-temporal dependencies and correlations among the continuous frames.

2. Related Work

SCNN: The Spatial Convolutional Neural Network (SCNN) was first proposed by (Pan et al., 2017b). The ”spatial” here means that the specially designed CNN can propagate spatial information via slice-by-slice message passing. The detailed structure of SCNN is demonstrated in Figure 1. SCNN has demonstrated its strengths in extracting spatial relationships in the image, which makes it suitable for detecting long continuous shape structures, e.g., traffic lanes, poles, and walls (Pan et al., 2018). However, using only one image for detection using SCNN still could not produce satisfying performance under extremely challenging conditions.

(Zou et al., 2020b) combines the CNN and RNN for lane

[1]Worcester Polytechnic Institute. Correspondence to: Dhruv Patel <dspatel@wpi.edu>, Fenil Desai <fdesai@wpi.edu>, Prajwal Poojari <pgpoojari@wpi.edu>, Samarth Shah <sshah5@wpi.edu>.

CS/DS 541 Deep Learning Final Project, Worcester Polytechnic Institute.

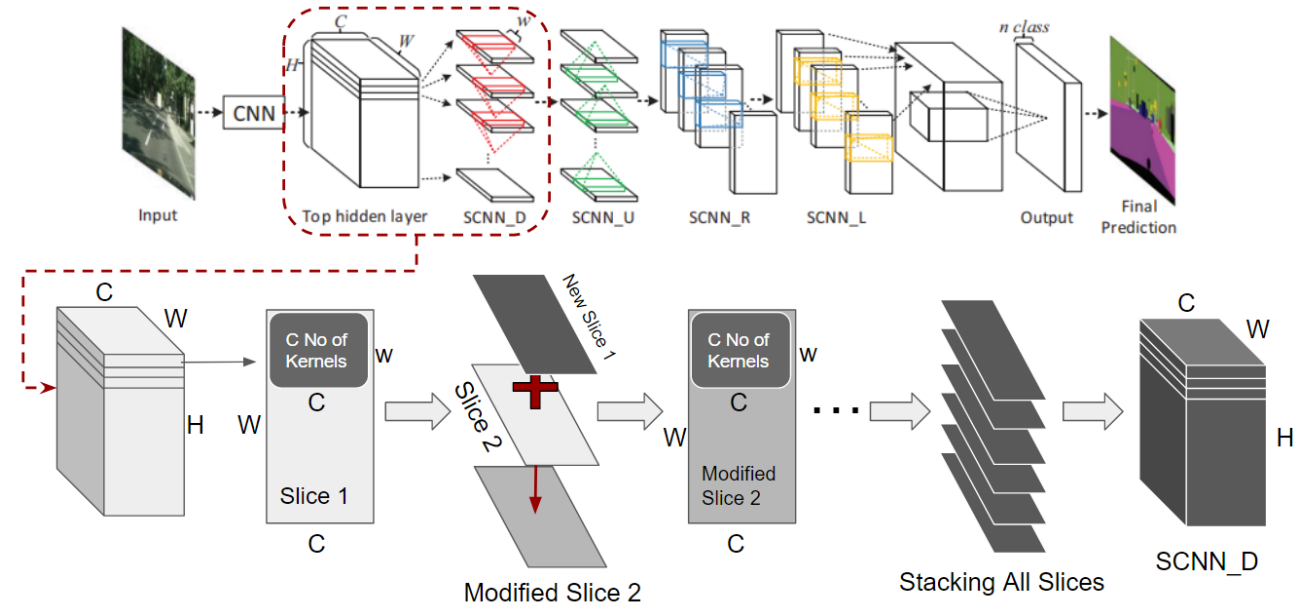

Figure 1. Spatial Convolutional Neural Network

detection, with a number of continuous frames of driving scene. In a continuous driving scene, the lanes in consecutive images captured by automobile cameras are commonly overlapped, which enables lane detection in a time-series prediction framework. In order to integrate CNN and RNN as an end-to-end trainable network, the authors **sandwich the RNN** in a convolutional encoder-decoder framework. With a number of continuous frames as input, the encoder processes each of them and get a time-series of feature maps. These feature maps are input to the LSTM network for lane-information prediction. The output of LSTM is then fed to the decoder CNN to produce a probability map for lane prediction. The lane probability map has the same size of the input image.

ConvLSTM Network: Referring to (Zou et al., 2020b), LSTM generally outperforms the traditional RNN model with its ability in forgetting unimportant information and remembering essential features, by using cells in the network to judge whether the information is important or not. A double-layer LSTM is applied, with one layer for sequential feature extraction and the other for integration. The traditional full-connection LSTM is slow and computationally expensive model. Therefore, the authors utilized the convolutional LSTM (ConvLSTM) (Wong et al., 2016) in (Zou et al., 2020b). The ConvLSTM replaces the matrix multiplication in every gate of LSTM with convolution operation, which is widely used in end-to-end training and feature extraction using time-series data.

However, this inclusion by the authors is still not able to utilize the inherent property of the lanes, i.e the lanes are long continuous lines having great spatial relationships. A specialized network for extracting spatial features from a single image can be utilized to make use of lanes being long continuous lines. Therefore, we decided to implement the network presented in (Dong et al., 2022) as it includes SCNN (Pan et al., 2017b) in a similar architecture with ST-RNN and Encoder-Decoder network for efficient lane detection.

3. Proposed Method

The proposed deep neural network architecture in (Dong et al., 2022) adopts a sequence-to-one end-to-end encoder-decoder structure as shown in Figure 2. Here ”sequence-to-one” means that the network gets a sequence of images as input and outputs the detection result of the last image (please note that essentially the network is still utilizing sequence-to-sequence neural networks); ”end-to-end” means that the learning algorithm goes directly from the input to the desired output(which refers to the lane detection result) after implementing the intermediate states; the encoder-decoder structure is a modular structure that consists of an encoder network and a decoder network. Here, the proposed network adopts encoder CNN with SCNN layers and decoder CNN using fully convolutional layers. The encoder takes a sequence of continuous image frames, i.e., time-series-images, as input and abstracts the feature map(s) in smaller sizes. To extract the knowledge that traffic lanes are solid- or dashed- line structures with a con-

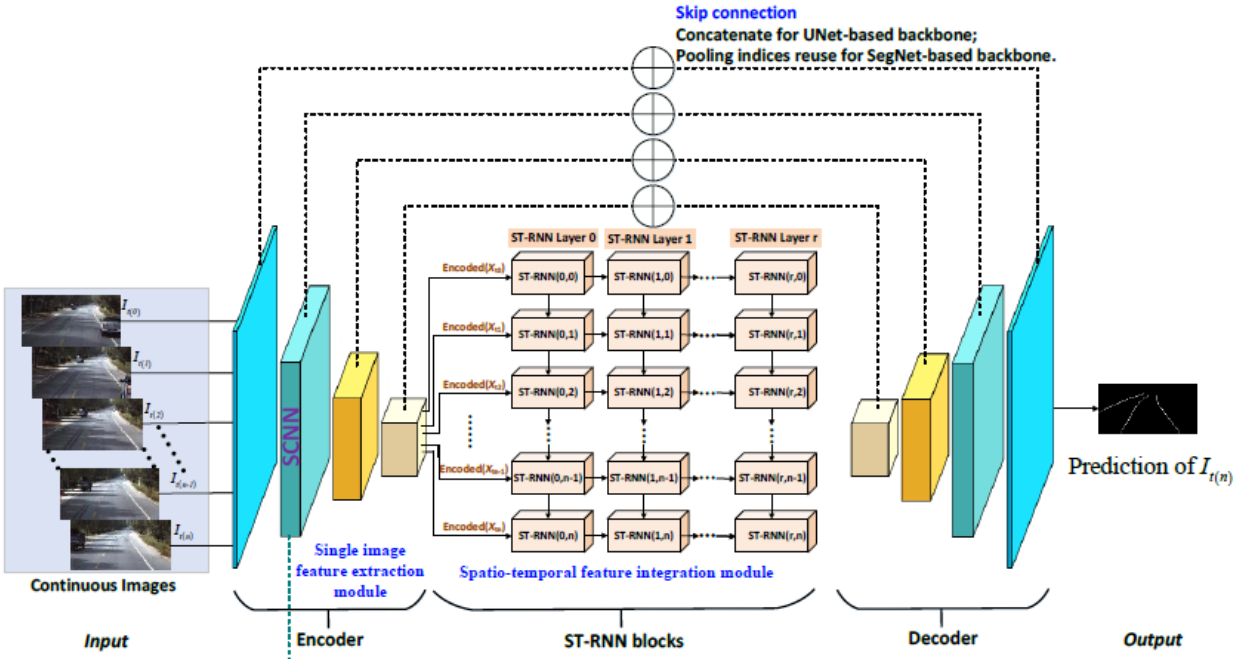

Figure 2. The architecture of proposed network

tinuous shape, one special kind of CNN, i.e., SCNN, is adopted after the first CNN hidden layer. With the help of SCNN, spatial features and relationships in every single image will be better extracted. Following this, the extracted feature maps of the continuous frames, constructed in a time-series manner, will be fed to ST-RNN blocks for sequential feature extraction and spatial-temporal information integration. Finally, the decoder network upsamples the abstracted feature maps obtained from the ST-RNN and decodes the content to the original input image size with the detection results.

4. Experiment

Multiple network consisting of the three modules (i.e, SCNN, ST-RNN and Encoder-Decoder) were implemented from scratch, trained and tested. However, the network designed is same as presented in (Dong et al., 2022). The various network implemented are UNet_ConvLSTM2, SCNN_UNet_ConvLSTM2, SCNN_SegNet_ConvLSTM2, SCNN_SegNet_ConvGRU2.

These networks were trained using the tvtLANE dataset, which is a modified version of the TuSimple dataset containing 19383 sequences. Sequence of images are incorporated as this is a sequence to one model. Here a sequence corresponds to five images and one ground truth corresponding to the 5th image. The dataset contains normal road condition as well as challenging driving scenes such as occlusion, blur, shadow, tunnel and degraded roads. For the training and testing of the networks both the personal laptop (Nvidia RTX3060 Ti, 4GB) and the Turing machine by WPI HPC (Nvidia Tesla P100-SXM2, 4GB) were utilized. The proposed networks are evaluated quantitatively in terms of various evaluation metrics. One of the metrics used is accuracy but since it is an imbalanced binary classification task(i.e., the lane pixels are fewer compared to the background pixels), accuracy is not a good metric for comparing the networks. Hence the other three metrics: Recall, Precision and F1-measure are introduced as indicated in the equations below:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN} \tag{1}$$

$$F1 - measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2}$$

Where,
TP is True Positive which indicates the correctly identified number of image pixels that are lane markings
FP is False Positive which indicates the number of background pixels being incorrectly identified as lane markings
FN is False Negative which indicates the incorrectly identified number of image pixels that are lane markings
F1-measure is a good metric for comparison of the networks.

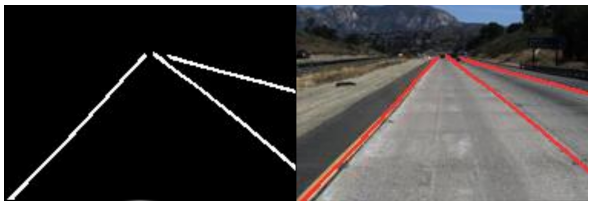

Figure 3. Lane Detection in Normal Condition

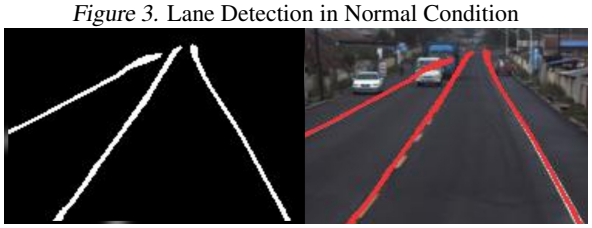

Figure 4. Lane Detection under occlusion

Figure 5. Lane Detection inside a tunnel

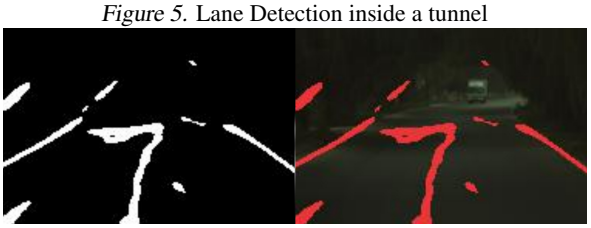

Figure 6. Failed Lane Detection in low lighting condition

5. Results

From the Experimentation and Results table (* represents using pre-trained weights) it is evident that after addition of SCNN module to the network, the accuracy and increases as it efficiently captures spatial information from a single image. It is also observed that ConvLSTM performs better as compared to ConvGRU on all the evaluation matrix since ConvLSTM can better extract spatial-temporal features as it has more control gates and thus more parameters as compared to ConvGRU. The results obtained using SCNN_UNet_ConvLSTM_2* as seen in Figure 3., 4., 5. and 6. The network performs well in normal conditions as seen in Figure 3. and also in adverse conditions such as lane occlusion and shadow as observed in Figure 4. and 5. However we can also observe that the network fails in certain conditions as seen in Figure 6.

6. Discussion

Due to limited computational resources and time, the networks were trained on 1/3 of the tvtLane dataset. The training time for this network was around 12 hours on Turing machine by WPI HPC cluster and 36 hours on personal machine for single epoch. Therefore, the networks were trained only for 3 epochs and hence the reported results does not fully justify the capability of the networks.

Based on the experiments and results we can observe that the SCNN does improve the network. However, it increases the training time. This is due to the fact that in SCNN, convolution is done on each slice which increases the number of operations and learnable parameters. This limitation can be tackled by incorporating a lighter encoder and decoder model such as E-Net which can result in faster training and also decrease the inference time while testing. However it could effect the evaluation metrics which can be interesting to observe.

7. Conclusions and Future Work

This project is an exact implementation of a novel deep learning architecture for lane detection presented in (Dong et al., 2022) utilizing (Zou et al., 2020a) and (Xingang Pan & Tang, 2018). As it is observed the addition of SCNN makes the model heavy, a future research direction might be to implement a light encoder-decoder architecture to reduce the training and inference time. Furthermore to enhance lane detection network a custom loss function, pre-trained techniques adopted in image-inpainting task and sequential attention models can be added to the network.

** Find the Code implementation HERE

Experimentation and Results							
Network	**Dataset**	**Epochs**	**Machine**	**Accuracy**	**Recall**	**Precision**	**F1-Measure**
UNet_ConvLSTM_2	Full	30	Turing	96.17%	0.9801	0.7560	0.8670
SCNN_UNet_ConvLSTM_2*	1/3	3	Personal	97.68%	0.9613	0.8320	0.8919
SCNN_UNet_ConvLSTM_2	1/3	3	Personal	95.18%	0.9750	0.8040	0.8812
SCNN_SegNet_ConvLSTM_2	1/3	3	Personal	95.03%	0.9501	0.8183	0.8792
SCNN_SegNet_ConvGRU_2	1/3	3	Personal	94.87%	0.9498	0.7982	0.8674

References

Badrinarayanan, Vijay, Kendall, Alex, and Cipolla, Roberto. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12): 2481–2495, 2017. doi: 10.1109/TPAMI.2016.2644615.

Dong, Yongqi, Patil, Sandeep, Arem, B., and Farah, Haneen. A hybrid spatial–temporal deep learning architecture for lane detection. Technical report, 02 2022.

Pan, Xingang, Shi, Jianping, Luo, Ping, Wang, Xiaogang, and Tang, Xiaoou. Spatial as deep: Spatial cnn for traffic scene understanding, 2017a. URL https://arxiv.org/abs/1712.06080.

Pan, Xingang, Shi, Jianping, Luo, Ping, Wang, Xiaogang, and Tang, Xiaoou. Spatial as deep: Spatial cnn for traffic scene understanding. 12 2017b.

Ronneberger, Olaf, Fischer, Philipp, and Brox, Thomas. U-net: Convolutional networks for biomedical image segmentation. In Navab, Nassir, Hornegger, Joachim, Wells, William M., and Frangi, Alejandro F. (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Wong, Wai Kin, Shi, Xingjian, Yeung, Dit-Yan, and WOO, Wang-chun. A deep-learning method for precipitation nowcasting. 07 2016.

Xingang Pan, Jianping Shi, Ping Luo Xiaogang Wang and Tang, Xiaoou. Spatial as deep: Spatial cnn for traffic scene understanding. In *AAAI Conference on Artificial Intelligence (AAAI)*, February 2018.

Zou, Q., Jiang, H., Dai, Q., Yue, Y., Chen, L., and Wang, Q. Robust lane detection from continuous driving scenes using deep neural networks. *IEEE Transactions on Vehicular Technology*, 69(1):41–54, 2020a.

Zou, Qin, Jiang, Hanwen, Dai, Qiyu, Yue, Yuanhao, Chen, Long, and Wang, Qian. Robust lane detection from continuous driving scenes using deep neural networks. *IEEE Transactions on Vehicular Technology*, 69(1):41–54, 2020b. doi: 10.1109/TVT.2019.2949603.