Semantic Stereo Matching with Pyramid Cost Volumes

Zhenyao Wu[1], Xinyi Wu[1], Xiaoping Zhang[2], Song Wang[1,3,*], Lili Ju[1,3,*]
[1]University of South Carolina, USA [2]Wuhan University, China [3]Farsee2 Technology Ltd, China
{zhenyao, xinyiw}@email.sc.edu, xpzhang.math@whu.edu.cn, songwang@cec.sc.edu, ju@math.sc.edu

Abstract

The accuracy of stereo matching has been greatly improved by using deep learning with convolutional neural networks. To further capture the details of disparity maps, in this paper, we propose a novel semantic stereo network named SSPCV-Net, which includes newly designed pyramid cost volumes for describing semantic and spatial information on multiple levels. The semantic features are inferred by a semantic segmentation subnetwork while the spatial features are derived by hierarchical spatial pooling. In the end, we design a 3D multi-cost aggregation module to integrate the extracted multilevel features and perform regression for accurate disparity maps. We conduct comprehensive experiments and comparisons with some recent stereo matching networks on Scene Flow, KITTI 2015 and 2012, and Cityscapes benchmark datasets, and the results show that the proposed SSPCV-Net significantly promotes the state-of-the-art stereo-matching performance.

1. Introduction

Stereo matching is indispensable for many computer vision applications, such as autonomous driving [5], 3D reconstruction [42], augmented realities [8], and robot navigation [1]. By finding pixel-level correspondence between two images, stereo algorithms aim to construct a disparity map from a pair of rectified stereo images. In traditional methods, hand-crafted reliable features are used to identify cross-image matching pixels or patches for computing the disparity map [3, 33]. Recently, as in many other computer vision tasks, convolutional neural networks (CNNs) have been applied to stereo matching with significant success.

When applying CNNs for stereo matching, many of the existing works construct a cost volume for computing the correspondence cost at each position by traversing a set of possible disparity values. A regression layer is then used to infer the optimal disparity map based on the cost volume. While early works calculate the cost in the original image domain [17, 18, 37], recent works construct the cost volume using the deep features extracted by the respective networks [22, 30, 35, 4]. For these prior works, the cost volume is constructed at a single level without considering multiscale spatial information separately underlying the stereo image pairs. However, for the considered feature map, a single-scale cost volume may not be sufficient to capture the spatial relationship between stereo images. One of our major ideas in this paper is to develop a new CNN network with multilevel cost volumes, which we call *pyramid cost volumes*, for better capturing the disparity details in stereo matching.

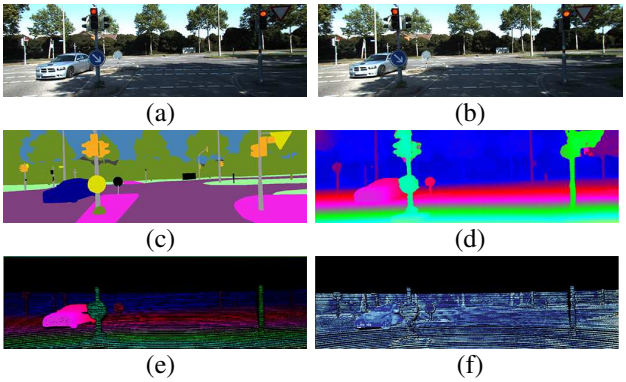

Figure 1. (a)&(b) the input stereo pairs (left and right images) from KITTI dataset; (c) the semantic segmentation; (d) the predicted disparity by the proposed SSPCV-Net; (e) the ground-truth of the disparity estimation; (f) the 3px-error map of the SSPCV-Net prediction.

Our work is also partly inspired by the recent work of SegStereo [38] that integrates semantic information to stereo matching through joint learning. As shown in Figure 1, semantic segmentation captures different objects and their boundaries in images and shows much spatial and intensity correlation with the disparity map. In particular, an accurate semantic segmentation can help rectify the disparity values along the object boundaries, which are usually more prone to error in stereo matching [2, 15]. Thus, our network will also integrate both the semantic and the spatial information in multiple levels for constructing pyramid

*Co-corresponding authors.

cost volumes, and we find that such an approach can improve the stereo-matching accuracy significantly.

More specifically, we design a new semantic stereo network named SSPCV-Net for stereo matching. In this network, after several initial convolutional layers, we take the extracted deep features as input for two separate branches. One of them performs the traditional spatial pooling, but with hierarchical multilevel processing. The other branch is a semantic segmentation subnetwork. We then build pyramid cost volumes by combining the outputs of these two branches from input stereo pairs such that these new pyramid cost volumes well represent both semantic and spatial information in multiple levels. Next, we design a 3D multi-cost aggregation module to integrate the extracted multilevel features and perform regression for predicting disparity maps. We employ a two-step strategy to train the SSPCV-Net: 1) supervised training of the semantic segmentation subnetwork; and 2) joint training of the whole network with supervision on both semantic segmentation and disparity estimation. We conduct comprehensive experiments, including a series of ablation studies and comparison tests of SSPCV-Net with existing state-of-the-art methods on Scene Flow, KITTI 2015 and KITTI 2012 benchmark datasets, and moreover, we also perform tests on Cityscapes dataset to compare their generalization abilities. It is observed that the proposed SSPCV-Net clearly outperforms many existing state-of-the-art stereo-matching methods. The major contributions of this paper are:

- We propose a new semantic stereo network of SSPCV-Net, in which we construct pyramid cost volumes for capturing semantic and multiscale spatial information simultaneously.
- We propose a 3D multi-cost aggregation module in SSPCV-Net to integrate the extracted multilevel features and perform regression for accurate disparity-map prediction.
- SSPCV-Net significantly promotes the state-of-the-art performance of stereo matching on the benchmark datasets of Scene Flow, KITTI 2015 and 2012, and CityScapes.

2. Related Work

Almost all recent state-of-the-art performances of stereo matching are achieved by using CNN-based methods. For example, in [27, 13], disparity value is discretized and disparity estimation is reduced to classification with CNN. In [28], CNN is used for computing disparity map and optical flow simultaneously. This result can be refined iteratively based on error maps [30]. In [34], the disparity is estimated by patch matching. In [23], the use of low-resolution cost volumes leads to sub-pixel matching accuracy and real-time speed. In [10], a new 3D convolutional module, as well as a sparse depth map, is used for improving stereo matching. All these methods construct single-scale cost volumes. In this paper, we build multilevel cost volumes for better disparity estimation.

More related to our work are EdgeStereo [35], GC-Net [22] and PSMNet [4]. In EdgeStereo [35], edge detection is incorporated to accurately estimate depth change across object boundaries, while in this paper, we incorporate semantic segmentation to achieve this goal. In GC-Net [22], cost volumes are regularized by 3D convolutions before used for disparity estimation. Based on GC-Net, PSMNet [4] extracts multiscale image information for constructing a single cost volume, which is then taken for regularization and disparity estimation. Following the general framework of GC-Net and PSMNet, we here construct multilevel cost volumes, together with a 3D multi-cost aggregation module, to better capture the global context information for disparity estimation.

Semantic information has been found to be useful when integrated to solve many important computer vision problems. For example, in [9] an integrated SegFlow model is developed to address optical flow and video segmentation together, leading to a win-win result. In [20, 43, 21], two tasks of monocular depth estimation and semantic segmentation are solved simultaneously by using weight-sharing sub-networks or joint CNN learning. One of our main goals in this paper is to integrate semantic segmentation into stereo matching. Related to our work is SegStereo [38], which combines semantic and image features into a single cost volume for disparity estimation. Different from SegStereo, we propose to construct cost volumes for semantic features and image features separately, as well as using multilevel cost volumes of image features. The experiments show that the proposed approach can improve the accuracy significantly.

Multiscale information has been used in many CNN-based computer vision applications. For example, PSPNet [44] and DeepLab [7, 6] embed multiscale features of scenes to improve semantic segmentation. SPyNet [32] calculates optical flow by warping images in multiple scales. PWC-Net [36] uses multiscale features to compute optical flow with a single branch. Different from these works, we here introduce the multiscale information into stereo matching, as in PSMNet [4]. But as discussed above, PSMNet constructs a single cost volume using multiscale features, while we construct multilevel cost volumes directly, resulting in much better disparity estimation.

3. Our Approach

The architecture of the proposed SSPCV-Net is shown in Figure 2. We can see that new pyramid cost volumes are built to incorporate semantic information and multilevel spatial context information. In addition, a 3D multi-cost

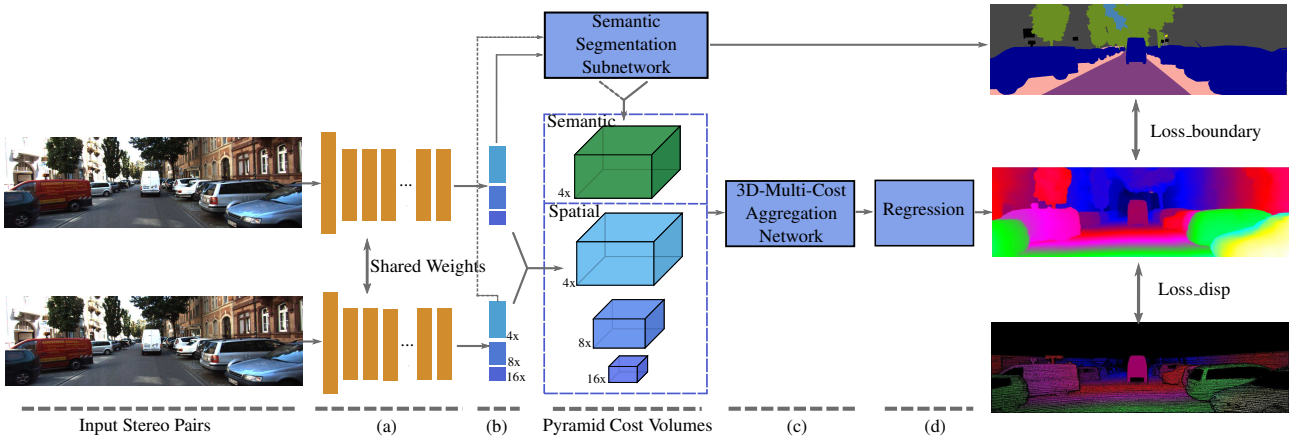

Figure 2. Architecture of the proposed semantic stereo network (SSPCV-Net) for disparity estimation. The main pipeline includes: (a) feature extraction: using ResNet50 [16]; (b) spatial pooling: using average pooling, and the resulted multilevel feature maps are fed into the semantic segmentation network; (c) multi-cost aggregation: fusing the pyramid cost volumes, and the details of this module is shown in Figure 4; (d) disparity regression: disparity map is estimated from the cost volumes using 3D convolution.

aggregation module is added for cost-volume aggregation and regularization.

3.1. Network architecture

We first use ResNet-50 [16] with the dilated network strategy [6, 40] to extract features from the input pair of images, and then adopt adaptive average pooling to compress features into three scales, followed by a 1×1 convolution layer to change the dimension of the feature maps. The resulting spatial features are simultaneously fed into two branches of the network – one branch produces spatial pyramid cost volumes directly and the other branch is a semantic segmentation subnetwork, which generates a semantic cost volume. The obtained semantic cost volume and the spatial cost volumes make up pyramid cost volumes. All these cost volumes are then fed into a 3D multi-cost aggregation module for aggregation and regularization. At the end, a regression layer produces the final disparity map. The pyramid cost volumes and the 3D multi-cost aggregation module are elaborated in the following sections.

3.2. Pyramid cost volumes

We design two branches to produce the cost volumes: the spatial branch generates spatial pyramid cost volumes and the semantic branch generates one semantic cost volume, as shown in the box of *Pyramid Cost Volumes* in Figure 2.

3.2.1 Spatial pyramid cost volumes

We propose to use the idea of pyramid cost volumes to learn the relationship between an object and its neighbors in space. Different from PSMNet, where only a single cost volume is generated from the pyramid features by first upsampling them to the same dimension and then performing concatenation, we instead use multilevel spatial features to build spatial pyramid cost volumes.

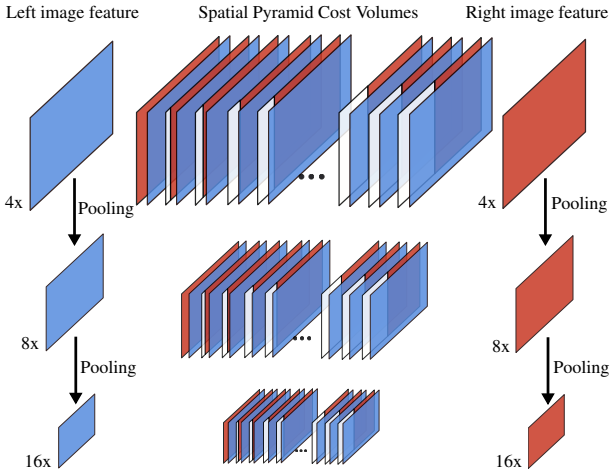

Figure 3. The construction process of spatial pyramid cost volumes from left and right image features by spatial pooling.

We use hierarchical scales of spatial features after different adaptive average pooling layers in feature extraction to form levels of cost volumes. Following the idea of GC-Net [22], for each level of the spatial feature maps, we form a cost volume by concatenating the corresponding unaries from the left and right image features and then packing them into a 4D volume, which contains all spatial context information for inferring disparity from this level. As shown in Figures 2 and 3, three hierarchical levels of feature maps are particularly used in our SSPCV-Net to form spatial pyramid cost volumes to represent different level of information, and the spatial pyramid cost volumes have sizes of $C \times \alpha W \times \alpha H \times \alpha D$ with $\alpha \in \{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}\}$ respectively at each level, where C is number of channels, W and H are

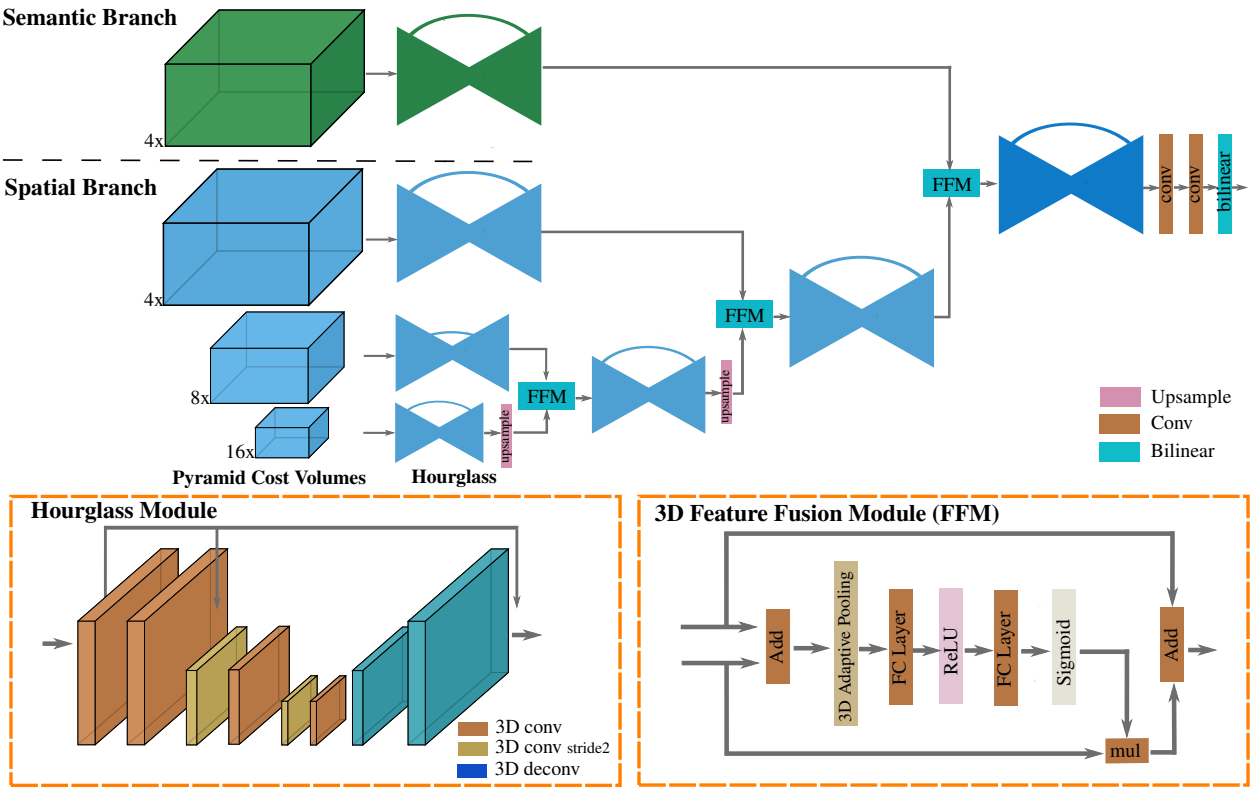

Figure 4. Details of the 3D multi-cost aggregation module with the hourglass and the 3D feature fusion.

the width and height of original images respectively, and D is the maximum disparity.

3.2.2 Semantic cost volume

For the semantic branch, the semantic segmentation subnetwork follows PSPNet [44]. With the extracted feature maps, the subnetwork upsamples the low-dimensional feature maps to the same size and concatenates all the feature maps. In the end, it is followed by a convolution layer to generate the final prediction of the semantic segmentation map.

To form the single semantic cost volume, we use the features before the classification layer. The use of semantic cost volume aims to capture context cues in a simple manner and learn the similarity of objects' pixels from the left and right semantic segmentation features. By concatenating each unary semantic feature with their corresponding unary from the opposite stereo image across each disparity level, and packing them into a 4D volume, we obtain a semantic cost volume with the size of $C \times \frac{1}{4}W \times \frac{1}{4}H \times \frac{1}{4}D$, which is the same size as the largest spatial cost volume.

3.3. 3D multi-cost aggregation module

As shown in Figure 4, both the spatial pyramid cost volumes and the semantic cost volume are fed into the 3D multi-cost aggregation module. We use a "Hourglass" module and a 3D feature fusion module (FFM) to learn different levels of spatial context information through the encoding/decoding process. As for the strategy, inspired by the MSCI (multiscale context intertwining) scheme [25] and RefineNet [26], we fuse the 4D spatial cost volumes from the lowest level to the higher ones in a recursive way: we first upsample the lower level volume to the same size as its immediately higher level one and feed them into FFM, then the fused cost volume is further fused with the next higher level cost volume after the hourglass module. Finally, the last level fused spatial cost volume is fused with the semantic cost volume and the result is then upsampled to the original image size $1 \times W \times H \times D$ via the bilinear interpolation.

Instead of concatenating the features as in BiSeNet [39], which includes a 2D feature fusion module to help the context information fusion, we develop a 3D feature fusion module specifically for fusing two cost volumes: first the two 3D cost volumes are summed up following the residual block structure in [16], next the adaptive average pooling is used to transform the concatenated features to a feature vector and then a weight vector is computed through a fc-ReLU-fc-sigmoid structure [19], finally, the upsampled one of the two cost volumes is multiplied by the weight vector and added with the other cost volume to form the output of

the FFM module.

3.4. Disparity regression and loss function

We take the disparity regression proposed in [22, 4] to estimate the continuous disparity map. The softmax operation $\sigma(\cdot)$ is first used to normalize the finally fused cost volume C_d to output a probability $P(d)$ for each disparity d, which is regarded to as a soft attention mechanism and often more robust than classification-based approaches. The predicted disparity $\hat{d}$ is then calculated as the sum of each disparity d weighted by its probability as

$$\hat{d} = \sum_{d=0}^{D_{max}} d \times P(d) \tag{1}$$

where D_{max} denotes the maximum disparity.

To train the proposed architecture, we rely on the following multi-task loss function.

$$L = \alpha L_{disp} + (1 - \alpha) L_{bdry} \tag{2}$$

which consists of the weighted sum ($0 \leq \alpha \leq 1$ is the weight) of two terms, namely the disparity loss (L_{disp}) and the boundary loss (L_{bdry}).

We use the $smooth_{L_1}$ as the basic loss function to train our proposed SSPCV-Net which has been widely used in many regression tasks [14, 22]. The disparity loss is defined as

$$L_{disp}(d^*, \hat{d}) = \frac{1}{N} \sum_{(i,j)} smooth_{L_1}(d^*_{i,j}, \hat{d}_{i,j}) \tag{3}$$

where N is the number of all the labeled pixels, d^* is the disparity ground-truth. Since the disparity discontinuity point is always on the semantic boundaries[31], we accordingly deploy the following boundary-loss function as

$$\begin{aligned} L_{bdry} = \frac{1}{N} \sum_{(i,j)} \big(& |\varphi_x(sem_{i,j})| e^{-|\varphi_x(\hat{d}_{i,j})|} \\ & + |\varphi_y(sem_{i,j})| e^{-|\varphi_y(\hat{d}_{i,j})|} \big) \end{aligned} \tag{4}$$

where sem is the semantic segmentation ground-truth label, and φ_x and φ_y are the intensity gradients between neighboring pixels along the x and y directions, respectively.

4. Experiments

4.1. Datasets and evaluation metrics

In this section, we use the following stereo datasets for performance evaluation and comparison of SSPCV-Net with several recent state-of-the-art networks for stereo matching:

Scene Flow [28]: This is a synthetic dataset consists of 35,454 training and 4,370 testing image pairs that can be used for evaluating optical flow and stereo matching performance. This dataset has dense and elaborate disparity maps as ground-truth for training.

KITTI 2015 & KITTI 2012 [29, 12]: These are two real-world datasets. KITTI 2015 contains 200 training stereo image pairs with sparse ground-truth disparities and another 200 testing image pairs without ground-truth disparities. The left (reference) images of the stereo image pairs have semantic labels. KITTI 2012 contains 194 training stereo image pairs with sparse ground-truth disparities and another 195 testing image pairs without ground-truth disparities. All these images have no semantic labels.

Cityscapes [11]: This is a large dataset of stereo image pairs focusing on urban street scenes. It contains 1,525 stereo image pairs for testing with ground-truth disparities precomputed using SGM.

Some metrics are used to evaluate the stereo matching performance. The measure of averaged end-point error (EPE) is defined by $EPE(d^* - \hat{d}) = ||d^* - \hat{d}||_2$. A pixel is considered to be an erroneous pixel when its disparity error is larger than t pixels, and the percentages of erroneous pixels in non-occluded and all areas are calculated. The percentages of erroneous pixel averaged over background & foreground regions and all ground-truth pixels are measured separately. For all error metrics, the lower the better.

4.2. Model specification

We implemented the proposed SSPCV-Net based on PyTorch, and the training was done on two Nvidia 1080 GPUs with Adam (momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$). The stereo image pairs were randomly cropped into two kinds of size ($256 \times 512, 256 \times 792$) before the training stage. The maximum disparity D_{max} was set to 256 for Scene Flow and 192 for KITTI 2015 & 2012.

For Scene Flow dataset, we trained our model from scratch using the training split with a constant learning rate of 0.001 and a batch size of 2 with $\alpha = 0.9$. The semantic segmentation subnetwork within SSPCV-Net was first trained for 40 epochs, where segmentation labels were transformed from object labels, then we did the joint training of the whole network for 40 epochs.

For KITTI 2015 & 2012, the model trained with SceneFlow was used (as pretrain) for further fine-tuning on the KITTI training dataset. The learning rate for both KITTI dataset trainings began at 0.01 and was reduced at a rate of 50% every 100 epochs. The semantic segmentation subnetwork was first trained with the KITTI 2015 dataset for 300 epochs. Then we did the joint training of the whole network for 400 epochs with $\alpha = 0.9$ for KITTI 2015, but with $\alpha = 1$ (i.e., the boundary loss term L_{bdry} was excluded from the loss function) for KITTI 2012 because of no semantic ground-truth available for use in KITTI 2012 dataset.

Table 1. Comparison of a number of different model variants for justification of SSPCV-Net on SceneFlow validation dataset (*20 epochs*) and KITTI 2015 validation datasets. The percentage of pixels with errors is used for KITTI 2015 evaluation and the averaged end-point error is used for Scene Flow evaluation.

	Semantic branch	Pyramid cost volumes	Dilated convoution	Scene Flow validation	KITTI 2015 validation
Single spatial cost volume				2.12	2.63
+Semantic branch	✓			1.76	2.42
+Semantic branch (Joint-train)	✓			1.78	2.37
+Spatial pyramid cost volumes		✓		1.21	2.11
+3D multiple cost volumes	✓	✓		1.04	1.99
SSPCV-Net (excluding FFM)	✓	✓	✓	1.07	2.10
SSPCV-Net (excluding boundary loss in joint training)	✓	✓	✓	1.01	1.93
SSPCV-Net	✓	✓	✓	0.98	1.85

The overall training process took about 120 hours for the Scene Flow dataset and 70 hours for each of two KITTI datasets. In our experiments, the Cityscapes dataset is only used to evaluate the *generalization ability* of the network.

4.3. Ablation studies

We first conducted ablation studies to compare a number of different model variants for SSPCV-Net on the Scene Flow dataset and the KITTI 2015 dataset (*without pretraining from Scene Flow*), respectively. For KITTI 2015, we divided the origin training set into a training split (80%) and a validation split (20%) since the original testing set has no disparity ground-truth provided. Importance of three key ideas in SSPCV-Net was evaluated: 1) adding semantic branch, 2) using pyramid cost volumes and 3) dilated convolution in feature extraction. The results are reported in Table 1 and clearly justify our design choices for SSPCV-Net: pyramid cost volumes and the semantic information can promote the accuracy of disparity estimation, and the feature extraction has been improved when the dilated convolution strategy was used in the network.

Some disparity maps regressed from SSPCV-Net by excluding certain cost volume of different branches or levels are illustrated in the Figure 5. The lowest-level spatial cost volume helps improve the accuracy in small objects region and the highest-level spatial cost volume contains more context information and helps detect more scenes. The semantic cost volume helps produce better edge and better shape cues. Finally, SSPCV-Net possesses all advantages from the semantic cost volume and spatial pyramid cost volumes.

4.4. Comparisons with some existing networks

We compared the performance of SSPCV-Net with some state-of-the-art networks for stereo matching, including MC-CNN [41], DispNet v2 [15], iResNet-i2 [24], GC-Net [22], CRL [30], PSMNet [4], EdgeStereo [35], and Segstereo [38].

On Scene Flow – As reported in Table 2 for performance evaluation results on the Scene Flow dataset, SSPCV-Net

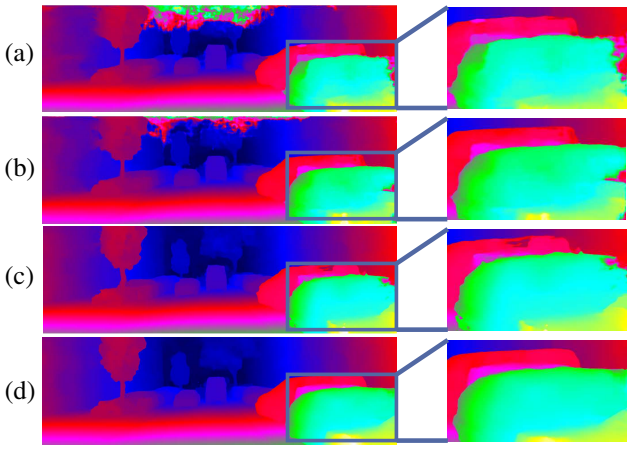

Figure 5. Disparity maps resulting from SSPCV-Net by excluding certain cost volume of different branches or levels. (a) Without the lowest-level spatial cost volume; (b) without the highest-level spatial cost volume; (c) without the semantic cost volume; (d) from the full-version SSPCV-Net.

obtained the best averaged EPE (0.87) and 3-pixel error in all pixels (D1-all) for all regions (3.1) and significantly outperformed all comparison methods in term of accuracy. The predicted disparity maps and corresponding errors of two examples by SSPCV-Net are illustrated together with the disparity maps by PSMNet in Figure 6, which visually demonstrates that SSPCV-Net can reach more accurate disparity maps especially at the edge of the objects.

On KITTI 2015 – Table 3 reports the performance evaluation results on the KITTI 2015 online leaderboard (by the KITTI evaluation server), in which the 3-pixel errors in estimated pixels (D1-est), background (D1-bg), foreground (D1-fg) and all pixels (D1-all) for all regions (ALL) and non-occluded regions (NOC) are computed. Clearly, SSPCV-Net achieved the best performance in terms of almost all error metrics except for the NOC D1-fg metric among all comparison methods. The leaderboard ranks the overall performance based on the ALL D1-all metric, and SSPCV-Net obtained 2.11%, which is much better than

Table 2. Results of the performance comparison on Scene Flow dataset.

Method	MC-CNN	GC-Net	iResNet-i2	CRL	PSMNet	EdgeStereo	SegStereo	SSPCV-Net
Averaged EPE	3.79	1.84	1.40	1.32	1.09	1.11	1.45	**0.87**
D1-all	-	9.7	5.0	6.7	4.2	-	3.5	**3.1**

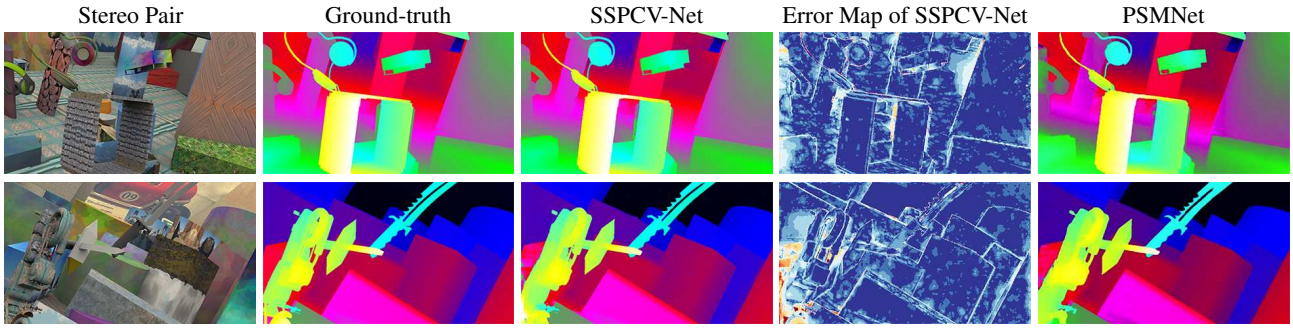

Figure 6. Two testing results from Scene Flow dataset. From left to right: the left input image of stereo image pair, the ground-truth disparity, the predicted disparity map by SSPCV-Net, the error map of SSPCV-Net prediction and the predicted disparity map by PSMNet.

Table 3. Results of the performance comparison on the KITTI 2015 dataset.

Method	ALL				NOC			
	D1-est	D1-bg	D1-fg	D1-all	D1-est	D1-bg	D1-fg	D1-all
MC-CNN [41]	3.88	2.89	8.88	3.89	3.33	2.48	7.64	3.33
DispNet v2 [15]	3.43	3.00	5.56	3.43	3.09	2.73	4.95	3.09
GC-Net [22]	2.87	2.21	6.16	2.87	2.61	2.02	5.58	2.61
CRL [24]	2.67	2.48	**3.59**	2.67	2.45	2.32	**3.12**	2.45
EdgeStereo [35]	2.59	2.27	4.18	2.59	2.40	2.12	3.85	2.40
PSMNet [4]	2.32	1.86	4.62	2.32	2.14	1.71	4.31	2.14
SegStereo [38]	2.25	1.88	4.07	2.25	2.08	1.76	3.70	2.08
SSPCV-Net	**2.11**	**1.75**	3.89	**2.11**	**1.91**	**1.61**	3.40	**1.91**

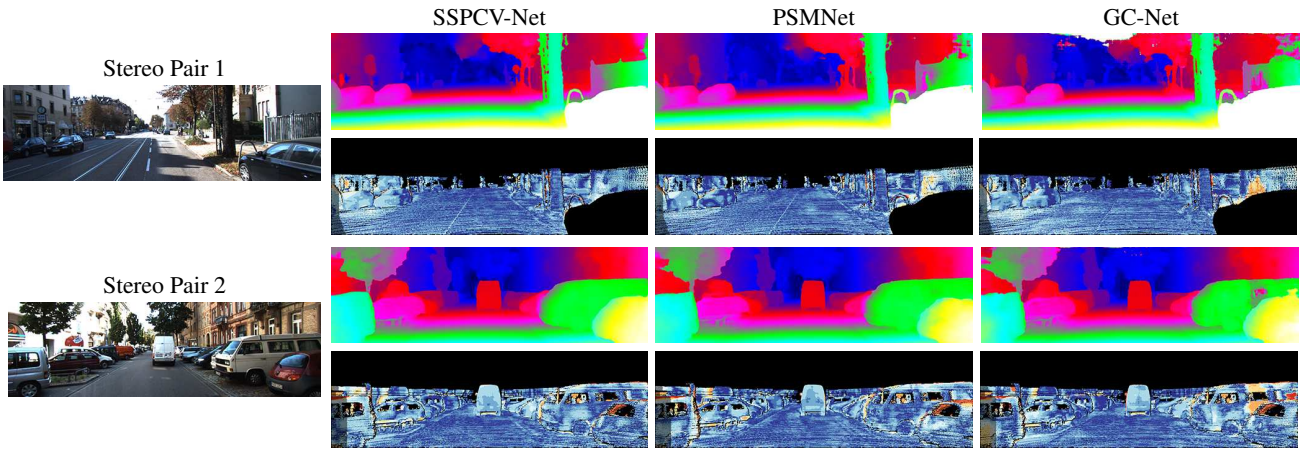

Figure 7. Two testing results from KITTI 2015 dataset. The left panel shows the left image of the input stereo image pair, and for each input image pair, the predicted disparity and corresponding error maps obtained by SSPCV-Net, PSMNet and GC-Net are presented.

other stereo matching networks. Moreover, we evaluated the semantic sub-network on KITTI 2015 and got average IoU of 56.43% for each class and 82.21% for each category.

For visual illustration, Figure 7 presents three examples of the disparity maps estimated by SSPCV-Net, PSMNet and GC-Net with the corresponding error maps.

Table 4. Results of the performance comparison on KITTI 2012 dataset.

Method	2px		3px		4px		5px	
	Out-Noc	Out-All	Out-Noc	Out-All	Out-Noc	Out-All	Out-Noc	Out-All
MC-CNN [41]	3.90	5.45	2.43	3.63	1.90	2.85	1.64	2.39
GC-Net [22]	2.71	3.46	1.77	2.30	1.36	1.77	1.12	1.46
PSMNet [4]	**2.44**	3.01	1.49	**1.89**	1.12	1.42	0.90	1.15
EdgeStereo [35]	2.79	**2.43**	1.73	2.18	1.30	1.64	1.04	1.32
SegStereo [38]	2.66	3.19	1.68	2.03	1.25	1.52	1.04	1.32
SSPCV-Net	2.47	3.09	**1.47**	1.90	**1.08**	**1.41**	**0.87**	**1.14**

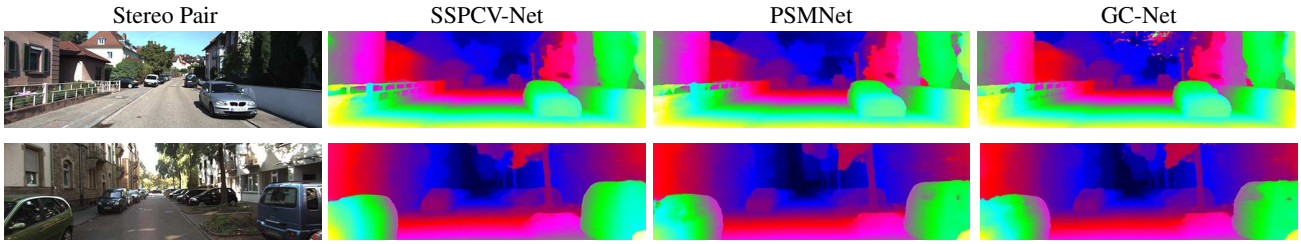

Figure 8. Two testing results from KITTI 2012 dataset. The left panel shows the left image of the input stereo image pair, and for each input image pair, the disparity maps obtained by SSPCV-Net, PSMNet and GCNet are presented.

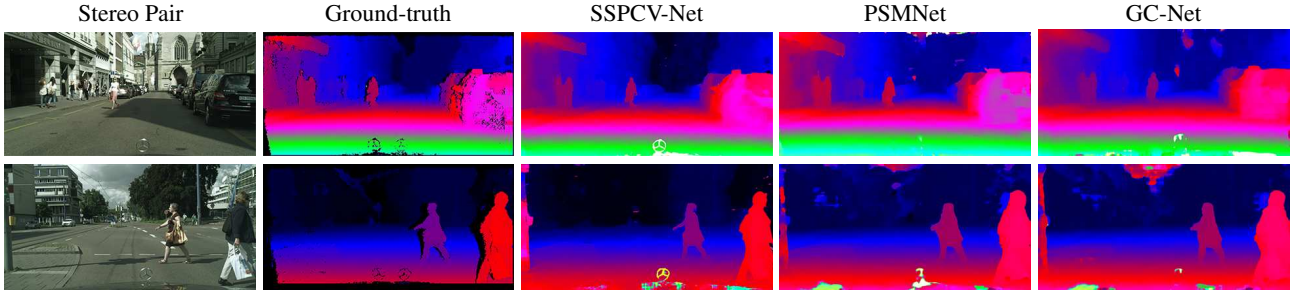

Figure 9. Two testing results from Cityscapes dataset by SSPCV-Net, PSMNet and GC-Net on the generalization ability.

On KITTI 2012 – Table 4 reports the performance evaluation results on the KITTI 2012 online leaderboard, in which the 2, 3, 4 and 5 pixel errors in all regions (Out-All) and non-occluded regions (Out-Noc) are evaluated. Although the boundary loss term was excluded from the loss function for joint training, in this case, SSPCV-Net still achieved the best performance in five error metrics out of a total of eight among all comparison methods, and did just very slightly worse than PSMNet in two and EdgeStereo in one of the remaining three error metrics. Figure 8 visually illustrates two examples of the predicted disparity maps produced by SSPCV-Net, PSMNet and GC-Net, and it again shows SSPCV-Net can give more reliable and accurate results, especially on ambiguous regions.

On Cityscapes – To evaluate the generalization ability, we used the test split of Cityscapes to test the models which were all trained on Scene Flow and KITTI 2015 (without any training on Cityscapes dataset). Note that the channel of cost volumes for all compared methods was set to be 16 in experiments. Figure 9 shows two examples of the disparity maps estimated by SSPCV-Net, PSMNet and GC-Net for visual comparison, which show SSPCVT-Net significantly outperformed PSMNet and GC-Net on the generalization ability. Predictions by the proposed SSPCVT-Net are able to capture the global layout and object details (shape & edge) quite well.

5. Conclusion

In this paper, we developed a new semantic stereo network of SSPCV-Net, in which pyramid cost volumes are constructed for describing semantic and spatial information in multiple levels and a 3D multi-cost aggregation module is proposed to integrate the extracted multilevel features. Comprehensive experiments on Scene Flow, KITTI 2015 and 2012, and Cityscapes stereo datasets demonstrated that the proposed SSPCV-Net can significantly improve the accuracy and generalization ability of stereo matching over many existing state-of-the-art neural networks.

References

[1] Joydeep Biswas and Manuela Veloso. Depth camera based localization and navigation for indoor mobile robots. In *RGB-D Workshop at RSS*, volume 2011, page 21, 2011.

[2] Michael Bleyer, Carsten Rother, Pushmeet Kohli, Daniel Scharstein, and Sudipta Sinha. Object stereo–joint stereo matching and object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[3] Matthew Brown, Gang Hua, and Simon Winder. Discriminative learning of local image descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):43–57, 2011.

[4] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[5] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[8] Qiuyu Chen, Ryoma Bise, Lin Gu, Yinqiang Zheng, Imari Sato, Jenq-Neng Hwang, Sadakazu Aiso, and Nobuaki Imanishi. Virtual blood vessels in complex background using stereo x-ray images. In *IEEE International Conference on Computer Vision Workshops*, pages 99–106, 2017.

[9] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[10] Xinjing Cheng, Peng Wang, and Ruigang Yang. Learning depth with convolutional spatial propagation network. *arXiv preprint arXiv:1810.02695*, 2018.

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361. IEEE, 2012.

[13] Spyros Gidaris and Nikos Komodakis. Detect, replace, refine: Deep structured prediction for pixel wise labeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[14] Ross Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[15] Fatma Guney and Andreas Geiger. Displets: Resolving stereo ambiguities using object knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[17] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2008.

[18] Heiko Hirschmuller and Daniel Scharstein. Evaluation of cost functions for stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007.

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[20] Jianbo Jiao, Ying Cao, Yibing Song, and Rynson Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *European Conference on Computer Vision (ECCV)*, pages 53–69, 2018.

[21] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7482–7491, 2018.

[22] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[23] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. *arXiv preprint arXiv:1807.08865*, 2018.

[24] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Linbo Qiao, Wei Chen, Li Zhou, and Jianfeng Zhang. Learning deep correspondence through prior and posterior feature constancy. *arXiv preprint arXiv:1712.01039*, 2017.

[25] Di Lin, Yuanfeng Ji, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Multi-scale context intertwining for semantic segmentation. In *European Conference on Computer Vision (ECCV)*, pages 603–619, 2018.

[26] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1925–1934, 2017.

[27] Wenjie Luo, Alexander G. Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[28] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[29] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[30] Jiahao Pang, Wenxiu Sun, Jimmy SJ Ren, Chengxi Yang, and Qiong Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *IEEE International Conference on Computer Vision Workshops*, volume 7, 2017.

[31] Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Geometry meets semantics for semi-supervised monocular depth estimation. In *Asian Conference on Computer Vision (ACCV)*, pages 298–313. Springer, 2018.

[32] Anurag Ranjan and Michael J. Black. Optical flow estimation using a spatial pyramid network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[33] Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Deepmatching: Hierarchical deformable dense matching. *International Journal of Computer Vision*, 120(3):300–323, 2016.

[34] Amit Shaked and Lior Wolf. Improved stereo matching with constant highway networks and reflective confidence learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[35] Xiao Song, Xu Zhao, Hanwen Hu, and Liangji Fang. Edgestereo: A context integrated residual pyramid network for stereo matching. In *Asian Conference on Computer Vision (ACCV)*, 2018.

[36] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[37] Federico Tombari, Stefano Mattoccia, Luigi Di Stefano, and Elisa Addimanda. Classification and evaluation of cost aggregation methods for stereo correspondence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008.

[38] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic information for disparity estimation. In *European Conference on Computer Vision (ECCV)*, September 2018.

[39] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *European Conference on Computer Vision (ECCV)*, September 2018.

[40] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations (ICLR)*, 2016.

[41] Jure Zbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(1-32):2, 2016.

[42] Chi Zhang, Zhiwei Li, Yanhua Cheng, Rui Cai, Hongyang Chao, and Yong Rui. Meshstereo: A global stereo model with mesh alignment regularization for view interpolation. In *IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[43] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. Joint task-recursive learning for semantic segmentation and depth estimation. In *European Conference on Computer Vision (ECCV)*, pages 235–251, 2018.

[44] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.