Scene-Adaptive Driving Area Prediction based on Automatic Label Acquisition from Driving Information

Takuya Migishima[1], Haruya Kyutoku[2], and Daisuke Deguchi[1]
Yasutomo Kawanishi[1] Ichiro Ide[1] Hiroshi Murase[1]

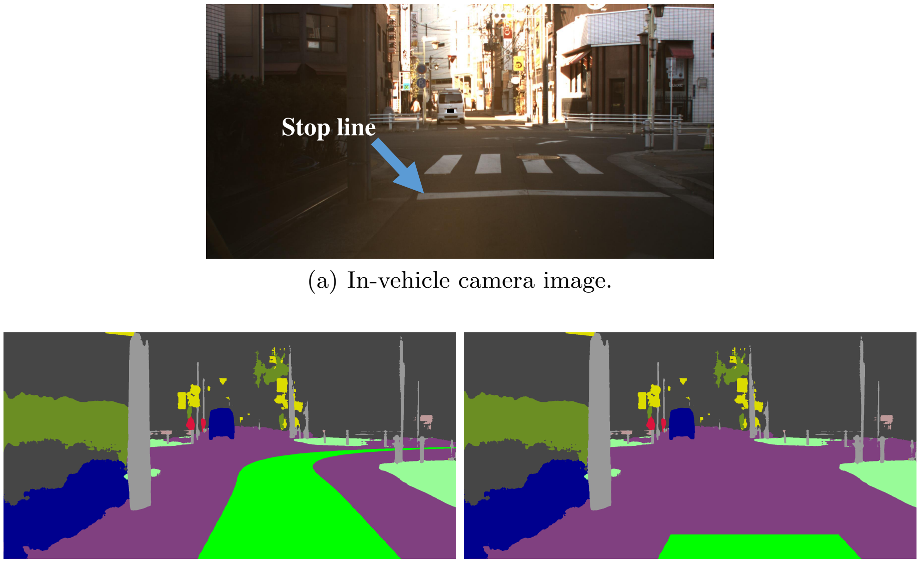
[1] Nagoya University, Nagoya, Aichi, Japan
migishimat@murase.is.i.nagoya-u.ac.jp
[2] Toyota Technological Institute, Nagoya, Aichi, Japan

Abstract. Technology for autonomous vehicles has attracted much attention for reducing traffic accidents, and the demand for its realization is increasing year-by-year. For safety driving on urban roads by an autonomous vehicle, it is indispensable to predict an appropriate driving path even if various objects exist in the environment. For predicting the appropriate driving path, it is necessary to recognize the surrounding environment. Semantic segmentation is widely studied as one of the surrounding environment recognition methods and has been utilized for drivable area prediction. However, the driver's operation, that is important for predicting the preferred drivable area (scene-adaptive driving area), is not considered in these methods. In addition, it is important to consider the movement of surrounding dynamic objects for predicting the scene-adaptive driving area. In this paper, we propose an automatic label assignment method from actual driving information, and scene-adaptive driving area prediction method using semantic segmentation and Convolutional LSTM (Long Short-Term Memory). Experiments on actual driving information demonstrate that the proposed methods could both acquire the labels automatically and predict the scene-adaptive driving area successfully.

Keywords: Semantic segmentation · Path prediction · Autonomous vehicle.

1 Introduction

Technology for autonomous vehicle has attracted much attention for reducing traffic accidents, and the demand for its realization is increasing year-by-year. Although some automotive manufacturers are already providing autonomous driving functions on expressways, it is still difficult to provide such functions on urban roads due to the diversity of the surrounding environment. To drive safely on urban roads, it is indispensable to predict an appropriate driving path even if various objects exist in the environment. Here, the path can be considered as a trajectory of future vehicle positions, which should be predicted appropriately since it will affect vehicle control during autonomous driving. Since vehicle



(a) In-vehicle camera image.

(b) Example of drivable area.

(c) Example of preferred drivable area.

Fig. 1. Difference between ordinary path and scene-adaptive path.

driving paths heavily depend on the environment, it is necessary to recognize the surrounding objects, the situation of pedestrians, and other vehicles in the environment in detail for path prediction. Therefore, the technology to predict a future vehicle path is strongly needed especially for autonomous driving in urban environment.

On the other hand, semantic segmentation, which is a task to predict object labels pixel-by-pixel in an image, is widely studied as one of the surrounding environment recognition methods [2, 7]. Barnes et al. [1] and Zhou et al. [8] tried to extend the problem of semantic segmentation to that of predicting the drivable area that cannot be observed as image features. They constructed the semantic segmentation model from the training data generated by projecting the trajectory of vehicle positions onto the road surface. For example, the green area in Fig. 1(b) indicates the drivable area label that is used as a ground-truth by Zhou's method. As shown in Fig. 1(a), which is the original image corresponding to Fig. 1(b), there is a stop sign on the path. However, as shown in Fig. 1(b), Zhou's method does not consider this sign and thus the necessity of braking. In addition, although the drivable area should be adaptively changed by considering the context of surrounding dynamic objects (e.g. pedestrians, vehicles, etc.), Zhou's method does not consider the necessity of braking against those objects. Thus, we tackle the problem of predicting the preferred drivable area considering the safety required for automated driving and the movement of surrounding dynamic objects. Figure 1(c) shows an example of the preferred drivable area in the context that can follow traffic rules around the intersection.

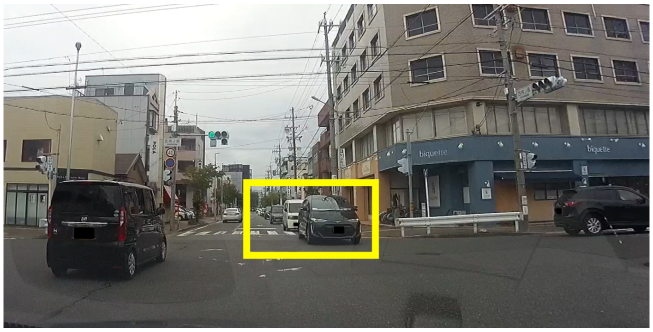


Fig. 2. Example of a scene containing oncoming vehicles. The yellow rectangle indicates the oncoming vehicles.

Since the preferred drivable area should adapt to the environment, we call this as the "scene-adaptive driving area", and propose a method for its prediction in this paper.

To predict the scene-adaptive driving area accurately, it is necessary to solve the following two issues: (i) Generation of the ground-truth labelling of the scene-adaptive driving area, (ii) How to handle the object movement in the vehicle front. As described above, since the scene-adaptive driving area should end before stop signs or other objects, it is necessary to prepare training data satisfying this requirement for training a semantic segmentation model that can predict a scene-adaptive driving area. On the other hand, the scene-adaptive driving area should change dynamically due to the existence and the movement of oncoming vehicles, pedestrians, and other objects. Figure 2 shows an example of a scene containing an oncoming vehicles. In this situation, we cannot cross the road safely without considering the moving state of the oncoming vehicles indicated by the yellow rectangle in the image. That is, it is necessary to consider the movement of surrounding objects for predicting an appropriate scene-adaptive driving area.

To tackle the first issue, we refer to the vehicle speed to generate the training data. If a driver operates the brake pedal, we can assume that the main cause for that should exist in the vehicle front. However, there are possibilities such as the existence of blind intersections, traffic signs, or red traffic signals. From these points-of-view, we try to generate appropriate labels for scene-adaptive driving area automatically by referring to the reduction of the own vehicle speed as a key.

To tackle the second issue, we introduce feature learning from frame sequences. In a deep learning framework, LSTM (Long Short-Term Memory), which is a variant of RNN (Recursive Neural Network), is one of the popular techniques to handle frame sequences. Generally speaking, LSTM can be used to learn sequential (temporal) information, but it loses spatial information. To preserve the spatial information in LSTM, we use ConvLSTM (Convolutional Long Short-Term Memory) proposed by Shi et al. [6]. The ConvLSTM is a network where fully connected layers in LSTM are replaced with convolutional layers. It can be applied to the semantic segmentation task for predicting labels of a future frame [4]. The proposed method incorporates ConvLSTM in the prediction of the scene-adaptive driving area to learn the movement of other objects.

Based on the above concept, we propose a method to automatically acquire training data from actual driving information and predict a scene-adaptive driving area. The main contributions of this paper can be summarized as follows:

- Proposal of the concept of "scene-adaptive driving area" considering the necessity of braking.
- Automatic label assignment for training a model to predict a scene-adaptive driving area referring to the own vehicle's speed.
- Prediction of a scene-adaptive driving area that considers the movement of other objects using ConvLSTM.

In the following, Sec. 2 proposes the automatic label assignment method, Sec. 3 describes the construction of the scene-adaptive driving area predictor using ConvLSTM, Sec. 4 reports the evaluation experiments, and Sec. 5 concludes this paper.

2 Automatic label assignment using driving information

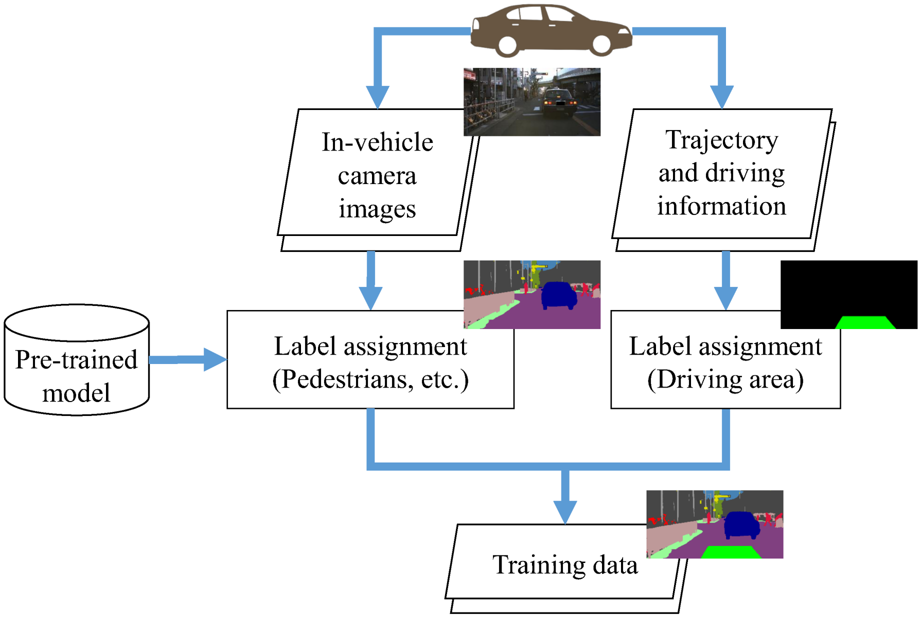

In-vehicle camera images
Trajectory and driving information
Pre-trained model
Label assignment (Pedestrians, etc.)
Label assignment (Driving area)
Training data

Fig. 3. Process flow of the automatic label assignment.

Figure 3 shows the process flow of the automatic label assignment. To output the training data automatically, the proposed automatic label assignment method receives driving trajectory, speed, and in-vehicle camera images of the own vehicle simultaneously. A state-of-the-art semantic segmentation model is used for assigning labels to the training data, such as roads, pedestrians, vehicles, etc. To assign the scene-adaptive driving area label, the actual driving trajectories are projected considering the reduction of the vehicle speed. Then, the labeled images of training data are generated by integrating the scene-adaptive driving area label and other labels. Figure 4 shows an example of an automatically generated ground-truth label image by the proposed method.

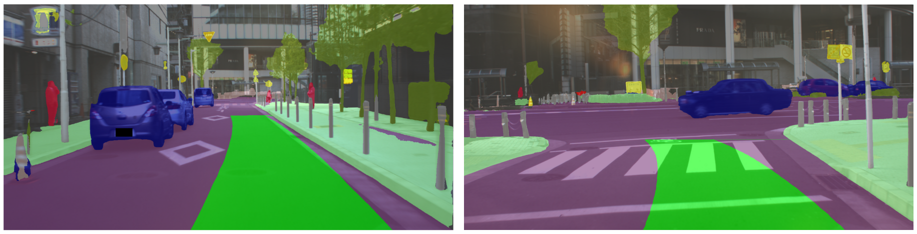


Fig. 4. Examples of scene-adaptive driving areas.

2.1 Assigning labels: Pedestrian, vehicle, roads, etc.

To train the semantic segmentation model, it is necessary to prepare images and pixel-wise annotation to each of them. Since drivers usually determine their driving path from the relationship between the own vehicle and its surrounding objects (e.g. pedestrians, vehicles, and roads), it is important to recognize those objects and scene-adaptive driving area simultaneously. Here, the labels of pedestrians, vehicles, and roads can be easily and accurately extracted by applying a state-of-the-art semantic segmentation model.

2.2 Assinging labels: Scene-adaptive driving area

The proposed method assigns the ground-truth label for scene-adaptive driving area from the own vehicle's speed and the actual trajectory of the own vehicle estimated by a LIDAR-based localization method. Here, let X_t be the vehicle center position at time $t \in T$ in the world coordinate system and $F_{t'}$ be a transformation matrix that converts the vehicle position from the world coordinate system to a coordinate system whose origin is $X_{t'}$ (t'-th coordinate system). Based on these notations, the vehicle position $\widetilde{X}_t$ in the t'-th coordinate system at time t is calculated as

$$\widetilde{X}_t = F_{t'} X_t. \tag{1}$$

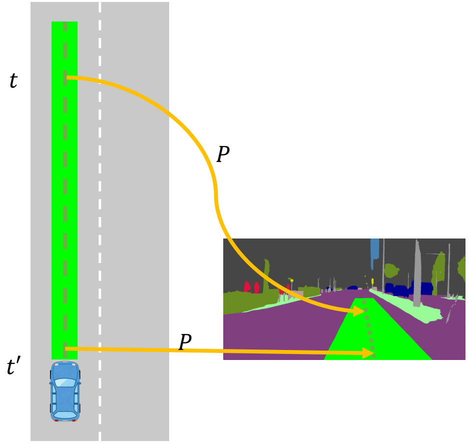


Fig. 5. Projection of vehicle positions onto an image at time t'.

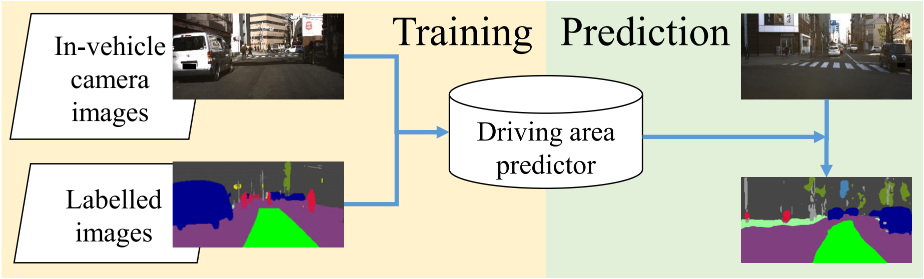


Fig. 6. Overview of the proposed scene-adaptive driving area predictor.

By iterating the above transformation until the following conditions (2) are met, a set of vehicle positions $\widetilde{\mathcal{X}}_{t'} = \{\widetilde{X}_{t'}, \widetilde{X}_{t'+1}, ...\}$ in the t'-th coordinate system is obtained.

$$\|X_t - X_{t'}\|_2 > D \quad \text{or} \quad \alpha_t \leq -2 \quad \text{or} \quad v_t = 0, \tag{2}$$

where α_t is the acceleration $[\mathrm{m/s^2}]$ at time t, and v_t is the velocity $[\mathrm{m/s}]$ at time t.

Here, since a scene-adaptive driving area far from the own vehicle cannot be observed in the image plane, the proposed method calculates the distance between the vehicle position $X_{t'}$ and each point X_t, and terminates the transformation process if the distance becomes larger than a threshold D. In addition, it is necessary to consider traffic signs and traffic signals requesting the vehicle to stop for safety. Through observation of actual driving data, we found that these situations are observed with a specific driving behavior like the reduction of the vehicle speed. From these points-of-view, the proposed method terminates the transformation when the vehicle's speed decreases.

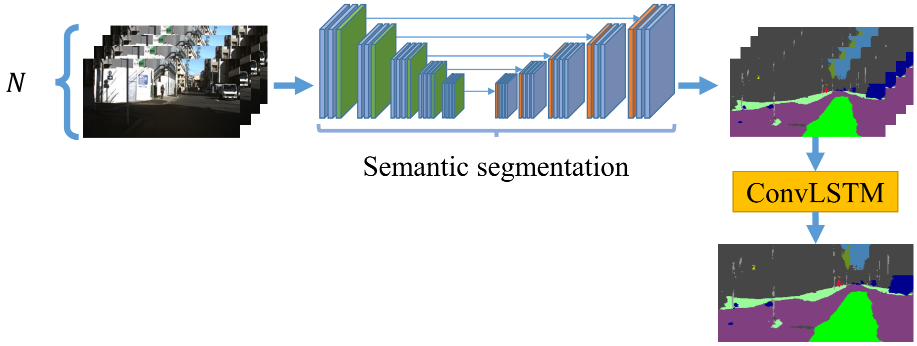


Fig. 7. Scene-adaptive driving area prediction model.

Subsequently, the vehicle trajectory is projected onto an image. Figure 5 shows the schematic diagram of the projection. The vehicle trajectory is obtained using the set of vehicle center position $\widetilde{\mathcal{X}}_{t'}$ and vehicle width. Here, the pixel position $\{x'_t, y'_t\}$ on the image is calculated as

$$\begin{bmatrix} {x'}_t \\ {y'}_t \\ 1 \end{bmatrix} = P \begin{bmatrix} \widetilde{X}_t \\ 1 \end{bmatrix}, \tag{3}$$

where P is a projection matrix. By filling with the corresponding class label between the line segments that indicate the trajectories of left and right tires, the scene-adaptive driving area is obtained.

3 Scene-adaptive driving area prediction model using semantic segmentation and ConvLSTM

We build a scene-adaptive driving area prediction model from the training data acquired by the above procedure. As shown in Fig. 6, the model is trained from pairs of an in-vehicle camera image and a generated label, and assigns label using only from the in-vehicle camera image for prediction. To predict the scene-adaptive driving area considering movement of the surrounding object, ConvLSTM is integrated into a scene-adaptive driving area prediction model as shown in Fig. 7. The proposed method employs a semantic segmentation model inspired by U-Net [5] that is based on an encoder-decoder architecture with skip connections; The encoder extracts the latent features from the in-vehicle camera images. Then, the pixel-wise label likelihoods are estimated by concatenating and restoring the features using the decoder. To obtain the movement of objects, we remove the logits layer in the segmentation model and join the end of the segmentation model to the ConvLSTM model. By applying the proposed model to N in-vehicle camera images, N label likelihoods are obtained. Here, the ConvLSTM model receives N label likelihoods. By learning the movements of

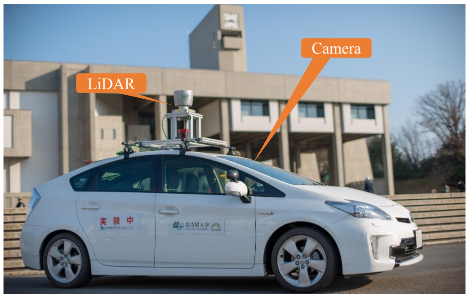


Fig. 8. Vehicle used for the experiment.

other objects from N label likelihoods, we can predict the scene-adaptive driving area accurately.

4 Experiment

We conducted an experiment for evaluating the proposed method. In this experiment, the accuracy of the semantic segmentation results was compared by changing the length of an input sequence N. We evaluated the proposed method based on Intersection over Union (IoU) that is a measure to calculate the overlap of the prediction results and their ground-truth. Here, IoU is calculated by

$$\mathrm{IoU} = \frac{\mathcal{A} \cap \mathcal{B}}{\mathcal{A} \cup \mathcal{B}}, \tag{4}$$

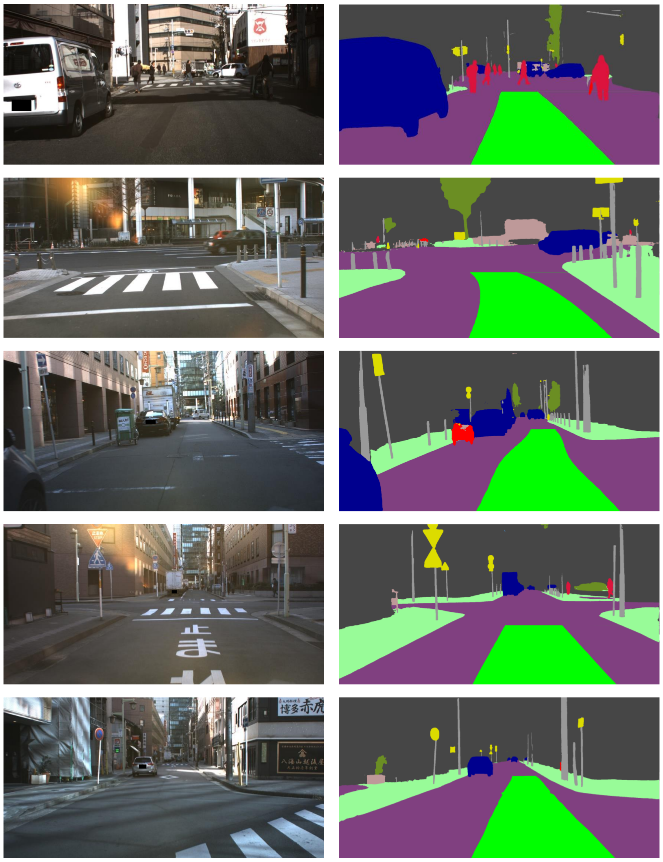
where $\mathcal{A}$ is the ground-truth constructed from the acquired dataset, and $\mathcal{B}$ is the prediction result.

4.1 Dataset

The data were collected by driving around Nagoya Station in Nagoya, Japan, with a special vehicle as shown in Fig. 8. DeepLabv3+ [2] model trained by Cityscapes dataset [3] was utilized to assign the labels other than the driving area. Here, we merged the labels of "car", "truck", "bus" in the Cityscapes dataset into a single "vehicle" label. Finally, the proposed method merged these labels with the label of the scene-adaptive driving area that is automatically acquired from driving information. Figure 9 shows an example of the constructed dataset. In the following evaluations, 4,245 data were used for training, and 1,405 data were used for evaluation.

4.2 Results and discussions

From Fig. 9, we confirmed that the proposed method could acquire training data considering stop signs and other objects.



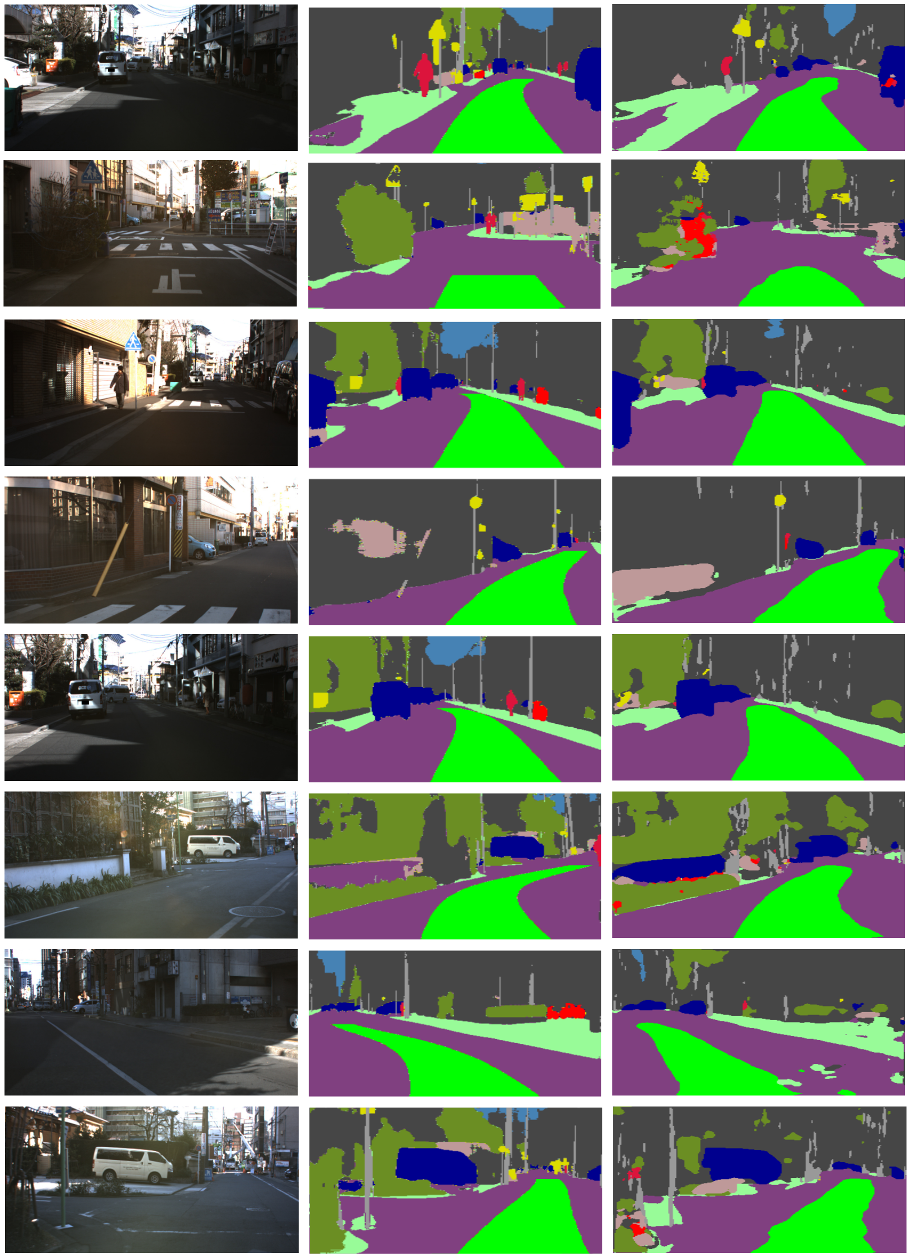
(a) In-vehicle camera images. (b) Generated labels.

Fig. 9. Examples from the dataset.

Table 1. Experiment results (IoU) of driving area prediction for each class.

Labels	$N = 1$	2	3	4	5	6	7
sky	8.5	12.1	12.8	**14.7**	**14.7**	9.9	14.4
building	76.4	**77.3**	77.2	77.1	77.1	77.0	76.7
pole	16.3	19.3	19.9	19.3	16.9	19.2	**21.6**
road	**76.0**	75.8	75.9	75.7	75.6	75.7	75.7
terrain	58.9	60.4	**60.8**	60.5	59.9	60.1	60.4
vegetation	**38.1**	36.0	34.0	33.6	34.3	32.5	35.1
traffic sign	12.5	12.6	13.2	14.2	13.9	13.4	**14.6**
fence	17.4	18.9	**19.7**	18.8	18.2	18.5	17.4
car	57.8	58.9	**59.3**	58.6	58.9	58.5	57.5
person	9.4	12.7	**13.0**	12.4	11.8	12.8	12.0
rider	8.9	8.7	8.9	8.6	7.5	8.1	**9.1**
drivable path	65.1	64.9	65.2	65.2	**65.3**	**65.3**	65.2
mIoU	37.1	38.1	**38.3**	38.2	37.9	37.6	**38.3**

Table 1 shows the IoU of each class and mean IoU, and the column of $N = 1$ indicates the prediction results by a conventional semantic segmentation model that does not use ConvLSTM. On the other hand, the $N \geq 2$ columns indicate the prediction results by the proposed scene-adaptive driving area prediction model employing ConvLSTM. From Table 1, we confirmed that the IoU of the scene-adaptive driving area and the mIoU improved by increasing the length of an input sequence N. The IoU of the scene-adaptive driving area at $N = 2$ was lower than that at $N = 1$. In the case of $N = 2$, since the frame-interval between the two input images was about 0.125 seconds, the change of object positions in the surrounding environment was small during this short interval. Therefore, we consider that objects' movements were not trained appropriately because sufficient information was not given. In addition, IoU of scene-adaptive driving area improved from $N = 2$ to $N = 6$, while the prediction result degraded by increasing N to 7. Comparing the experimental results when $N = 6$ and $N = 7$, we can see that the IoU of the person and the vehicle degraded at $N = 7$ in addition to the scene-adaptive driving area. Since the movement of the surrounding environment objects increases as the length of an input sequence increases, as a result, we consider that it could not be learned correctly. This indicates that it is necessary to investigate the optimal value of N. Figure 10 shows an example of prediction results using the proposed model. We confirmed that the proposed predictor can estimate these scene-adaptive driving area with 65.3%. From these results, we confirmed that the proposed method could acquire the labels automatically and predict highly accurate scene-adaptive driving area using the proposed semantic segmentation model incorporating ConvLSTM.



(a) In-vehicle camera images. (b) Ground-truth labelling. (c) Predicted labelling.

Fig. 10. Examples of prediction results.

5 Conclusion

In this paper, we proposed a method to acquire training labels of scene-adaptive driving area automatically from driving information of the in-vehicle camera image, own vehicle's speed, and location. Here, the training labels are acquired by using the actual trajectory of the own vehicle. According to the own vehicle speed, the proposed method generates the label of the scene-adaptive driving area reflecting the driving context. We also proposed a scene-adaptive driving area predictor based on the semantic segmentation model introducing ConvLSTM trained with the acquired training data.

To evaluate the proposed method, we created 5,650 labels and predicted the scene-adaptive driving area by the proposed method with 65.3%. We also confirmed the effectiveness of considering the movement of the surrounding object.

Future work will include an improvement of scene-adaptive driving area prediction, an enhancement of the network structure, and experiments using a larger dataset including various patterns of the scene-adaptive driving area.

Acknowledgements. Parts of this research were supported by MEXT, Grant-in-Aid for Scientific Research 17H00745, and JST-Mirai Program Grant Number JPMJMI17C6, Japan.

References

1. Barnes, D., Maddern, W., Posner, I.: Find your own way: Weakly-supervised segmentation of path proposals for urban autonomy. In: Proceedings of 2017 IEEE International Conference on Robotics and Automation. pp. 203–210 (2017)
2. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of 2018 European Conference on Computer Vision. pp. 833–851 (2018)
3. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes dataset for semantic urban scene understanding. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3213–3223 (2016)
4. Nabavi, S., Rochan, M., Wang, Y.: Future semantic segmentation with convolutional LSTM. In: Proceedings of 2018 British Machine Vision Conference. pp. 137-1–137-12 (2018)
5. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of 2015 International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241 (2015)
6. Shi, X., Chen, Z., Wang, H., Yeung, D., Wong, W., Woo, W.: Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In: Advances in Neural Information Processing Systems 28. pp. 802–810 (2015)
7. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. pp. 6230–6239 (2017)
8. Zhou, W., Worrall, S., Zyner, A., Nebot, E.M.: Automated process for incorporating drivable path into real-time semantic segmentation. In: Proceedings of 2018 IEEE International Conference on Robotics and Automation. pp. 1–6 (2018)