MULTI-TASK SELF-SUPERVISED VISUAL REPRESENTATION LEARNING FOR MONOCULAR ROAD SEGMENTATION

Jaehoon Cho, Youngjung Kim, Hyungjoo Jung, Changjae Oh, Jaesung Youn, and Kwanghoon Sohn[†]

School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
E-mail : rehoon, read12300, coolguy0220, ocj1211, jasonyoun, khsohn@yonsei.ac.kr

ABSTRACT

Training deep networks commonly follows the supervised learning paradigm, which requires large-scale semantically-labeled data. The construction of such dataset is one of the major challenges when approaching to Advanced Driver Assistance Systems (ADAS) due to the expense of human annotation. In this paper, we explore whether unsupervised stereo-based cues can be used to learn high-level semantics for monocular road detection. Specifically, we estimate drivable space and surface normals from stereo images, which are used for pseudo ground-truth to train a convolutional neural network (CNN) as a multi-task learning scheme. Combining these multiple self-supervision tasks enables CNN to jointly encode the knowledge of obstacle and ground-plane into a single frame. We demonstrate that the feature representation learned by our multi-task approach synergistically provides a rich knowledge about geometrical characteristics. Experiments on the KITTI road dataset show that our representation outperforms state-of-the-art road detection approaches.

Index Terms— Multi-task visual learning, self-supervised learning, convolutional neural network, monocular road detection.

1. INTRODUCTION

Detecting road regions from a scene has remained a core technique in the computer vision and robotics. This can be attributed to the fact that road detection is a crucial task for autonomous vehicles [1], mobile robot navigation [2], and driver assistance system [3]. Traditional approaches for road detection rely on several monocular cues, such as color [4], edge [5], and texture information [6]. However, these methods suffer from handling various road semantics under complex environments, due to manually designed constraints imposed on the prediction models. Alternatively, the road detection can be performed by using either stereo camera [7] or LiDAR [8]. Depth information provides insight into the geometric layout of a scene, which is invaluable for road detection. This process, however, requires an additional camera calibration, and is non-trivial when utilizing a single camera.

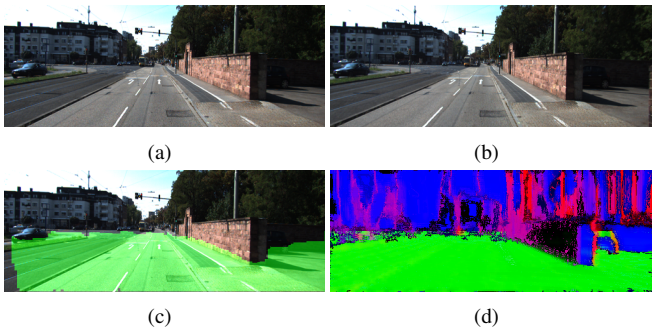


Fig. 1. (a) Left image, (b) right image, (c) pseudo drivable space from [27], and (d) pseudo surface normal obtained by [23]. We use unsupervised stereo-based cues to train CNN to predict obstacle mask and surface normal, and show that the network learns effective feature representations for road detection.

Recent progress in convolutional neural networks (CNNs) has dramatically advanced the accuracy of road detection thanks to its property of enabling a higher level of scene understanding. Brust *et al.* [9] devised the convolutional patch network with spatial prior to classify road and non-road patches. The StixelNet [10] divided an image into columns and solved the detection as a regression problem using CNNs. Mohan [11] has combined the CNNs with deconvolutional networks. These methods are suboptimal since predictions are made on a relatively small local patch only. A network thus should be run on the overlapping patches, which leads to a computational overhead. In contrast, Oliveria *et al.* [12] addressed these drawbacks by using fully convolutional networks (FCN) which can be trained in an end-to-end manner [13]. This network consists of two contrasting parts: a feature encoder, and a corresponding decoder part that expands the representation to a high-resolution output. It takes a whole image as an input and produces detection results proportional to the size of the input image. Teichmann *et al.* [14] developed a unified FCN architecture that is able to jointly reason about classification, road detection, and semantic segmentation.

This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT (NRF-2017M3C4A7069370). ([†]Corresponding author : khsohn@yonsei.ac.kr.)

978-1-5386-1737-3/18/$31.00

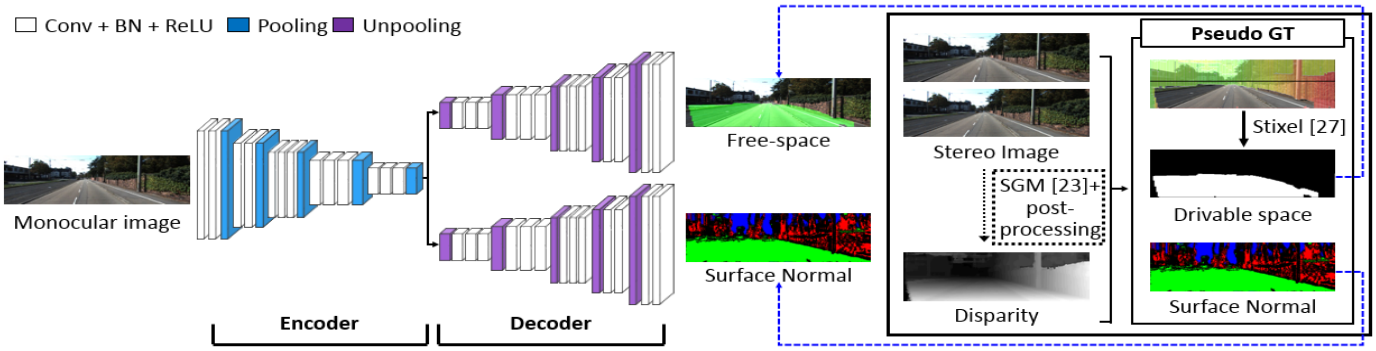


Fig. 2. The overall framework of the proposed multi-task and self-supervised representation learning. We automatically estimate drivable space and surface normals from stereo images, which are used for pseudo ground-truth to train a CNN as a multi-task learning. Combining these multiple self-supervision tasks enables CNN to jointly encode the knowledge of obstacle and ground-plane into a single frame.

A key factor for the recent success of CNNs is its ability to extract complex, high-level features from massive training data [15]. However, for the road detection, the situation becomes a bit more tricky as the existing dataset [16] consists only of hundreds of images, which is several orders of magnitude smaller than other commonly used datasets such as image classification [17] and segmentation [18]. To address this problem, the authors of [12, 14] initialized the encoder part using VGG-net [19] pretrained on ImageNet classification [17], and included dropout layers to avoid over-fitting. However, the pretrained VGG-net can be biased towards the representations from ImageNet, which is a common issue in supervised learning. It is thus unclear whether the weights from the pretrained VGG-net serves an effective initialization for road detection. We have extensively experimented and found that such strategy has limited ability to obtain a good visual representation for road detection. One of the recurring themes to address these problems in supervised learning is the use of self- (or meta-) supervision. That is, a net trained for a pretext task, which is not of direct interest, but relates the most to the final high-level vision tasks. These pretext tasks are difficult to solve without understanding image semantics, such as learning CNNs to solve colorization [20], image inpainting [21], or jigsaw puzzles [22]. CNNs pretrained with these pretext tasks have shown good performance on various vision tasks.

In this paper, we propose a framework for self-supervised learning that naturally learns high-level road semantics without an annotated dataset. We design a novel pretext task which generates drivable space and surface normals from a single image by incorporating an additional modality, i.e, stereo image (Fig. 1(a) and (b)). Specifically, our framework begins by extracting drivable space (Fig. 1(c)) and surface normal maps (Fig. 1(d)) with unsupervised stereo matching algorithm [23]. Combining these tasks in a multi-task learning objective, we then train our networks to predict these pseudo ground-truth labels from single (monocular) images. Our learning paradigm jointly encodes the knowledge of obstacle and ground-plane into a single frame, which naturally enables CNNs to focus on geometric configurations for road detection. Note that our method is closely related to the recent work of [24]. However, it differs from [24] in the sense that we design a multi-task and self-supervised learning scheme to identify high-level road semantics. We will demonstrate that, when transferred to road detection, our representation significantly outperforms previous approaches. Performance comparisons according to the amount of pseudo ground-truth used in training are also provided.

2. PROPOSED METHOD

In this section, we introduce our multi-task and self-supervised representation learning scheme in detail. The overall framework is illustrated in Fig. 2. Given a single (monocular) image, the goal of the CNN is to predict a drivable space and a surface normal map, using unsupervised stereo matching to provide pseudo ground-truth. We then transfer our network to road detection by fine-tuning on the KITTI [16] road detection benchmark.

2.1. Pseudo ground-truth generation

Given a rectified stereo image pair (I_l, I_r), we begin by extracting the disparity map using the unsupervised stereo matching method, called semi-global matching (SGM) [23]. In order to alleviate the estimation errors in SGM, we apply the joint bilateral filter [25] and left-right consistency check [26], which are purely unsupervised post-processing[1]. We use 15,000 stereo pairs in the KITTI object detection dataset [16] for generating pseudo ground-truth.

[1]Our pseudo ground-truth for representation learning is completely unsupervised, and does not use any form of supervision.

The SGM [23] often fails under adverse conditions such as textureless region and illumination. This can lead to errors in pseudo ground-truth (drivable space and surface normal). Nevertheless, we find that our representation is resilient to these kinds of degradations (see Sections 2.2 and 3.1.).

2.1.1. *Obstacle mask*

We adopt the Stixel World algorithm [27] to generate pseudo labels for drivable space. From an RGB image and its corresponding disparity map, the Stixel World [27] simplifies a scene into piecewise planar segments, i.e., rectangular sticks. These mid-level representations inherently considers common geometric layouts in road scenes. The MAP estimation is then performed over columns with disparity maps to extract the drivable space (Fig. 1(c)).

2.1.2. *Surface normal*

Given the baseline distance and the camera focal length, we convert the SGM disparity to depth and compute the surface normal using the same method as in Siberman *et al.* [28]. Concretely, the method [28] first projects the depth points from the image plane to the 3D world coordinates. The normal map is then estimated by fitting the least-squares planes to neighboring sets of pixels in the point cloud (see Fig. 1(d)).

2.2. Learning multi-task and self-supervised CNN

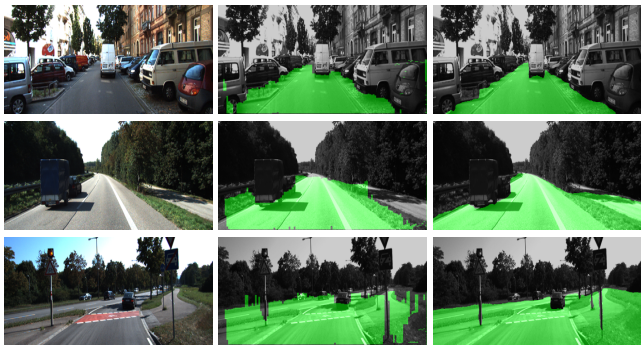
After generating the pseudo ground-truth, we use the I_l as an input to train the multi-task CNN as shown in Fig. 2. The basic architecture is similar to [12], except that we have two decoders to simultaneously predict both drivable space and surface normal. The encoder consists of the repeated application of two or three 3×3 convolutions and element-wise rectified linear unit (ReLU), followed by 2×2 max-pooling. Note that the parameters of the encoder are shared across two tasks as shown in Fig. 2. Since the output of the encoder has low-resolution feature representations, we add the decoder networks to derive the high-resolution output. The decoder progressively enlarges the spatial resolution of feature representations through a sequence of bilinear interpolation. We use softmax and L_1 loss functions for drivable space and surface normal, respectively. Joint training is performed by merging the gradients computed by each loss (with equal weights).

As stated earlier, the generated pseudo ground-truth is often noisy and inaccurate since it is from unsupervised algorithms [23], [27]. This raises a natural question: Can the CNN learn from an imperfect ground-truth? To answer this question, we measure the accuracy[2] of drivable spaces predicted by our pretrained CNN. As shown in Table 1 and Fig. 3, our prediction is surprisingly better than the Stixel World [27] which was used to generate pseudo ground-truth. This result demonstrates that CNN can capture high-level semantics for driving sequences, although it is learned from noisy ground-truth. Furthermore, our multi-task learning scheme improves the performance both for maximum F1-measurement (Fmax) and average precision (AP).

[2]For this measure, we manually segment the drivable spaces of 200 test images in the KITTI road benchmark [16].

Table 1. Results for the drivable space estimation

Method	**Fmax**	**AP**
Stixel World [10]	86.75	85.36
Ours (drivable space only)	95.63	94.27
Ours (multi-task)	**96.15**	**95.37**



(a) (b) (c)

Fig. 3. (a) Input monocular image, (b) the output of Stixel World [27] that was used to train our multi-task CNN, and (c) the output of our multi-task CNN. The generated pseudo ground-truth is often noisy. However, our network can still be learned from this noisy data, and produce better drivable space estimation.

2.3. Transfer to road detection

In this section, we fine-tune our pre-trained model on a smaller and, manually-annotated labels. We use [12] as our baseline architecture and copy the weights partly from our pre-trained model. The weights of encoder part are copied from our pre-trained model, while rest of the layers (decoder part) are randomly initialized using a uniform distribution in the range $[-0.1,0.1]$. The softmax loss function is used for fine-tuning the network.

In this stage, noting that only a single decoder is needed, the number of parameters is equivalent to [12]. During fine-tuning, data augmentation is performed on the fly. We flip the input images horizontally with a 50% chance, and randomly scale the image by a factor between $[0.7, 1.4]$. Color augmentations are also implemented, by adding a value between -0.1 and 0.1 to the hue channel of the HSV representation.

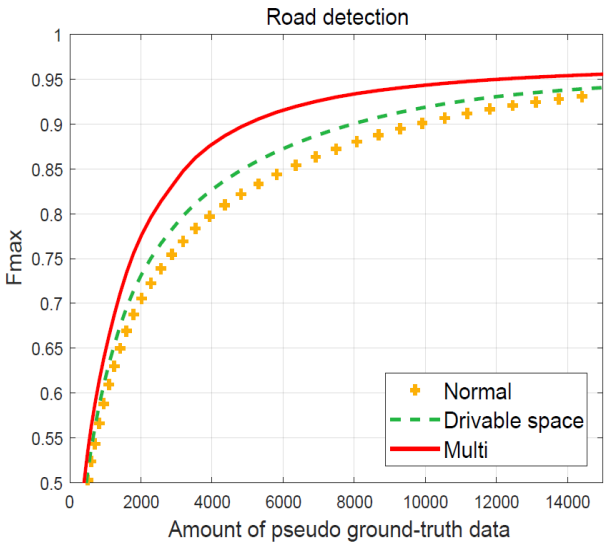


Fig. 4. KITTI [16] road detection performance using our representations. Our representation quality increases logarithmically with the amount of data, and our multi-task learning scheme synergistically improves the performance of road detection.

2.4. Implementation details

We use `VLFeat MatConvNet library` [29] with 12 GB NVIDIA Titan. The Adam solver [30] is adopted for an efficient stochastic optimization. For the pre-training phase, i.e., learning the multi-task networks with designed pretext task, we trained our multi-task networks with the 15,000 training set which consists of monocular images with their corresponding pseudo ground-truth labels. The multi-task network was trained on 60 epochs with batch size 2. In the transfer learning, we used the KITTI road benchmark [16] which contains 289/290 training/test images with manually-annotated labels. The mini-batch size in this task is set to 4. The source code will be publicly available later.

3. EXPERIMENTAL RESULTS

In this section, we validate the effectiveness of our pre-trained model by fine-tuning our network to KITTI road detection [16]. We first analyze the pretext task through several ablations and varying the amount of pseudo ground-truth data. We then compare the performance with other state-of-the-art methods.

3.1. Analysis of the learned representation

We vary the amount of pseudo ground-truth available for training, and evaluate the learned representation on road detection. The results are shown in Fig. 4. The quality of our representation grows logarithmically with the number of pseudo ground-truths, suggesting that good representations will need large amounts of data. Training for our pretext task leads to strong features even with imprecise pseudo ground-truth. Furthermore, the learned representation from multi-task approach outperforms other single tasks, demonstrating that it synergistically improves the performance of road detection. Using our multi-task self-supervised representation, only few hundreds of annotated road images are needed to achieve superior performance.

Table 2. Quantitative comparison for road detection.

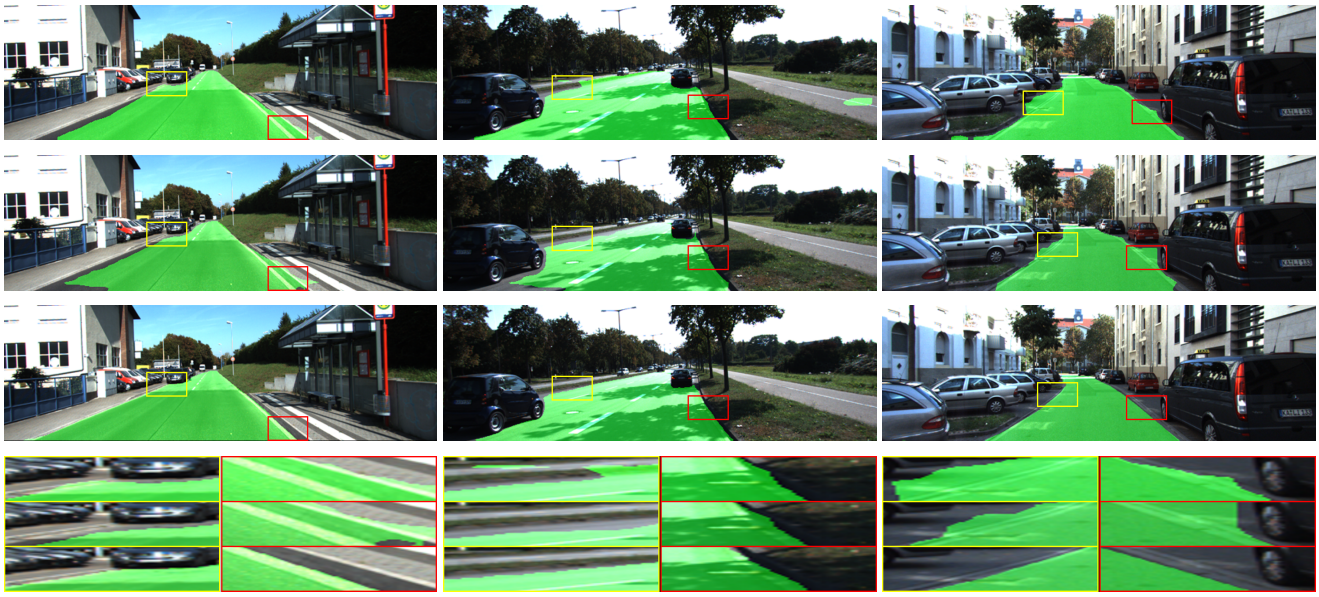
Benchmark	Fmax	AP	PRE
StixelNet [10]	89.12	81.23	85.80
FNC-LC [31]	90.79	85.83	90.87
DDN [11]	93.43	89.67	**95.09**
Oliveria *et al.* [12]	93.83	90.47	94.05
SPL [24]	94.40	93.05	93.87
Teichmann *et al.* [14]	94.88	<u>93.71</u>	94.84
Ours (surface normal)	94.11	93.14	93.01
Ours (drivable space)	<u>94.93</u>	93.25	94.13
Ours (multi-task)	**95.32**	**94.23**	<u>95.07</u>

3.2. Comparison with the state-of-art methods

We validate the performance of the proposed method against several state-of-the-art methods, including StixelNet [10], DDN [11], FNC-LC [31], Oliveria *et al.* [12], Teichmann *et al.* [14], and SPL [24]. All methods including ours use the CNNs for road detection. However, the first two methods rely on patch training, and the SPL [24] adopts the self-paced learning technique. The remaining methods initialize their networks using VGG-net [19] pretrained on ImageNet classification [17]. The results for the comparison are obtained from source codes provided by the authors, or are taken from their project websites. We do not include any post-processing for a fair comparison.

The quantitative results from 290 test images are reported in Table 2. The first- and second-ranked scores for each metric are highlighted in bold and underlined, respectively. This table shows that the proposed method outperforms other state-of-the-art methods. Although our network architecture is similar to [12], our representation achieves significant performance improvements in all metrics. This result implies that the pretrained VGG-net [19] can be biased towards the representations from ImageNet dataset [15].

Figure 5 shows a visual comparison for the three different categories: single lane road with markings (UM), single-lane road without markings (UU), and multi-lane road with markings (UMM). It shows that the proposed method is effective at segmenting roads, and successfully distinguishes road and sidewalk.



(a) UM (b) UMM (c) UU

Fig. 5. Road detection results on different road categories [16]: (From top to bottom) Oliveira *et al.* [12], Teichmann *et al.* [14], ours (multi-task), and snippets. The proposed method is effective at detecting roads, and successfully distinguishes road and sidewalk.

4. CONCLUSION

We have presented a multi-task self-supervised learning of visual representations for road detection, and have shown that it can effectively encode high degree of road semantics without human annotations. We have employed the idea of pretext task which learns the CNNs to predict the drivable space and surface normals from a single image. The proposed method enables CNNs to imply potential road regions during self-supervised learning. Experiments on the KITTI road dataset has shown that our representation outperforms state-of-the-art road detection approaches. In the future, we plan to investigate the potential gain of incorporating other sensors, such as near- and far-infrared images.

5. REFERENCES

[1] A. Lookingbill, J. Rogers, D. Lieb, J. Curry, and S. Thrun, "Reverse optical flow for self-supervised adaptive autonomous robot navigation," *Int. J. Comput. Vis.*, vol.74, no.3, pp.287-302, Jan. 2007.

[2] C. Siagian, CK. Chang, and L. Itti, "Mobile robot navigation system in outdoor pedestrian environment using vision-based road recognition," in *Proc. IEEE Int. Conf. Robot. Autom.*, pp.510-517, 2015.

[3] YC. Kuo, NS. Pai, and YF. Li, "Vision-based vehicle detection for a driver assistance system," in *Comput. Math. Appl.*, vol.61, no.8, pp.2096-2100, Apr. 2011.

[4] J.M. Alvarez and A.M. Lopez, "Road Detection Based on Illuminant Invariance," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 184-193, Mar. 2010.

[5] Y. He, H. Wang, and B. Zhang, "Color-based road detection in urban traffic scenes," in *IEEE Trans. Intell. Transp. Syst.*,vol. 5, no. 4, pp. 309-318, Dec. 2004.

[6] J. Zhang and HH. Nagel, "Texture-based segmentation of road images," in *Proc. IEEE Intell. Vehicle Symp.*, 1994.

[7] C. Oh, B. Kim, and K. Sohn, "Automatic illumination Invariant Road Detection with Stereo Vision," in *IEEE Conf. Ind. Electron. Appl.*, Jan. 2012.

[8] R. Femandes, C. Premebida, and P. Peixoto, "Road Detection Using High Resolution LIDAR," in *IEEE Vehicle Power and Propulsion Conference.*, 2014.

[9] CA. Brust, S. Sickert, M. Simon, and E. Rodner, "Convolutional Patch Networks with Spatial Prior for Road Detection and Urban Scene Understanding," in *Int. Conf. on Comput. Vis. Theory and Appl.*, pp.510-517, 2015.

[10] D. Levi, N. Garnett, and E. Fetaya, "StixelNet: A Deep Convolutional Network for Obstacle Detection and Road Segmentation," in *British Mach. Vis. Conf.*, 2015.

[11] R. Mohan, "Deep Deconvolutional Networks for Scene Parsing," in *arXiv*, 2014.

[12] GL. Oliveira, W. Burgard, and T. Brox, "Efficient Deep Methods for Monocular Road Segmentation," in *IEEE/RSJ Int. Conf. on Intell. Robot. and Syst.*, 2016.

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

[14] M. Teichmann, M. Weber, M. Zoellner, and R. Cipolla. "MultiNet: Real-time Joint Semantic Reasoning for Autonomous Driving," in *CoRR*, 2016.

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211-252, Mar. 2015.

[16] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012.

[17] A. Krizhevsky, I. Sutskever, and GE. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Proc. Syst.*, 2012.

[18] TY. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and CL. Zitnick, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *arXiv*,2014.

[20] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *Proc. Eur. Conf. Comput. Vis.*, 2016.

[21] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and AA. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.

[22] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.*, 2016.

[23] H. Hirschmuller, "Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005.

[24] A. Kendall, H. Martirosyan, S. Dasgupta, and P. Henry, "Self-paced cross-modality transfer learning for efficient road segmentation," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2017.

[25] J. Kopf, MF. Cohen, and D. Lischinski, "Joint bilateral upsampling," *ACM Trans. Graph*, vol. 26, no. 3, pp. 96, Mar. 2010.

[26] S. Joung, S. Kim, B. Ham, and K. sohn, " Unsupervised Stereo Matching Using Correspondence Consistency ," in *Proc. IEEE Int. Conf. Image Process.*, 2017.

[27] D. Hernandez-Juarez and A. Espinosa, "GPU-accelerated real-time stixel computation," in *IEEE Winter Conf. Appl. Comput. Vis.*, 2017.

[28] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, " Indoor segmentation and support inference from rgbd images," in *Proc. Eur. Conf. Comput. Vis.*, pp. 746-760, 2012.

[29] http://www.vlfeat.org/matconvnet.

[30] D. Kingma and J. Ba, " Adam: A method for stochastic optimization," in *arXiv*, 2014.

[31] C. Mendes, V. Fremont, and D. FernandoWolf, "Exploiting fully convolutional neural networks for fast road detection," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2016.