# Multimodal Neural Networks: RGB-D for Semantic Segmentation and Object Detection

Lukas Schneider[1,2(✉)], Manuel Jasch[3], Björn Fröhlich[1], Thomas Weber[3], Uwe Franke[1], Marc Pollefeys[2,4], and Matthias Rätsch[3]

[1] Daimler AG, Stuttgart, Germany
lukas.schneider@daimler.com
[2] ETH Zurich, Zurich, Switzerland
[3] Reutlingen University, Reutlingen, Germany
[4] Microsoft Corporation, Seattle, USA

**Abstract.** This paper presents a novel multi-modal CNN architecture that exploits complementary input cues in addition to sole color information. The joint model implements a mid-level fusion that allows the network to exploit cross-modal interdependencies already on a medium feature-level. The benefit of the presented architecture is shown for the RGB-D image understanding task. So far, state-of-the-art RGB-D CNNs have used network weights trained on color data. In contrast, a superior initialization scheme is proposed to pre-train the depth branch of the multi-modal CNN independently. In an end-to-end training the network parameters are optimized jointly using the challenging Cityscapes dataset. In thorough experiments, the effectiveness of the proposed model is shown. Both, the RGB GoogLeNet and further RGB-D baselines are outperformed with a significant margin on two different task: semantic segmentation and object detection. For the latter, this paper shows how to extract object-level groundtruth from the instance level annotations in Cityscapes in order to train a powerful object detector.

## 1 Introduction

Semantic interpretation of image content is one of the most fundamental problems in computer vision and is of highest importance in various applications. The availability of extremely large datasets has pushed the development of strongly data-driven machine learning methods. In particular, convolutional neural networks (CNNs) have pushed the state of the art in image understanding in various different tasks and applications. Simultaneously, the costs for cameras with increasing resolution have decreased substantially in the last years. We expect this trend to continue and thus focus on methods that can deal with such high resolution images. At the same time, we are interested in efficient methods that can meet high real-time requirements as in e.g. robotics or autonomous driving. Naturally, the main focus in the computer vision community has been in the interpretation of color images which neglects the availability of complementary


P. Sharma and F.M. Bianchi (Eds.): SCIA 2017, Part I, LNCS 10269, pp. 98–109, 2017.
DOI: 10.1007/978-3-319-59126-1_9

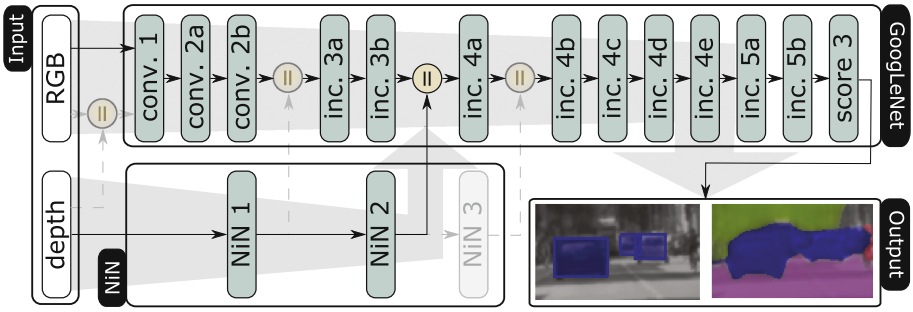



**Fig. 1.** Structure of our mid-level fusion approach. A GoogLeNet is used on top of the RGB input image and a NiN for the depth information. Alternative processing paths to fuse both networks are shown grayed.

inputs from other domains, e.g. depth, infrared, or motion. In this work we focus on depth data as additional input to CNNs. However, the presented approach is easily adaptable to other modalities.

Only using state-of-the-art CNN approaches for the multi-modal data is not optimal, since huge datasets such as ImageNet [33], MS COCO [28] or places [40] only provide color images and do thus not allow the training of large multi modal CNNs. Two main different approaches have emerged to deal with this problem. Either, only the small amount of data is used for training while accepting the resulting degraded performance. Or existing RGB networks are simply applied to the new domain and fused with those fork responsible for the color domain.

This paper proposes a novel network architecture, c.f. Fig. 1, that implements a mid-level fusion of features from the individual input domains. This combines both advantages of the previous approaches: first, the network can exploit highly complex intra-domain dependencies through the joint feature processing in order to maximize the semantic accuracy of the network Secondly, it allows the reuse of the existing initialization on large datasets. Furthermore, we demonstrate that using a network designed and trained for color inputs is suboptimal in the depth domain and propose a superior adapted architecture together with an initialization scheme yielding significant improvements in terms of semantic accuracy. The experiments show that filters learned on depth data with this approach differ substantially to those obtained by a training on RGB data.

Overall, this paper presents a simple yet effective novel network architecture together with an initialization scheme that exploits depth data in addition to sole color information. This approach leads to a significant improvement on two different common tasks in computer vision: semantic segmentation and object detection. It is based on an standard state-of-the-art network architecture and is easily adaptable to different modalities as well as tasks.

## 2 Related Work

The vast amount of relevant literature can be split into three different groups. The first comprises methods that use CNNs for semantic segmentation and can be further split into methods using an additional graphical model [2,3] and those methods without [1,8,19,39]. Due to their computational efficiency we opt for a purely CNN-based approach. Using an ensemble of CNNs [18,19,38] can lead to significant performance gains, however, with the cost of high computational burden. The scope of this paper is to show how to benefit of multi-modal data. To this end, we don't use ensembles and restrict ourselves to standard network architectures and training schemes.

The second line of work is formed by CNN's for object detection. Current literature differs between two basic approaches. First, methods such as RCNN [14], Fast RCNN [13], or R-FCN [23] require a previous hypotheses generation step and finally classify each hypothesis. As another group of methods there are e.g. Overfeat [34], YOLO [32] or SSD [29] without extra hypotheses generation. For a more detailed overview and comparison of multiple state of the art object detection methods we refer to [22]. Due to the excellent trade off between computational time and performance, we focus in this work on SSD.

Finally, we identify those methods basing on CNNs that exploit depth data, as the most related line of work. Apparently, most works use additional input features such as height, depth or angle of gradient [4,15,21,26]. Instead, we simply rely on inverse depth as input in addition to the color image. Some methods use graphical models to increase the semantic accuracy with the cost of more computational demands [9,24]. The depth input has been used to select the scale in a scale pyramid [25]. This way, however, no depth features such as depth discontinuities can be exploited. This method serves as baseline in our experimental section, c.f. Section 4. A further distinction between methods in this group is the level of fusion. Fusing color and depth data at an early level, i.e. concatenating the inputs directly, has been studied by [4,7,37]. But [4] report better results with a late fusion. We address this observation to the little availability of labeled multi-modal data. Most existing works, on the other hand, opt for a late fusion, i.e. separate network streams for depth and color data. Either a classifier is applied on the independently trained networks [4,15,16] or the networks are fused in one of the last layers and a joint training is carried out [10,17,21,26]. In the spirit of end-to-end learning, we also perform joint training. In contrast to these methods, this work shows the benefit of a mid-level fusion of learned features from the depth and color domain.

Most related to this paper is recent work that also implements a mid-level fusion with RGB-D data [17]. However, some significant differences exist: first, they use a decoder architecture with unpooling layers based on the SegNet architecture [1]. Due to the poor reported results on the Cityscapes dataset [5], we opt for a learned transpose convolutional (also referred to as deconvolutional) decoder instead. It seems that the SegNet architecture suffers from high-resolution input images. Also [17] use a small resolution of $224 \times 224$ px as input although the dataset Sun RGBD [36] provides varying resolutions around

$640 \times 480$ px. This paper focuses on high resolution images - we use input resolutions up to $2048 \times 1024$ px during both: training and testing. Finally, an initialization of the depth branch with ImageNet pre-trained weights is needed in that work. In this paper, we show that this is non-optimal: the parameters trained for depth data lead to different filters in the CNN and to superior results. We hope that this paper will stimulate more works that exploit depth data on the challenging high-resolution dataset Cityscapes [6].

We consider the following as main contributions of this paper: (1) A novel generic mid-level fusion network architecture is proposed together with an experimentally grounded initialization scheme for the additional modality. This network is simple yet effective and can be easily adapted to different modalities and tasks. (2) In thorough experiments on the Cityscapes dataset, the effectiveness of the proposed approach as well as the influence of the important design choices is demonstrated. Both, the RGB as well as an RGB-D baseline are outperformed with a significant margin on two different challenging tasks: semantic segmentation and object detection. (3) Finally, we show how to use the pixel-level annotations of the Cityscapes dataset to train an powerful neural network for object detection. To this end, the well-known SSD approach is adapted to the GoogLeNet and extended to the proposed multi-modal CNN.

## 3 Method

In this work, we propose a novel deep neural network architecture that can exploit other modalities such as depth images in addition to sole color information. Since in many cases no large datasets like ImageNet exist for the new modalities, c.f. Sect. 2, simply using an existing state-of-the-art CNN architecture and performing a training for multi-modal data is unfortunately not possible. Instead, we adapt the frequently used GoogLeNet [38] and fuse it with a network branch optimized for depth data. Note that the modifications described in this work are easily adaptable to other modalities, e.g. optical flow or infrared, as well as other network architectures, e.g. Network-in-Network (NiN) [27], VGG [35] or ResNet [18].

**Depth Network.** For the depth branch, we train and adapt a NiN [27] variant for sole depth data and use the large semi-supervised part of the Cityscapes dataset [6] for initialization. A NiN consists of multiple modules, each being further composed of one convolutional layer with a kernel size larger than one that captures spatial information and multiple $1 \times 1$ convolutional kernels. Such a module is equivalent to a multi-layer perceptron (MLP). For classification, a global average pooling layer yields one score per class. We follow [31] and discard the global average pooling resulting in a FCN [30] that predicts one score per pixel and class.

We argue that depth data requires filters that differ significantly from those obtained via training on RGB data. For instance, we expect edge and blob filters to be wider in order to be robust to the noisy depth estimates. For this reason

we use random initialization for training and reduce the channel count in each layer to $\frac{1}{3}$ considering that depth yields only one instead of the three color input channels.

**RGBD Network.** The GoogLeNet consists of a first part with convolutional and max pooling layers that quickly reduce the spatial resolution. This part is followed by nine inception modules including further pooling layers each halving the spatial dimension as illustrated in Fig. 1. We identify different points for joining the depth and RGB network. First, the RGB and depth input can be concatenated directly resulting in a new first convolutional layer. We will refer to this model as *early fusion*. Second, the scores of the RGB network and depth branch can be concatenated at the end, followed by a $1 \times 1$ convolutional as classifier. We will refer to this as *late fusion*. Finally, scores of the depth branch can be merged in the RGB network before one of it's max-pooling layer, again followed by a $1 \times 1$ convolutional layer. The number of NiN modules used in this mid-level fusion approach is determined by the required spatial dimensional in the RGB network. Thus, we call these models according to the number of NiN modules, e.g. *NiN 1*.

In theory, a multi-modal CNN with an early fusion as described above can develop independent network streams by learning features that only take one input modality into account. Thus, an early fusion is generally more expressive then a mid-level fusion, it can exploit correlations between the modalities already on a low-level of CNN computation. However, the higher expressivity comes with the price that larger amounts of data might be required for training. The benefit of a later fusion is that most of the network initialization can be reused directly without the necessity to adapt the network weights to the additional input cue. Unfortunately, it does not allow the network to learn this high-level interdependencies between the individual input modalities, since only the resulting scores on classification level are fused.

## 4 Experiments

We evaluate our proposed model on two different tasks: semantic segmentation, the task to assign a semantic label to each pixel in an image c.f. Sect. 4.2, and object detection, c.f. Sect. 4.3. The initialization of the depth network branch is described and evaluated in Sect. 4.1.

**Dataset.** Throughout our experiments, we use the Cityscapes dataset that provides a high number of pixel-level semantic annotations with 19 classes, e.g. person, car, road etc., in challenging inner city traffic scenarios. In addition to this fine annotations, 20 000 coarsely annotated images are provided. The coarse labels are more quickly and thus more cheaply annotated images where objects are labeled via polygons. Although this way many pixels remain unlabeled, each annotated pixel is defined to be correct.

### 4.1 Depth Network

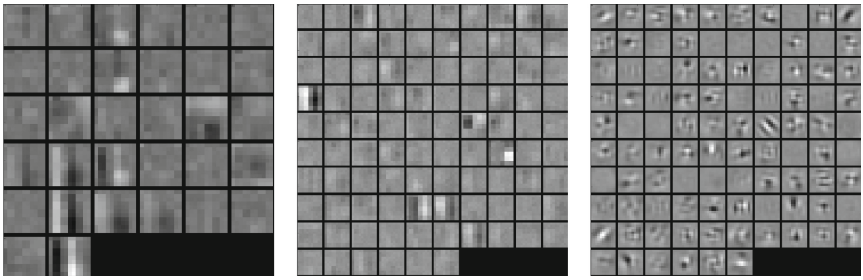
**Evaluation and Training Details.** We use the 20 000 coarsely annotated images to train a NiN for scene labeling c.f. Sect. 3, consisting of three NiN modules with two $1 \times 1$ convolutional layers each. We follow [30] and add two skip layers in order to exploit low-level image features for the expanding part of the network. We opt for a batch size of ten and use random crops during training to account the GPU memory limitations. As depth input, the publicly available stereo data obtained via semi-global matching [20] is used. More precisely, we follow [31] and encode depth as disparities, i.e. inverse depth. Missing measurements are encoded as $-1$, the mean value is subtracted. After this initialization phase, the network is fine-tuned on the 2975 finely annotated training image of Cityscapes. For evaluation, we use the 500 validation images. As evaluation metric, we use Intersection-over-Union (IoU) [6] defined as $IoU = \frac{TP}{TP+FP+FN}$, where TP, FP, and FN are the numbers of true positive, false positive and false negative pixels determined over the whole dataset.

The 19 Cityscapes classes are grouped to seven categories: flat, construction, nature, vehicle, sky, object, and human. In addition to the IoU on the 19 classes ($IoU\ class$), we measure the performance in terms of IoU on this seven categories.

**Table 1.** Impact of the initialization scheme for the disparity network on the semantic segmentation performance. The upper part of the table shows results with random weight initialization. The high amount of channels in the original NiN prevents a successful training. The influence of different initialization schemes is shown in the lower part. The proposed initialization on the coarsely annotated images yields significantly improved results compared to a variant initialized on ImageNet [33].

| | Initialization | IoU class [%] | IoU category[%] |
|---|---|---|---|
| Original NiN [27] | Random | $< 5.0$ | $< 10.0$ |
| NiN (ours) | Random | 30.5 | 61.3 |
| Original NiN [27] | ImageNet | 35.0 | 64.0 |
| Original NiN [27] | Cityscapes coarse | $< 5.0$ | $< 10.0$ |
| **NiN (ours)** | Cityscapes coarse | **37.3** | **66.5** |

**Initialization Method.** In Sect. 3, we argued that a CNN for depth data should significantly differ from a CNN on RGB data. First, we train a CNN solely on depth data discarding the available RGB input. The results of this proposed model in comparison to the original NiN are given in Table 1. The first observation from the upper half of the table is that we were not able to train the original NiN on the cityscapes dataset only. The proposed variant with $\frac{1}{3}$ of the channels, however, yields surprisingly good results. Second, initialization with the weights trained on the RGB ImageNet data guides the learning process and yields an improvement of 4.5%. Nevertheless, an initialization on actual depth data leads to significant improvements. Overall, the number of parameters of



(a) Our model (b) ImageNet finetuned (c) ImageNet initialization

**Fig. 2.** Filters of the first convolutional layer in the NiN architecture. The filters initialized randomly and trained on Cityscapes coarse labels (left) differ significantly of those trained on color data (right). Mainly gradient, mean and blob filters are developed during training. Fine-tuning of the color filters on depth data (middle) yields smaller amount of sharp filters.

the network was reduced to $\frac{1}{3}$ leading to $\frac{1}{3}$ of the computational costs. On the other hand the results were improved significantly. The resulting filters in the first convolutional layer differ substantially between the depth and color input respectively, c.f. Fig. 2. Observably, the amount of meaningful filters is higher in our model which we address to the reduced number of filters in the network.

### 4.2 RGBD Semantic Segmentation

**Evaluation and Training Details.** We leverage the 5000 finely annotated images, i.e. 2975 for training, the 500 validation images for testing, of the Cityscapes dataset and IoU as evaluation metric as before, c.f. Sect. 4.1. We do not use the images in the validation set for training. A batch-size of two and the maximal learning rate were used. After convergence, we decrease the learning rate step by step with a factor of $\frac{1}{10}$ until no further improvements on the validation set are observed. For each method, we report the best results according to the IoU on the validation set. The results of best model according to the following experiments is submitted to the Cityscapes benchmark server for evaluation on the remaining ∼1500 test images.

**Level of Fusion.** First, the optimal level for fusing the color and depth branch is determined. To this end, we train and evaluate the early-fusion, late-fusion and all five mid-level fusion models NiN-1 to NiN-7 and compare it to the RGB baseline. The results in Fig. 3 first show that the additional depth input helps in all fusion variants significantly. The RGB baseline achieves 63.9% IoU (class wise) compared to the 69.1% of the NiN-2 model. This is a considerable relative improvement of about 10%. Furthermore, it becomes apparent that a mid-level fusion after 2 NiN modules leads to the best results, the most frequently used late fusion only yields 67.1% IoU. Surprisingly, the NiN-1 and NiN-7 variants perform worst. The feature concatenation in the NiN-1 model takes place directly

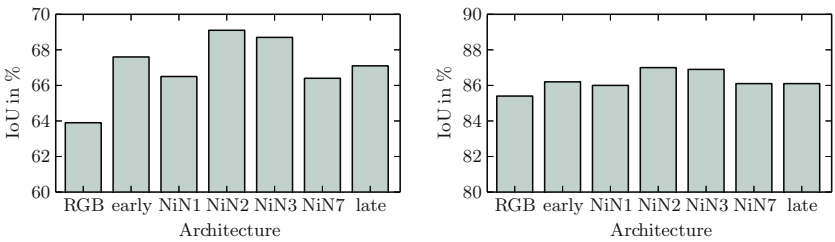
**Table 2.** Comparison to baselines on the Cityscapes dataset [5]. Both: the RGB baseline as well as an external RGB-D baseline (with and without additional CRF) are outperformed by the proposed model without the need for a CRF in terms of semantic accuracy on all 19 classes respectively the seven categories.

| Method | Input | IoU class [%] | IoU category[%] |
|---|---|---|---|
| [25] w/o CRF | RGB-D | 62.5 | N/A |
| [25] with CRF | RGB-D | 66.3 | 85.0 |
| GoogleNet | RGB | 63.0 | 85.8 |
| **Ours** | RGB-D | **67.4** | **87.5** |

after a local-response-normalization that might harm the interplay with the non-normalized features of the depth branch (Fig. 3).

**Comparison to Baselines.** So far, all experiments have been carried out on the validation set, for the comparison with external baselines, the test set is used. On the Cityscapes dataset, the results of only one work has been reported that exploits depth information: [25], c.f. Sect. 2. We report the results according to the Cityscapes benchmark server [5]. Secondly, the GoogLeNet trained with the same scheme naturally serves as additional baseline.

Visually, the proposed RGB-D model outperforms the RGB baseline particularly at objects in farther distance, e.g. the car in the left image as well as the pedestrian and traffic sign in the right image of Fig. 5. Although the detection of these objects can be of highest important for e.g. autonomous vehicles, the influence on the pixel-level IoU score is rather low.




(a) class-wise IoU (b) category-wise IoU

**Fig. 3.** Where is the optimal level for modal fusion? The semantic accuracy for different levels of fusion of RGB and depth data with the proposed CNN.

### 4.3 RGBD Object Detection

**Dataset, Evaluation, and Training Details.** For object detection, we also use the Cityscapes dataset [6]. Due to the highly accurately labeled instances of all object types, bounding boxes can simply be extracted from the pixel-wise

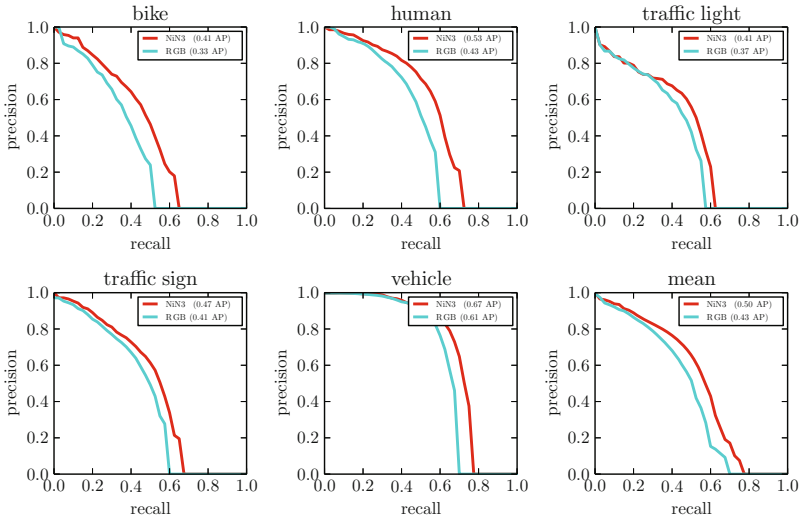



**Fig. 4.** Precision-recall curves for all object classes as well as the mean precision-recall curve for both: the RGB only baseline and our approach.

annotation. For training, we use the publicly available training data set with 2975 fully annotated images. Since the groundtruth for the test data is not publicly available, we test on the 500 images of the validation data set. Since not all classes are "object-like", we only use a subset of Cityscapes: *vehicle* (in Cityscapes: *car*, *truck*, *bus*), *bike* (in Cityscapes: *motorcycle*, *bicycle*), *traffic sign*, *traffic light*, *human* (in Cityscapes: *person*, *rider*).

For evaluation, the overlap of the groundtruth and the predicted bounding box must be larger than 0.5 for a true positive detection (TP), otherwise the prediction counts as false positive (FP). If more than one predicted bounding box overlaps with the same groundtruth box, each additional box will be counted as FP. Each missed groundtruth box is called false negative (FN). Due to the nature of bounding box detection, there are no true negatives (TN). Therefore, we follow Geiger *et al.* [12] and use the Pascal VOC measures recall, precision, and average precision (AP) [11]. The recall is the class-wise average of $\frac{TP}{TP+FN}$ and the precision $\frac{TP}{TP+FP}$. The average precision is the area under the precision-recall curve, whereby a piecewise constant interpolation was used.

For our experiments, we use the state of the art "Single Shot Multibox detector" frame work (SSD) [29]. Following our experiments from Sect. 4.2, we use a fully convolutional approach based on GoogLeNet and extend the RGB framework with the proposed and pre-trained NiN architecture for depth images.

**Results.** First, we adapt SSD to the GoogLeNet architecture with RGB input only and second, we add a depth branch as proposed in Sect. 3. Both the class-wise and the mean precision-recall curves are shown in Fig. 4. The classification of all classes benefits in similar fashion from depth data. Especially the performance


**Fig. 5.** Incorporating depth information (RGB-D, third line for semantic segmentation and fifth line for detection) leads to better segmentation of small objects, more details in the classification results, and tighter bounding boxes.

for the classes human and bike increases significantly. As shown in Fig. 5 objects in far distances are detected more robustly and more accurately by using our approach in comparison to the traditional RGB only approach.

## 5 Conclusion

This paper presented a novel generic CNN architecture that exploits input cues from other modalities in addition to sole color information. To this end, the GoogLeNet was extended with a branch specifically adapted to depth as complementary input. Together, the joint network implemented a mid-level fusion that allowed the network to exploit cross-modal interdependencies already on a medium feature-level. So far, state-of-the-art RGB-D CNNs have used network weights pre-trained on color data. In contrast, a superior initialization scheme was proposed to pre-train the depth branch of the multi-modal CNN independently. In an end-to-end training the network parameters were optimized jointly using the challenging Cityscapes dataset. The evaluation is carried on two different common computer vision tasks namely semantic segmentation and object detection. For the latter this paper furthermore showed how to extract object-level groundtruth from the instance level annotations in Cityscapes in order to

train a powerful SSD object detector. In thorough experiments, the effectiveness of the proposed multi-modal CNN was shown. Both, the RGB GoogLeNet and further RGB-D baselines were outperformed significantly.

## References


1. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. In: CVPR (2015)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected CRFs. In: ICLR (2014)
3. Chen, L.C., Yuille, A.L., Urtasun, R.: Learning deep structured models. In: ICML (2015)
4. Chen, X., Kundu, K., Zhu, Y., Berneshawi, A., Ma, H., Fidler, S., Urtasun, R.: 3D object proposals for accurate object class detection. In: NIPS (2015)
5. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: Benchmark of Cityscapes dataset. www.cityscapes-dataset.com/benchmarks/, Accessed 27 Aug 2016
6. 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: CVPR (2016)
7. Couprie, C., Farabet, C., Najman, L., LeCun, Y.: Indoor semantic segmentation using depth information. In: ICLR (2013)
8. Couprie, C., Najman, L., Lecun, Y.: Learning Hierarchical features for scene labeling. Trans. PAMI **35**(8), 1915–1929 (2013)
9. Deng, Z., Todorovic, S., Jan Latecki, L.: Semantic segmentation of RGBD images with mutex constraints. In: ICCV (2015)
10. Eitel, A., Springenberg, J.T., Spinello, L., Riedmiller, M., Burgard, W.: Multimodal deep learning for robust RGB-D object recognition. In: IROS (2015)
11. Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: a retrospective. IJCV **111**(1), 98–136 (2015)
12. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: CVPR (2012)
13. Girshick, R.: Fast R-CNN. In: ICCV (2015)
14. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR (2014)
15. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from RGB-D images for object detection and segmentation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 345–360. Springer, Cham (2014). doi:10.1007/978-3-319-10584-0_23
16. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: CVPR (2015)
17. Hazirbas, C., Ma, L., Domokos, C., Cremers, D.: Fusenet: incorporating depth into semantic segmentation via fusion-based CNN architecture. In: ACCV (2016)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2015)
19. Zhao, H., Jianping Shi, X.Q., Wang, X., Jia, J.: Pyramid scene parsing network. ArXiv (2016)
20. Hirschmüller, H.: Stereo processing by semiglobal matching and mutual information. Trans. PAMI **30**(2), 328–341 (2008)

21. Hou, S., Wang, Z., Wu, F.: Deeply exploit depth information for object detection. In: CVPRW (2016)
22. Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., Murphy, K.: Speed/accuracy trade-offs for modern convolutional object detectors. ArXiv (2016)
23. Jifeng Dai, Y.L., He, K., Sun, J.: R-FCN: object detection via region-based fully convolutional networks. In: NIPS (2016)
24. Khan, S.H., Bennamoun, M., Sohel, F., Togneri, R., Naseem, I.: Integrating geometrical context for semantic labeling of indoor scenes using RGBD images. IJCV **117**(1), 1–20 (2016)
25. Krešo, I., Čaušević, D., Krapac, J., Šegvić, S.: Convolutional scale invariance for semantic segmentation. In: Rosenhahn, B., Andres, B. (eds.) GCPR 2016. LNCS, vol. 9796, pp. 64–75. Springer, Cham (2016). doi:10.1007/978-3-319-45886-1_6
26. Li, Z., Gan, Y., Liang, X., Yu, Y., Cheng, H., Lin, L.: LSTM-CF: unifying context modeling and fusion with LSTMs for RGB-D scene labeling. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 541–557. Springer, Cham (2016). doi:10.1007/978-3-319-46475-6_34
27. Lin, M., Chen, Q., Yan, S.: Network in network. In: ICLR (2013)
28. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). doi:10.1007/978-3-319-10602-1_48
29. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). doi:10.1007/978-3-319-46448-0_2
30. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2014)
31. M. Jasch, T. Weber, M.R.: Fast and robust RGB-D scene labeling for autonomous driving. In: ICSCC, JCP (2016, to appear)
32. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: CVPR (2016)
33. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. IJCV **15**(3), 211–252 (2015)
34. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., Lecun, Y.: Overfeat: integrated recognition, localization and detection using convolutional networks. In: ICLR (2014)
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Ecology (2015)
36. Song, S., Lichtenberg, S.P., Xiao, J.: SUN RGB-D: a RGB-D scene understanding benchmark suite. In: CVPR (2015)
37. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.A.: Semantic scene completion from a single depth image. In: CVPR (2017, to appear)
38. Szegedy, C., Liu, W., Jia, Y., Sermanet, P.: Going deeper with convolutions. In: CVPR (2014)
39. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: ICLR (2015)
40. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: NIPS (2014)