Semantic Segmentation of Street Scenes Using Disparity Information

Hanwen Hu and Xu Zhao(✉)

Department of Automation, Shanghai Jiao Tong University, Shanghai, China
zhaoxu@sjtu.edu.cn

Abstract. In this work, we address the task of semantic segmentation in street scenes. Recent approaches based on convolutional neural networks have shown excellent results on several semantic segmentation benchmarks. Most of them, however, only exploit RGB information. Due to the development of stereo matching algorithms, disparity maps can be more easily acquired. Structural information encoded in disparity can be treated as supplementary information of RGB images, which is expected to boost performance. Therefore, in this work we propose to fuse disparity information in street scene understanding task. And we design four methods to incorporate disparity information into semantic segmentation framework. They are summation, multiplication, concatenation and channel concatenation. Besides, disparity map can be utilized as ground truth of a regression task, guiding the learning of semantic segmentation as a loss term. Comprehensive experiments on KITTI and Cityscapes datasets show that each method can achieve performance improvement. The experimental results validate the effectiveness of disparity information to street scene semantic segmentation tasks.

Keywords: Semantic segmentation · Disparity

1 Introduction

Semantic segmentation is one of the fundamental topics and challenging tasks in computer vision. The goal is to make pixel-wise prediction for a given image.

Since the introduction of the fully convolutional networks (FCN) [16], FCN based approaches [7,15,23] have achieved great success in semantic segmentation tasks. However, most of them only exploit RGB information. In addition to using the appearance information provided by RGB images, semantic segmentation tasks can benefit from structural information of the scene, *e.g.* depth information.

In this work, we utilize both RGB images and disparity maps to address street scene semantic segmentation tasks. The difficulty of street scene segmentation mainly lies in the complex scenes, occlusion and illumination variation, which make it still an open problem at present. Only using RGB appearance information may lead to incorrect predictions in two situations: (1) adjacent pixels that share the same semantic categories but different appearances; (2)

Y. Zhao et al. (Eds.): ICIG 2019, LNCS 11901, pp. 169–181, 2019.
https://doi.org/10.1007/978-3-030-34120-6_14

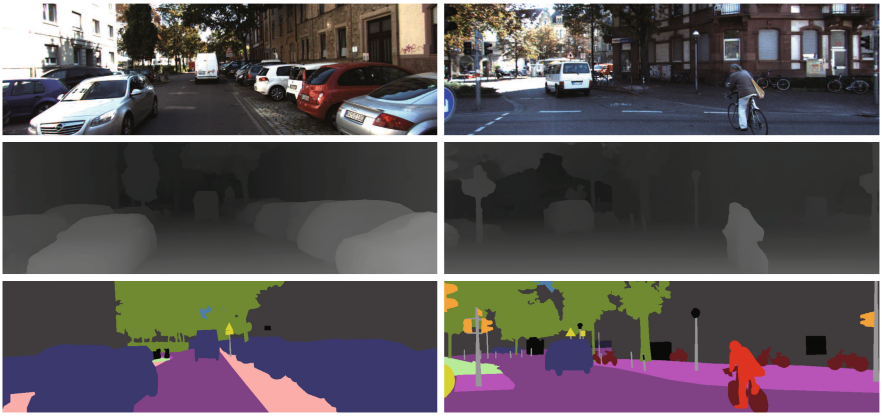

Fig. 1. Training examples from KITTI [2] dataset. From top to bottom: original RGB images, disparity maps generated by PSMNet [3] and ground truths. Disparity maps can be used as supplementary information to RGB images.

adjacent pixels that have different semantic labels but with similar appearances. Disparity maps provide structural information as well as hierarchical relationships between objects of the scenes, which can be exploited as supplementary information to RGB images.

Previous approaches utilizing depth information mainly focus on indoor scene understanding tasks. Several approaches [11,21] treat depth map as an additional input channel, utilizing FCN based models to segment RGB-D images. Gupta et al. [12] transform depth maps to HHA image[1] and [13] employs two branches of CNNs to extract RGB and HHA image features respectively. In our work, we use disparity maps as depth information to address the task of semantic segmentation in street scenes. Disparity maps are acquired by applying PSMNet [3], which is a state-of-the-art stereo matching algorithm. Training examples are illustrated in Fig. 1.

In this work, we employ MobileNetV2 [19] and Xception [8] to extract RGB image features. VGG [20] like fully convolutional network is utilized as disparity feature extractor. We propose four fusion methods, *i.e. summation*, *multiplication*, *concatenation* and *channel concatenation*, to incorporate disparity information into semantic segmentation framework without introducing much computation complexity. Besides, we treat disparity map as ground truth of a regression task, imposing extra constraints to guide training process along with semantic segmentation task. Each method is evaluated on two street scene datasets: KITTI [2] and Cityscapes [9]. The experimental results indicate that each fusion method can improve the performance of semantic segmentation and validate the effectiveness of disparity information to street scene understanding tasks.

[1] HHA image consists of three channels: horizontal disparity, height of the pixels and norm angle.

2 Related Work

2.1 Semantic Segmentation

Long et al. [16] transforms classification-purposed CNN models into Fully Convolutional Networks (FCN) by replacing fully connected layers with convolutional layers. Recent methods [7,15,23] mainly focus on: (1) encoding context information; (2) decoding semantic features and (3) providing structured outputs.

Encoding Context Information. It is of great importance to utilize context information of an image in semantic segmentation tasks. Dilated convolution is utilized in [22] to control the resolution of extracted features and aggregate multi-scale contextual information. Besides, PSPNet [23] and Deeplab [4–7] exploiting spatial pyramid pooling modules to encode context information. PSPNet implements spatial pyramid pooling at several grid scales. While Deeplabv2 [5] introduces *Atrous Spatial Pyramid Pooling* (ASPP) module that applies several parallel dilated convolutions with different dilation rates.

Decoding Semantic Features. Because of pooling layers and strided convolutional layers, encoder module gradually reduces the resolution of feature maps and captures high-level semantic information. Several approaches have been proposed to recover resolution from semantic features. Bilinear interpolation [6] and transposed convolution [17] are commonly used operations to enlarge resolution. Skip connections [15,18] between encoder parts and decoder parts have been adopted to acquire accurate information.

Providing Structured Outputs. Conditional Random Fields (CRF) is often utilized to generate sharper results. Deeplabv1 [4] applies DenseCRF [14] as a post-processing method to refine predictions along object boundaries. CRF-as-RNN [24] interprets mean-field of CRFs as a recurrent neural network, allowing for training CNN and CRF parameters in an end-to-end differentiable network.

In this work, we utilize dilated convolution to enlarge receptive field without reducing the resolution of feature maps. Moreover, we adopt ASPP module to aggregate multi-scale and context information. Bilinear interpolation is used in decoding stage to recover the resolution.

2.2 RGB-D Semantic Segmentation

The availability of low-cost range sensors advances the progress of RGB-D semantic segmentation. However, RGB-D semantic segmentation mainly focuses on indoor scene understanding tasks. Several approaches [11,21] treat depth map as an additional input channel, utilizing FCN based models to segment RGB-D images. Gupta et al. [12] transform depth to HHA image, which consists of three channels: horizontal disparity, height of the pixels and norm angle. [13] employs two branches of CNNs to extract RGB and HHA image features respectively, achieving promising results on several indoor scene semantic segmentation benchmarks.

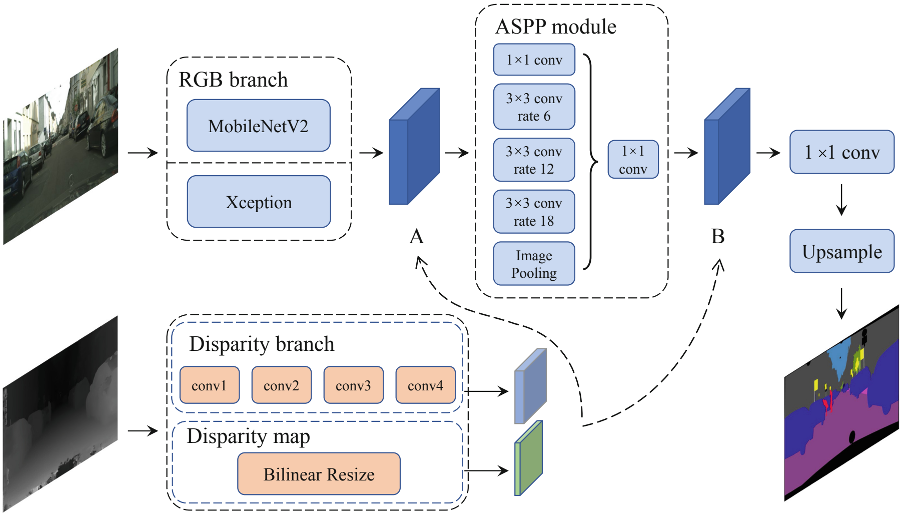

Fig. 2. Network structure. Two branches of CNNs are employed to extract RGB and disparity features respectively. MobileNetV2 and Xception are employed as RGB feature extractor, followed by ASPP module. The disparity features or resized disparity maps are fused with RGB features. Two possible fusion options are marked with A and B. Position A stands for *early fusion* while position B strands for *late fusion*.

In this work, we focus on street scene semantic segmentation instead of indoor scenes. Although both disparity maps and depth maps encode 3D information, disparity maps can be effectively computed using stereo matching algorithms with no need of other sensors. Hence, we exploit disparity map as an additional information to perform street scene semantic segmentation.

3 Methods

3.1 Network Architecture

The architecture of our network is illustrated in Fig. 2. We have two branches of networks to extract features from RGB images and disparity maps, respectively.

In RGB branch, we employ two CNNs as backbone networks: MobileNetV2 [19] and Xception [8]. Both of them leverage depthwise separable convolution as basic building block, greatly reducing the number of parameters. And residual connections are essential in helping with convergence. Besides, we enlarge the resolution of the final feature maps using dilated convolution. We set the stride of last pooling or strided convolutional layer to 1 and replace all subsequent convolutional layers with dilated convolutional layers. This allows us to extract denser feature maps without sacrificing receptive field. Although convolutional features have shown remarkable ability to implicitly represent scale, explicitly accounting for object scale can further improve the ability to handle

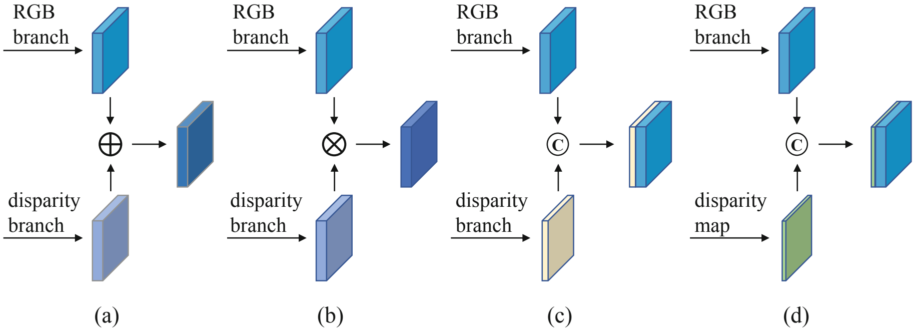

Fig. 3. Illustration of fusion methods. (a) Element-wise summation. (b) Element-wise multiplication. (c) Concatenation. (d) Channel concatenation. $\oplus$, $\otimes$, ⓒ represent summation, multiplication and concatenation respectively.

objects with various size. Thus, we adopt *atrous spatial pyramid pooling* (ASPP) module, which introduces dilated convolution into spatial pyramid pooling, to aggregate multi-scale and context information. We stack ASPP module on the top of backbone networks following the spirit of [7].

In disparity branch, we apply a light-weight fully convolutional network to extract disparity features. The network follows the VGG [20] style and has fewer channels in convolutional layers. It consists of alternating convolutional blocks and max pooling layers. Disparity features are then fused with RGB features.

3.2 Fusion Strategies

As illustrated in Fig. 2, we have two options to fuse disparity information:

Early Fusion. Since ASPP module is used to extract multi-scale features, fusion of disparity and RGB features before ASPP module is expected to simultaneously capture context information of RGB features and disparity information.

Late Fusion. The second possible fusion position is after ASPP module. Fused features are only passed through one 1×1 convolutional layer to obtain final logits, which are the feature maps before softmax activation. The motivation is to let disparity information directly influence prediction process.

Four fusion methods are proposed to incorporate disparity information into semantic segmentation framework without introducing much computation complexity. Let f_{RGB}, f_{Disp} and F denote RGB features, disparity features and fused features respectively. Proposed methods are illustrated in Fig. 3 and described as follows.

Summation. Inspired by [13], we apply element-wise summation to fuse features. Both disparity branch and ASPP module generate 256-channel features. When *late fusion* is adopted, element-wise summation is performed directly upon disparity features and RGB features. However, when *early fusion* is applied, the output of RGB branch is first reduced by a 1 × 1 convolutional layer, generating 256-channel feature maps. We formulate this fusion method as follows:

$$F = f_{RGB} \oplus f_{Disp}. \tag{1}$$

Multiplication. In this method, we regard disparity feature maps as masks, multiplying RGB features element-wisely. Because of ReLU non-linearity, the value of each element in feature maps is greater or equal to 0. Hence, it is expected that disparity features would provide additional information to corresponding areas in RGB features. Opreation of *early fusion* follows the same operation in *summation* method. This method can be written as:

$$F = f_{RGB} \otimes f_{Disp}. \tag{2}$$

Concatenation. Another intuitive method is to concatenate RGB feature maps and disparity feature maps together, automatically fusing them by convolutional layers. Therefore, the most useful disparity features would be assigned higher weights in convolution. As additional information, the proportion of disparity features should not exceed the proportion of RGB features. Thus, we modify the disparity branch to generate 128-channel features, which are then concatenated to 256-channel RGB features. While in *early fusion* the disparity features are directly concatenated to the output of RGB branch. The fused feature is as below:

$$F = [f_{RGB}, f_{Disp}]. \tag{3}$$

Channel Concatenation. This method treats disparity map as an addition channel to RGB features. Disparity maps are bilinearly resized to the resolution of RGB feature maps, *i.e.* 1/16 of input image size, and then directly concatenated to them. This method utilizes original disparity maps and introduces minimal computation complexity. It is noteworthy that the only operation we perform on disparity maps is normalization, so that the range of disparity maps is $[-1, 1]$. This method can be represented by the following equation with *Disp* standing for bilinearly resized disparity maps:

$$F = [f_{RGB}, Disp]. \tag{4}$$

All the four methods are applied at each fusion position. Fused feature F is utilized to compute final output. Thus we incorporate disparity information into semantic segmentation. In conclusion, we propose four methods and two options to fuse disparity information with RGB features. We also combine RGB images and disparity maps into four-channel RGB-Disparity images, which are directly fed into the network. The network has no disparity branch in this case.

3.3 Disparity Loss Regularization

The disparity information can also guide learning of semantic segmentation as a loss term. We treat disparity map as ground truth of a regression task. Based on the final RGB feature map, we employ a convolutional layer to predict disparity of the image. The total loss is expressed as

$$L = \frac{1}{N} \sum_i L_{cls}(y^i_{label}, y^i_{pred}) + \lambda \cdot L_{reg}(d^i_{label} - d^i_{pred}), \tag{5}$$

where L_{cls} is cross-entropy loss and L_{reg} is disparity regression loss. y^i_{label} and y^i_{pred} are ground truth class and predicted class of semantic segmentation task respectively. d^i_{label} and d^i_{pred} are ground truth disparity and predicted disparity of disparity regression task respectively. We adopt Huber loss as the disparity regression loss, which is expressed as below:

$$L_{reg}(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| < 1 \\ |x| - \frac{1}{2} & \text{oterwise.} \end{cases} \tag{6}$$

The hyper-parameter λ controls the balance between the two task losses and we set $\lambda = 10.0$ empirically.

When training the network, the disparity regression loss L_{reg} is propagated back to the RGB feature extractor. Along with the basic cross-entropy loss L_{cls}, disparity regression loss L_{reg} imposes extra constraints to guide training process. We note that the network has no disparity branch in this case.

4 Experiments

4.1 Datasets and Implementation Details

We evaluate our approach on two challenging semantic segmentation datasets: KITTI [2] and Cityscapes [9]. Both of them focus on street scenes segmentation and provide rectified stereo image pairs. We apply PSMNet [3] to calculate dense disparity maps for both datasets.

KITTI. KITTI [2] semantic segmentation benchmark consists of 200 semantically annotated train images as well as 200 test images. The images were recorded while driving around a mid-size city, in rural areas and on highways. And the resolution of images is mainly 375 × 1242. We further split the whole training set into a training set (160 images) and a validation set (40 images).

Cityscapes. Cityscapes [9] dataset is a large-scale dataset that focuses on semantic understanding in urban street scenes. It consists of 5000 images (2975, 500, 1525 for the training, validation, and test sets respectively) with fine annotations and another 20000 images with coarse annotations. In this work, experiments are only conducted on the fine-annotated dataset.

Implementation Details. We employ MobileNetV2 and Xception, which have been pretrained on ImageNet dataset [10], as network backbone to extract features of RGB images. The parameters in disparity branch are initialized randomly. We follow the training protocol as in [6]. Input images are scaled randomly from 0.5 to 2.0 and are flipped randomly left-right during training stage.

Table 1. Results on KITTI validation set. The name of the network indicates network backbone, fusion position and fusion method.

Network	mIoU	Network	mIoU
MobileNetV2_baseline	49.40	Xception_baseline	52.76
MobileNetV2_RGB-D	50.27	Xception_RGB-D	54.54
MobileNetV2_earlyfusion_sum	51.88	Xception_earlyfusion_sum	56.66
MobileNetV2_earlyfusion_mul	51.62	Xception_earlyfusion_mul	**58.96**
MobileNetV2_earlyfusion_concat	51.77	Xception_earlyfusion_concat	57.62
MobileNetV2_earlyfusion_channel	**55.74**	Xception_earlyfusion_channel	58.75
MobileNetV2_latefusion_sum	50.62	Xception_latefusion_sum	57.37
MobileNetV2_latefusion_mul	52.63	Xception_latefusion_mul	56.90
MobileNetV2_latefusion_concat	**54.15**	Xception_latefusion_concat	**59.71**
MobileNetV2_latefusion_channel	50.47	Xception_latefusion_channel	59.04

The final logits output by the network are upsampled to the size of ground truths to preserve annotation details. SGD optimizer with momentum 0.9 is used to train the network. And the initial learning rate is multiplied by $(1-\frac{iter}{max\_iter})^{0.9}$ to decrease learning rate. The performance is measured in terms of mIoU, which is the mean value of classwise intersection-over-union. All experiments are built with TensorFlow [1] framework on a single NVIDIA Titan X (Pascal) GPU.

4.2 Experimental Results

KITTI. Baseline network only adopts RGB branch and is trained with the training protocols described above. In addition, we also conduct experiment with RGB-Disparity input where disparity map is treated as an additional input channel. We note that the network has no disparity branch in this case and parameters of the first convolutional layer are randomly initialized.

The results of MobileNetV2 and Xception based networks are summarized in Table 1. Furthermore, mIoU curves on KITTI validation set during training are provided in Fig. 4. The curves indicate that networks have converged at the end of training. We observe that each fusion method, no matter *early fusion* or *late fusion*, outperforms the performance of baseline. The four-channel RGB-D input method achieves the smallest performance improvement compared with baseline. Best performing *MobileNetV2_earlyfusion_channel* and *Xception_latefusion_concat* increase performance by 6.34% and 6.95%, respectively. Interestingly, performance gain can be achieved as well by the simplest *channel concatenation* fusion method. Thus, experimental results reveal that disparity information can be exploited as supplementary information to street scene understanding tasks.

Comparing the results of MobileNetV2 and Xception based networks, performance gain can be observed on both of them and Xception based networks

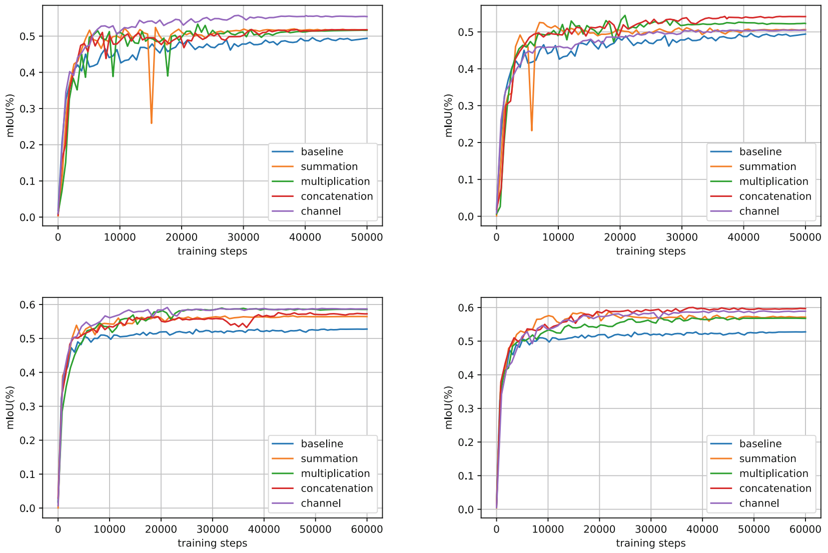

Fig. 4. mIoU curves on KITTI validation set. The first row stands for MobileNetV2 based networks and the second row stands for Xception based networks. The first column represents *early fusion* and the second column represents *late fusion*.

perform better because of more representative features. The results are accord with our expectation and also validate that disparity information is effective to different RGB convolutional features. Although each fusion method can improve performance, it is hard to tell which fusion method is the best. However, *channel concatenation* can be used to boost performance in resource constrained environments. At last, qualitative results on KITTI validation set are illustrated in Fig. 5. It shows that cars and poles are segmented better with the help of disparity maps.

Cityscapes. Experimental results are summarized in Table 2. Compared with the baseline, the performance of RGB-Disparity input method decreases slightly by 0.52%. We speculate that randomly initialized parameters of the first convolutional layer make the network difficult to learn representative features. Except RGB-Disparity method, each fusion method obtains performance gain on Cityscapes dataset, which is consistent with the results on KITTI dataset. *Xception_latefusion_sum* achieves best result and improves the performance by 4.08% over baseline. Besides, the results of *summation*, *concatenation*, and *channel concatenation* fusion methods are close to each other, surpassing the performance of baseline by about 3–4%. Hence, it can be concluded that disparity maps provide useful information to street scene understanding tasks.

We note that the disparity maps of Cityscapes dataset are calculated by PSMNet, which is trained on KITTI 2015 stereo dataset. Better results can be expected if we obtain more accurate disparity maps of Cityscapes.

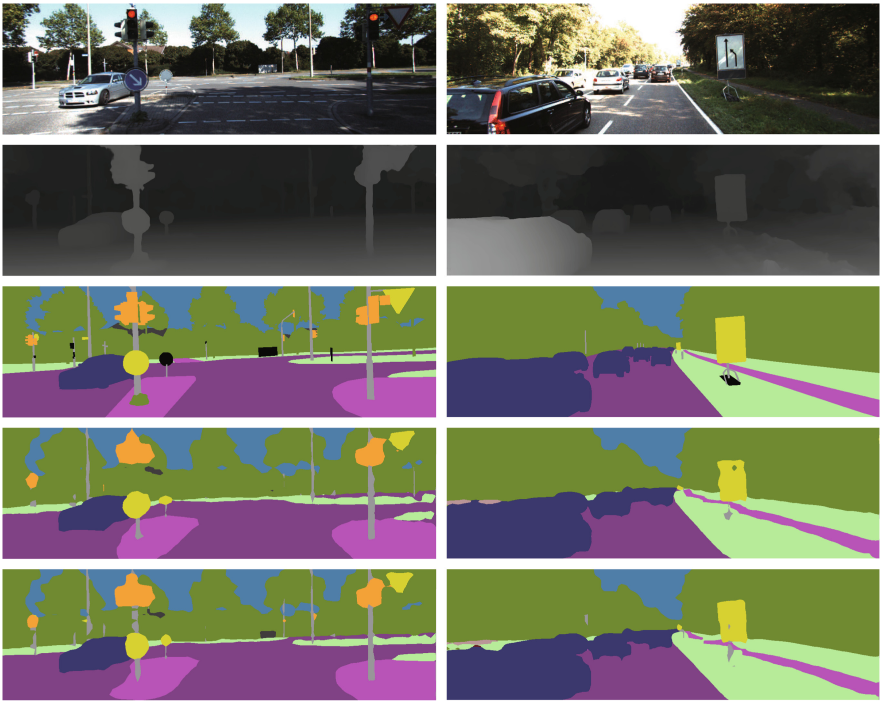

Fig. 5. Visualization results on KITTI validation set. From top to bottom: RGB images, disparity maps, ground truths, results of *Xception_baseline*, results of *Xception_latefusion_concat*.

Table 2. Results on Cityscapes validation set. The name of the network indicates network backbone, fusion position and fusion method.

Network	mIoU
Xception_baseline	69.34
Xception_RGB-Disparity	68.82
Xception_earlyfusion_sum	72.27
Xception_earlyfusion_mul	71.56
Xception_earlyfusion_concat	72.92
Xception_earlyfusion_channel	**73.23**
Xception_latefusion_sum	**73.42**
Xception_latefusion_mul	71.86
Xception_latefusion_concat	73.31
Xception_latefusion_channel	72.27

Disparity Loss Regularization. When we treat disparity map as ground truth of a regression task, a 3 × 3 convolutional layer is employed on the final RGB feature map of Xception to predict disparity of the image. In this case the network has only RGB branch. Table 3 shows experimental results on KITTI and Cityscapes datasets. Compared with baseline, the performance of disparity regularized methods increase 1.62% and 0.89% on KITTI and Cityscapes respectively. Experimental results manifest that disparity information can guide the learning of semantic segmentation in street scenes.

Table 3. mIoU results of disparity regularized method on KITTI and Cityscapes dataset.

Network	KITTI	Cityscapes
Xception_baseline	52.76	69.34
Xception_regression	54.38	70.23

5 Conclusions

We leverage disparity maps as supplementary information to address semantic segmentation in street scenes. Two branches of CNNs are employed to extract RGB and disparity features respectively. We propose four methods and two options to incorporate disparity information into semantic segmentation framework. Experimental results on KITTI and Cityscapes validate the effectiveness of disparity information to street scene understanding tasks. Moreover, each fusion method can improve the performance of semantic segmentation over baseline without introducing much computation complexity. We note that the proposed fusion methods are easy to implement and can be applied to fuse other information. Besides, we treat disparity map as ground truth of a disparity regression task and employ a convolutional layer on the final RGB feature map to predict disparity of the scene. Experimental results validate that disparity information can guide learning of street scene semantic segmentation.

Acknowledgments. This work is supported by: National Natural Science Foundation of China (U1764264, 61673269, 61273285).

References

1. Abadi, M., et al.: Tensorflow: a system for large-scale machine learning. In: OSDI 2016, pp. 265–283 (2016)
2. Alhaija, H.A., Mustikovela, S.K., Mescheder, L., Geiger, A., Rother, C.: Augmented reality meets deep learning for car instance segmentation in urban scenes. In: British Machine Vision Conference, vol. 1, p. 2 (2017)

3. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5410–5418 (2018)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.: Semantic image segmentation with deep convolutional nets and fully connected CRFs. In: ICLR, San Diego, United States, May 2015. https://hal.inria.fr/hal-01263610
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)
6. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
7. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv preprint arXiv:1802.02611 (2018)
8. Chollet, F.: Xception: deep learning with depthwise separable convolutions. arXiv preprint arXiv:1610.02357 (2017)
9. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: CVPR, pp. 3213–3223 (2016)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 248–255. IEEE (2009)
11. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2650–2658 (2015)
12. 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). https://doi.org/10.1007/978-3-319-10584-0_23
13. Hazirbas, C., Ma, L., Domokos, C., Cremers, D.: FuseNet: incorporating depth into semantic segmentation via fusion-based CNN architecture. In: Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y. (eds.) ACCV 2016. LNCS, vol. 10111, pp. 213–228. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54181-5_14
14. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected CRFs with Gaussian edge potentials. In: Advances in Neural Information Processing Systems, pp. 109–117 (2011)
15. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: multi-path refinement networks for high-resolution semantic segmentation. In: CVPR (2017)
16. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
17. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: ICCV, pp. 1520–1528 (2015)
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
19. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)

20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
21. Wang, J., Wang, Z., Tao, D., See, S., Wang, G.: Learning common and specific features for RGB-D semantic segmentation with deconvolutional networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 664–679. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_40
22. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
23. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2881–2890 (2017)
24. Zheng, S., et al.: Conditional random fields as recurrent neural networks. In: ICCV, pp. 1529–1537 (2015)