Semisupervised and Weakly Supervised Road Detection Based on Generative Adversarial Networks

Xiaofeng Han, Jianfeng Lu, Chunxia Zhao, Shaodi You, and Hongdong Li

***Abstract*—Road detection is a key component of autonomous driving; however, most fully supervised learning road detection methods suffer from either insufficient training data or high costs of manual annotation. To overcome these problems, we propose a semisupervised learning (SSL) road detection method based on generative adversarial networks (GANs) and a weakly supervised learning (WSL) method based on conditional GANs. Specifically, in our SSL method, the generator generates the road detection results of labeled and unlabeled images, and then they are fed into the discriminator, which assigns a label on each input to judge whether it is labeled. Additionally, in WSL method we add another network to predict road shapes of input images and use them in both generator and discriminator to constrain the learning progress. By training under these frameworks, the discriminators can guide a latent annotation process on the unlabeled data; therefore, the networks can learn better representations of road areas and leverage the feature distributions on both labeled and unlabeled data. The experiments are carried out on KITTI ROAD benchmark, and the results show our methods achieve the state-of-the-art performances.**

***Index Terms*—Generative adversarial networks, road detection, semi-supervised learning.**

I. Introduction

NOWADAYS, road detection has become one of the crucial tasks in autonomous driving. A large variety of methods have been proposed based on different kinds of sensors, such as cameras, radars, and Lidars [4]–[7]. Recently, driven by the great success of the convolutional neural networks, fully convolutional network (FCN) was proposed for image segmentation by Long *et al.* [1]. After that, the road detection results have been greatly improved by lots of FCN-based variations [8]–[11].

Most of the existing road detection methods are fully supervised, and the main drawback of them is that they need massive labeled training data. To expand the training dataset that only contains a limited number of samples, data augmentation is used in the state-of-the-art methods. However, performance can still be inadequate if there are distribution variations between training and testing datasets. For example, Fig. 1(a) presents a testing image, in which the road surface texture is rarely seen in training set. As a result, the trained model is overfitting to this small training dataset, and lots of false negatives arise in the result of a fully supervised learning method, as shown in Fig. 1(b).

Fig. 1. Different textures on road appearances between training and testing sets can lead to false segmentation. (a) Input image. (b) Result of a fully supervised segmentation network. (c) and (d) Results of our SSL and WSL methods. The predicted road areas are denoted in red.

In this letter, we propose a semisupervised learning (SSL) road detection method based on generative adversarial networks (GANs) [3] and a weakly supervised learning (WSL) method based on conditional GANs (CGANs) [13], [14]. In our methods, the unlabeled images can be directly used for training without annotating them manually. This helps the network to be more adaptive to different scenes and resist to overfit to the labeled training dataset. Fig. 1(c) and (d) show our successes in overcoming the above-mentioned problem.

There have been some SSL image classification and segmentation methods [15]–[19], and some of them are based on GANs too. However, unlike them, we do not generate fake samples from noises. In our SSL method, the generator is a semantic segmentation network to detect road on both labeled and unlabeled images. These images and their segmentation results are fed into the discriminator to predict if the input data are from labeled or unlabeled images. In WSL method, the generator and discriminator are constrained with road shapes of input images. We annotate road shapes of labeled images and train an additional network to predict road shapes of unlabeled images. This road shape prediction network is trained end-to-endly as a part of generator in CGANs, and the outputs are fed into the discriminator as well as the color images and their segmentation results.

Overall, our main contributions are as follows: 1) We propose the SSL road detection method based on GANs to train directly

Manuscript received December 11, 2017; revised February 2, 2018; accepted February 19, 2018. Date of publication February 27, 2018; date of current version March 16, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joao Paulo Paulo Papa. *(Corresponding author: Xiaofeng Han.)*

X. Han, J. Lu, and C. Zhao are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu 210096, China (e-mail: xf.han@foxmail.com; lujf@njust.edu.cn; zhaochx@njust.edu.cn).

S. You and H. Li are with the College of Engineering and Computer Science, Australian National University, Canberra, ACT 0200, Australia (e-mail: shaodi.you@anu.edu.au; hongdong.li@anu.edu.au).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2018.2809685

with unlabeled data and overcome the overfitting problem. 2) We extend our SSL method to WSL method by using the road shapes as extra constraints in CGANs.

This letter is organized as follows: The Section II introduces the details of our proposed methods. Section III reveals the experimental results, and it is followed by a section in which we conclude our letter.

II. Our Approach

In this section, we briefly introduce the background of GANs and CGANs, and then give details of our proposed methods. Generally speaking, in our methods, labeled data are used in generator to train the segmentation network and road shape prediction network. The unlabeled data are used to enhance the road feature representations by using a discriminator to distinguish the labeled and unlabeled inputs.

A. GANs and Conditional GANs

GANs [3] were proposed to train a generator model to produce realistic samples, and a discriminator model to distinguish them from real samples. Let x be a real sample, and z be a random noise. Then, we can see a GAN as a minmax game for two players with the following formulation:

$$\min_G \max_D V(D,G) = E_{x\sim p_{\mathrm{data}}(x)}[\log(D(x))] + E_{x\sim p_z(z)}[\log(1-D(G(z)))] \quad (1)$$

in which E is the empirical estimation of expected value of the probability. G and D are the generator and discriminator, respectively. For an input noise z, the generator G outputs a fake sample $G(z)$ according to the data distribution p_z. Then, the real sample x and fake sample $G(z)$ are fed into the discriminator D, and D estimates the probability that the input is drawn from real data distribution p_{data}.

The CGANs [13] were proposed based on GANs. In CGANs the optimization of parameters is directed by constraining both the generator and discriminator on additional information y. Therefore, the objective function is written as follows:

$$\min_G \max_D V(D,G) = E_{x\sim p_{\mathrm{data}}(x)}[\log(D(x|y))] + E_{x\sim p_z(z)}[\log(1-D(G(z|y)))] \quad (2)$$

B. SSL and WSL Road Detection

In our SSL road detection method, only a small fraction of labeled images are used for training, while a large number of unlabeled images are used to make the extracted features more robust in different scenes. We adopt the GANs to achieve this goal. The network architecture is shown in Fig. 2(a).

To be more specific, let us denote the labeled images as x_l, the unlabeled images as x_u, the ground truths of labeled images as y_l, the generator as G , which contains a segmentation network F_1, and the discriminator as D. The x_l and x_u are fed into F_1 to get the segmentation results $F_1(x_l)$ and $F_1(x_u)$, and then $F_1(x_l)$, x_l, $F_1(x_u)$, and x_u are convoluted into the same dimension. After that, x_l and $F_1(x_l)$ are concatenated together as the labeled input of D, while the x_u and $F_1(x_u)$ are concatenated as the unlabeled input. Here, we denote these inputs as I_l and I_u. The predictions of D are the probabilities that the inputs come from labeled images. Then, the discriminator loss L_D and the generator loss L_G are defined as follows:

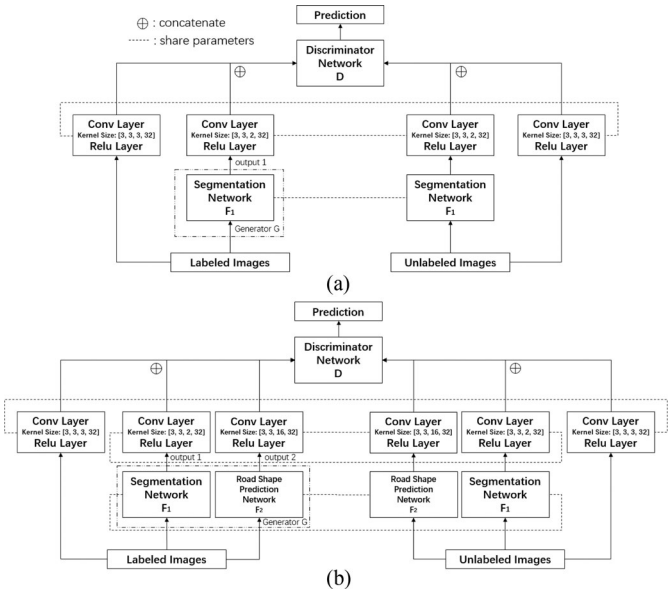

Fig. 2. Network architectures of (a) SSL road detection based on GANs; and (b) WSL road detection based on CGANs.

$$L_D = \min_D \left(\mathrm{CE}(1, D(I_l)) + \mathrm{CE}(0, D(I_u))\right) \quad (3)$$

$$L_G = \min_G \left(\mathrm{CE}(y_l, F_1(x_l)) + \alpha \times \mathrm{CE}(1, D(I_u))\right) \quad (4)$$

where CE is the cross entropy function, and α is a tradeoff parameter.

Additionally, recent works [13], [20] show that the performances of GANs can be improved by extending them to CGANs via directing the learning progress with extra information. By exploiting CGANs, our SSL method is then extended to a WSL method, in which the road shapes are used as weak supervision. We annotate the road shapes of labeled images and use them to train a network F_2 to predict the road shapes of unlabeled images. Therefore, the generator G now contains two networks F_1 and F_2. This WSL architecture is shown in Fig. 2(b). Here, the road shapes of labeled images are remarked as $y_{\mathrm{roadshape}}$. The labeled and unlabeled images are fed into G to output their segmentation results $F_1(x_l)$ and $F_1(x_u)$, as well as the road shape predictions $F_2(x_l)$ and $F_2(x_u)$. $F_2(x_l)$ and $F_2(x_u)$ are then convoluted and concatenated too into I_l and I_u. Therefore, the discriminator loss L_D remains the same, while the generator loss L_G is now defined as

$$L_G = \min_G \left(\mathrm{CE}(y_l, F_1(x_l)) + \mathrm{CE}(y_{\mathrm{roadshape}}, F_2(x_l)) + \alpha \times \mathrm{CE}(1, D(I_u))\right). \quad (5)$$

The architecture details of networks F_1, F_2, and D are given in Section III–A.

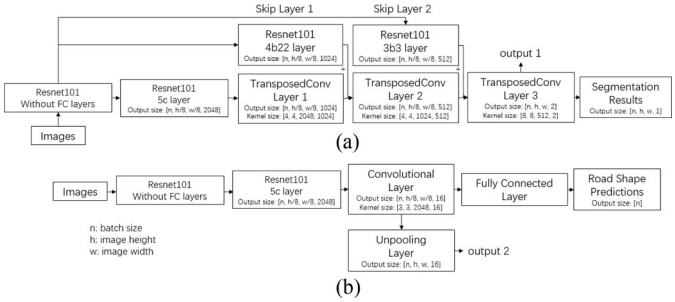

Fig. 3. Network architectures of segmentation network and road shape prediction network. (a) Segmentation network architecture. (b) Road shape prediction network architecture.

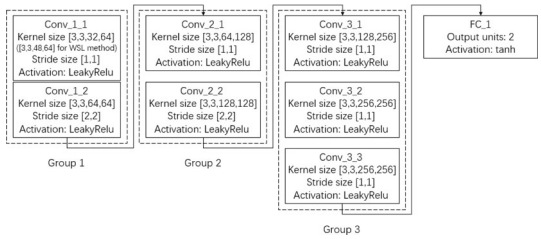

Fig. 4. Network architecture of discriminator network.

III. Experiments

A. Implement Details

Our networks are implemented with Tensorflow [26] framework and trained on an NVIDIA TITAN X GPU with 12 GB memory. The resolution of input RGB images is 375×1242 pixels. Due to the physical memory limitation on the GPU, we can only manage to feed 1 labeled image and 1 unlabeled image into the network per training iteration. The learning rates for generator and discriminator are fixed to 10e$-$5 and 10e$-$6, respectively. α is set to be 1. Training is terminated when maximum iteration number 20k is reached. In inference process, we take the outputs of segmentation network as the pixel-wise road detection results.

The details of our network architectures are given as follows.

Generator: In both SSL and WSL road detection methods, we use a pretrained Resnet101 model by DEEPLAB [22] as our backbone. In the segmentation network, we remove all the fully connected (FC) layers in Resnet101 and add three transposed convolutional layers to predict dense labels. We also add two skip layers. This network architecture is shown in Fig. 3(a). In WSL method, apart from the segmentation network, a road shape prediction network is added alongside to provide weak supervision. This network shares all the used layers of Resnet101 model with segmentation network; however three transposed convolutional layers are replaced by the FC layer to generate road shape labels. We show this network architecture in Fig. 3(b).

Discriminator: The network architectures of discriminators in both methods are shown in Fig. 4, and the only difference is the number of input feature channels in Conv_1_1 layer. As suggested in [12], max-pooling layers are replaced by strided convolutional layers to downsample feature maps for extracting features in different scales. Besides, an FC layer is added at last to assign "labeled/unlabeled" labels to inputs. Since using a bounded activation can allow the model to learn more quickly to saturate and cover the color space of the training distribution [12], the last activation function is "tanh" while all the other activation functions are "leakyRelu".

TABLE I
Performances on Different Amounts of Labeled and Unlabeled Training Data

Labeled data	Unlabeled data	Accuracy	Baseline accuracy
100	432	93.52%	87.19%
200	332	96.48%	93.66%
300	232	97.14%	95.05%
400	132	97.41%	96.22%
500	32	97.44%	97.36%
All (532)	0	–	**97.53%**

B. KITTI ROAD Dataset

To evaluate performances of our proposed road detection methods, we carry out experiments on KITTI ROAD dataset [21]. This dataset contains 289 frames training data and 290 frames testing data. We only exploit monocular color images, which are divided into three categories: UM (urban marked), UMM (urban multiple marked lanes), and UU (urban unmarked). We flip training images horizontally to expand the training dataset to 578 images.

C. SSL Road Detection

In order to demonstrate how training with unlabeled data can improve the road detection results, we train our SSL network with different amounts of labeled and unlabeled data. Specifically, we randomly select 46 images from training set as the validation images, and then divide the remaining training images into two categories as *labeled* and *unlabeled*. Furthermore, we investigate the effect ratio changes of sample numbers in these two categories have on subsequent detection performances. Besides, in each experiment a baseline segmentation network is trained with only labeled data. Table I presents the results.

As we can see from Table I, in each experiment, the result of SSL method is better than that of the baseline method, especially when only a little of labeled training data is available. Notably, the SSL method only needs 400 labeled images to achieve almost the same accuracy as that of the fully supervised learning method trained with all the labeled images.

D. WSL Road Detection

In our WSL method, we annotate all the training and validation images with three road shape labels: *branch*, *curve*, and *straight*. The number of training samples of three categories is 62, 92, and 378, respectively. The validation set consists of 8 *branch* images, 12 *curve* images, and 26 *straight* images. As for the unlabeled data, we collect images from 5 KITTI raw datasets which totally comprise 1572 unlabeled images. We also train a fully supervised baseline segmentation network and an SSL network on the same dataset. The experimental results

TABLE II
PERFORMANCES OF DIFFERENT METHODS ON VALIDATION DATASET

Method	MaxF	ACC	PRE	Recall
Segmentation method	96.90%	97.15%	96.56%	97.24%
SSL method	97.50%	97.68%	96.98%	98.03%
WSL method	97.61%	97.80%	97.31%	97.91%

TABLE III
PERFORMANCES ON DIFFERENT CATEGORIES OF KITTI ROAD BENCHMARK

Baseline road segmentation network				
Benchmark	MaxF	AP	PRE	REC
UM ROAD	92.32%	88.26%	91.79%	92.85%
UMM ROAD	94.63%	92.15%	94.59%	94.67%
UU ROAD	91.70%	87.72%	91.51%	91.89%
URBAN ROAD	93.23%	89.59%	93.11%	93.35%
Semisupervised learning road detection method				
Benchmark	**MaxF**	**AP**	**PRE**	**REC**
UM ROAD	94.62%	89.50%	95.32%	93.93%
UMM ROAD	96.72%	92.99%	97.05%	96.40%
UU ROAD	94.40%	87.84%	94.17%	94.63%
URBAN ROAD	95.53%	90.35%	95.84%	95.24%
Weakly supervised learning road detection method				
Benchmark	**MaxF**	**AP**	**PRE**	**REC**
UM ROAD	94.73%	89.22%	95.01%	94.45%
UMM ROAD	96.95%	92.87%	96.92%	96.98%
UU ROAD	94.54%	87.70%	94.01%	95.09%
URBAN ROAD	95.70%	90.17%	95.64%	95.77%

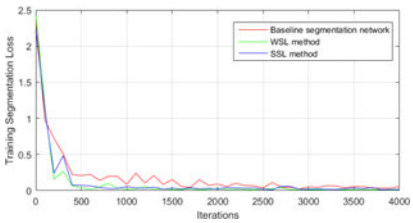

Fig. 5. Convergence times of segmentation loss. Red, blue, and green curves represent baseline segmentation network, SSL method, and WSL method, respectively.

in Table II show that SSL and WSL methods can both improve the performance effectively, and the WSL method outperforms slightly.

In addition, Fig 5 shows the segmentation loss of baseline segmentation network converges to less than 0.1 after almost 3000 iterations, while the segmentation losses of SSL and WSL methods converge to that level only after 1000 iterations. This proves that training with unlabeled images and weak supervision can accelerate the convergence of segmentation loss.

E. Road Detection Results on KITTI ROAD Benchmark

We train a baseline segmentation network, our SSL method and the WSL method on KITTI ROAD dataset. All the training images and validation images are used as labeled images. The unlabeled training images are identical as we mentioned above.

Table III shows that the results of SSL and WSL methods are much better than that of baseline segmentation network for each category.

TABLE IV
PERFORMANCES OF DIFFERENT METHODS ON KITTI ROAD BENCHMARK

method	MaxF	AP	PRE	REC
StixelNet II	94.88%	87.75%	92.97%	96.87%
MultiNet	94.88%	93.71%	94.84%	94.91%
LoDNN	94.07%	92.03%	92.81%	95.37%
DEEP-DIG	93.98%	93.65%	94.26%	93.69%
Up-Conv-Poly	93.83%	90.47%	94.00%	93.67%
DDN	93.43%	89.67%	95.09%	91.82%
Our SSL method	95.53%	90.35%	95.84%	95.24%
Our WSL method	95.70%	90.17%	95.64%	95.77%

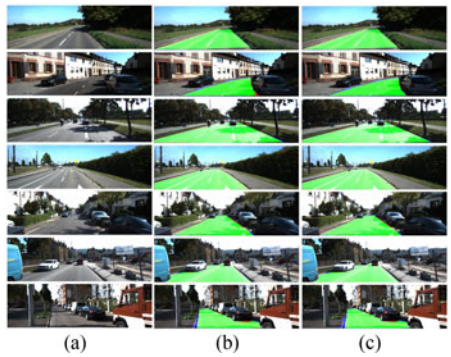

Fig. 6. Results on KITTI ROAD benchmark. (a) Input Images. (b) Results of our SSL method. (c) Results of our WSL method. Red areas denote false negatives, blue areas correspond to false positives, and green area represent true positives.

To manifest the performances of our methods, we list the results of our SSL method, our WSL method and the first six real name submitted methods in Table IV. They are StixelNet II [23], MultiNet [8], LoDNN [9], DEEP-DIG [2], Up-Conv-Poly [24], and DDN [25]. All these methods are under deep learning frameworks. The Max F1-measures of our SSL method and the WSL method are 95.53% and 95.70%, which are better than the others. Our proposed SSL and WSL road detection methods are named as "SSLGAN" and "WSLGAN" on the web page, respectively. Fig. 6 shows the results of our methods. As we can see, our methods can learn robust representation of road areas.

IV. CONCLUSION

In this letter, we propose an SSL road detection method based on GANs, and the WSL road detection method based on CGANs. In each method, we have a generator network and a discriminator network. In the SSL method, the generator network is a segmentation network, while in WSL method, it also contains a road shape prediction network. In both methods, the original input images, together with their segmentation results and road shape labels (if they are available) are convoluted into the same dimension and concatenated as the inputs of discriminator networks. The predictions given by discriminator networks show their judgments of whether the inputs come from labeled images or unlabeled images. Our methods are trained with a limited number of labeled images and large amounts of unlabeled images. Experimental results on KITTI ROAD benchmark show our methods can resist overfitting problem, accelerate the convergence speed, and achieve the state-of-the-art performances.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 2015, pp. 3431–3440.

[2] J. Muoz-Bulnes, C. Fernandez, I. Parra, D. Fernndez-Llorca, and M. Sotelo, "Deep fully convolutional networks with random data augmentation for enhanced generalization in road detection," *Proc. Workshop Deep Learn. Auton. Driving IEEE 20th Int. Conf. Intell. Transp. Syst.*, 2017.

[3] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inform. Process. Syst.*, 2014, pp. 2672–2680.

[4] K. Lu, J. Li, X. An, and H. He, "A hierarchical approach for road detection," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2014, pp. 517–522.

[5] L. Chen, J. Yang, and H. Kong, "Lidar-histogram for fast road and obstacle detection," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2017, pp. 1343–1348.

[6] L. Xiao, B. Dal, D. Liu, T. Hu, and T. Wu, "CRF based road detection with multi-sensor fusion," in *Proc. IEEE Intell. Vehicles Symp.*, 2015, pp. 192–198.

[7] A. B. Hillel, R. Lerner, D. Levi, and G. Raz, "Recent progress in road and lane detection: A survey," *Mach. Vis. Appl.*, vol. 25, no. 3, pp. 727–745, 2014.

[8] M. Teichmann, M. Weber, J. Zoellner, R. Cipolla, and R. Urtasun, "MultiNet: Real-time joint semantic reasoning for autonomous driving," arXiv:1612.07695, 2016.

[9] L. Caltagirone, S. Scheidegger, L. Svensson, and M. Wahde, "Fast lidar-based road detection using convolutional neural networks," in *Proc. IEEE Int. Vehicles Symp.* (IV), 2017, pp. 1019–1024.

[10] J. Gao, Q. Wang, and Y. Yuan, "Embedding structured contour and location prior in siamesed fully convolutional networks for road detection," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2017, pp. 219–224.

[11] A. Laddha, M. K. Kocamaz, L. E. Navarro-Serment, and M. Hebert, "Map-supervised road detection," in *Proc. IEEE Intell. Vehicles Symp.* (IV), 2016, pp. 118–123.

[12] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv:1511.06434, 2015.

[13] M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv:1411.1784, 2014.

[14] P. Luc, C. Couprie, S. Chintala, and J.Verbeek, "Semantic segmentation using adversarial networks," arXiv:1611.08408, 2016.

[15] E. Denton, S. Gross, and R. Fergus, "Semi-supervised learning with context-conditional generative adversarial networks," arXiv:1611.06430, 2016

[16] M. Koziski, L. Simon, and F. Jurie, "An adversarial regularisation for semi-supervised training of structured output neural networks," arXiv:1702.02382, 2017.

[17] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," arXiv:1610.02242, 2016.

[18] D. Pathak, P. Krahenbuhl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1796–1804.

[19] N. Souly, C. Spampinato, and M. Shah, "Semi and weakly supervised semantic segmentation using generative adversarial network," arXiv:1703.09695, 2017.

[20] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.,* 2017, pp. 5967–5976.

[21] J. Fritsch, T. Kuhnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *Proc. IEEE Conf. Intell. Transp. Syst.*, 2013, pp. 1693–1700.

[22] L.-C. Chen *et al.*, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," arXiv:1606.00915, 2016.

[23] N. Garnett *et al.*, "Real-time category-based and general obstacle detection for autonomous driving," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, 2017, pp. 198–205.

[24] G. Oliveira, W. Burgard, and T. Brox, "Efficient deep methods for monocular road segmentation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 4885–4891.

[25] R. Mohan, "Deep deconvolutional networks for scene parsing," arXiv:1411.4101, 2014.

[26] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. [Online]. Available: tensorflow.org