SGNet: Semantics Guided Deep Stereo Matching

Shuya Chen[1][0000−0003−4114−9066], Zhiyu Xiang[2*][0000−0002−3329−7037], Chengyu Qiao[1][0000−0003−2413−6837], Yiman Chen[1][0000−0001−6809−4291], and Tingming Bai[1][0000−0002−1516−1558]

[1] College of Information and Electronic Engineering, Zhejiang University, Hangzhou, China
[2] Zhejiang Provincial Key Laboratory of Information Processing, Communication and Networking, Zhejiang University, Hangzhou, China
xiangzy@zju.edu.cn

Abstract. Stereovision has been an intensive research area of computer vision. Based on deep learning, stereo matching networks are becoming popular in recent years. Despite of great progress, it's still challenging to achieve high accurate disparity map due to low texture and illumination changes in the scene. High-level semantic information can be helpful to handle these problems. In this paper a deep semantics guided stereo matching network (SGNet) is proposed. Apart from necessary semantic branch, three semantic guided modules are proposed to embed semantic constraints on matching. The joint confidence module produces confidence of cost volume based on the consistency of disparity and semantic features between left and right images. The residual module is responsible for optimizing the initial disparity results according to its semantic categories. Finally, in the loss module, the smooth of disparity is well supervised based on semantic boundary and region. The proposed network has been evaluated on various public datasets like KITTI 2015, KITTI 2012 and Virtual KITTI, and achieves the state-of-the-art performance.

1 Introduction

As a low cost 3D sensing module, stereo vision is widely used in lots of applications like robot navigation [1] or unmanned vehicles [2]. The main challenge of stereo vision lies in stereo matching, i.e., obtaining an accurate disparity map for the scene given a pair of stereo images.

Although the stereo matching performance has been greatly improved in recent years due to the development of deep learning [3, 4], there still exists problems caused by the lack of reliable scene clues, large changes in illumination, object occlusion or low texture. As a bio-prototype of stereo vision, our eyes can judge the distance of objects by combining multiple clues, such as object categories, the integrity of objects, the judgment of foreground/background and so on. To some extent, it is capable of fusing more global semantic information to help the distance determination.

In computer vision, semantic segmentation is an important high-level task in scene understanding [5, 6]. The goal is to assign a correct category label to each

pixel in the image and provide a high level understanding of the scene. These years have seen some successful fusion of semantics into different low level tasks such as optical flow [7], monocular depth estimation [8–10], depth completion [11] and stereo matching [12–14], etc. Generally, there are two ways to incorporate semantics. Fusion of convolutional semantic features is the most popular way [7–9, 12–14], since semantic features are heterogeneous high level ones different from low level features, it can be complementary for stereo matching. The second way uses the geometric layout of semantic results [9, 10, 14], since both of semantic and disparity maps can display the general layout and object boundaries.

In this paper, we present a novel semantics guided deep stereo matching network (SGNet) with new semantic constraints embedded. As shown in Fig. 1, the entire model is built upon PSMNet [15], a mature 3D cost volume based deep stereo matching network. We design three novel modules which embed semantic feature consistency, semantic label based optimization and semantic smooth priori into disparity computation, respectively. Based on the observation that if two pixels in the stereo pair don't have the same semantic labels, they are unlikely to be the correct matching, a confidence module is designed. This module computes the consistency between the correlation obtained from the semantic and disparity features upon input images and takes it as the confidence level of the disparity cost volume. Then a residual module is proposed to further optimize the initial disparity results from the regression step. In this module, the initial disparity map is divided into multiple channels according to their semantic categories where depthwise convolution is adopted to obtain a semantic-dependent disparity residual. Finally, with a priori similarity between the semantics and disparity layout, two loss functions are proposed in the loss module to guide the smooth of disparity under the semantic supervision.

In summary, our contributions are:

(1) A novel semantic guided deep stereo matching network with residual based disparity optimization structure is proposed. Within the residual module, semantics based depthwise convolution operation is presented to obtain category-dependent disparity residual in order to refine the initial disparity;

(2) A confidence module based on semantic-disparity consistency is proposed to help the initial disparity regression. Two improved loss functions based on the similarity between semantic and disparity map are also presented to guide the smooth of disparity prediction;

(3) The entire module is implemented end-to-end and comprehensively evaluated in various public datasets. The experimental results achieve state-of-the-art performance, demonstrating the success of our semantic guidance policy.

2 Related Work

Traditional stereo matching algorithms usually consist of four steps [16]: matching cost computation, cost aggregation, disparity computation and refinement. Local algorithms like SSD [16] and SAD [17] aggregate the cost within windows and select the disparities with minimal cost. Global algorithms construct a

global cost function to seek final disparities instead of using the aggregated cost, some examples are graph cuts [18] and belief propagation [19]. However, these methods need hand-crafted features and are still limited in ill-posed regions.

In contrast, learning based methods have a strong feature representation capability to deal with the problems in stereo matching. Early deep learning algorithms [3, 4] only replace parts of steps in traditional pipelines. Different from these approaches which still require human involvement, end-to-end algorithms are becoming more popular in recent years. The structure of end-to-end networks can be roughly divided into two categories according to the different forms of computing matching cost, i.e., correlation based and 3D cost volume based methods. The former requires less memory and can directly obtain the similarity between extracted feature maps. The latter preserves more complete features and can usually achieve better performance.

Correlation Based Methods. FlowNet [20] first introduces the correlation layer which directly calculates the correlation between two images by inner product and demonstrates its success in optical flow computation. Upon FlowNet [20], DispNet [21] is proposed for the task of stereo matching. Based on DispNet [21], CRL [22] constructs a two-stage model, the initial disparity is corrected by using the residual multiscale signal in the second stage. AANet [23] adopts multiscale correlation layers and proposes intra-scale and cross-scale modules to further refine the disparity.

3D Cost Volume Based Methods. GC-Net [24] manages to employ contextual information with 3D cost volume and adopts 3D convolution to regress disparity directly. Yu et al. [25] propose a learnable matching cost volume. PSMNet [15] introduces a pyramid pooling module to incorporate global and local features, and adopts stacked hourglass structures to regularize cost volume. To further fuse the constraints of image edges, EdgeStereo [26, 27] trains an edge detection sub-network to refine the disparity by concatenating the edge features and the edge-aware loss. GA-Net [28] designs a semi-global aggregation layer and a local guided aggregation layer inspired by SGM [29]. HSM [30] proposes a network for high-res images by a coarse-to-fine hierarchy and adopts three asymmetric augmentation about imaging condition, calibration and occlusion.

Our method also takes advantages of both correlation and 3D cost volume presentation. We construct 3D cost volume to regress the initial disparity and adopt the correlation layer to embed the constraints from semantics.

Methods Combined with Semantics. Semantic segmentation is a high-level pixel-wise classification task, which can provide valuable semantic information to many low level tasks. It has been successfully incorporated into many tasks, e.g., optical flow [7], depth estimation [8–10] and depth completion [11].

However, few semantics-combined stereo matching networks are published until recently. SegStereo [12] concatenates semantics with correlation features. DispSegNet [13] proposes an unsupervised algorithm which concatenates semantics with initial disparity to refine the prediction. SSPCV-Net [14] constructs heterogeneous and multiscale cost volumes combined with semantic features and proposes a 3D multicost aggregation module. It also uses gradient-related loss

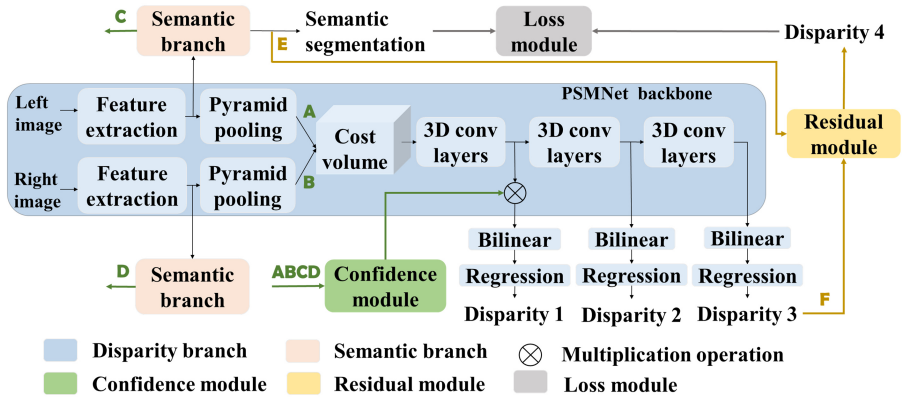

Fig. 1. Architecture of the proposed SGNet.

to constrain the smooth of disparity. These methods mainly embed semantics from the aspects of semantic features and geometric layouts.

Our solution utilizes the consistency between the semantic and disparity correlation as the cost confidence instead of directly adopting the concatenation operation to fuse the semantic features, as in [12] and [13]. In addition, we explicitly use the semantic category in our residual module by dividing the disparity map into multiple category channels and compute the category-dependent disparity residual. In spired by EdgeStereo [27] and SSPCV-Net [14] which adopt gradient-related loss to guide disparity prediction, we further supervise the smooth of disparity by boundary and inner region of semantic maps under the specialized category-dependent constraints.

3 Approach

3.1 Architecture Overview

The architecture of the proposed SGNet for stereo matching is illustrated in Fig. 1. It's built upon PSMNet [15], which uses 3D cost volume to compute stereo matching and possesses competitive performance. Given the left and right images, this baseline first gets high-dimensional features by the weight-sharing encoder layers. Then a pyramid pooling layer is followed to obtain the multiscale features and fuse them. The features are then translated in horizontal (x-disparity) direction and concatenated to construct 3D cost volume with different disparities. Disparity output is obtained by regression with stacked hourglass structures. Three outputs with gradually supervised refinement are used, each of which is the sum of the disparity value weighted by predicted probabilities. In [15], the last output, i.e., $disp3$ in Fig. 1, is regarded as the final output.

The baseline PSMNet [15] is taken as the disparity branch in our SGNet. Apart from it, a semantic branch is added after the feature extraction layer. It

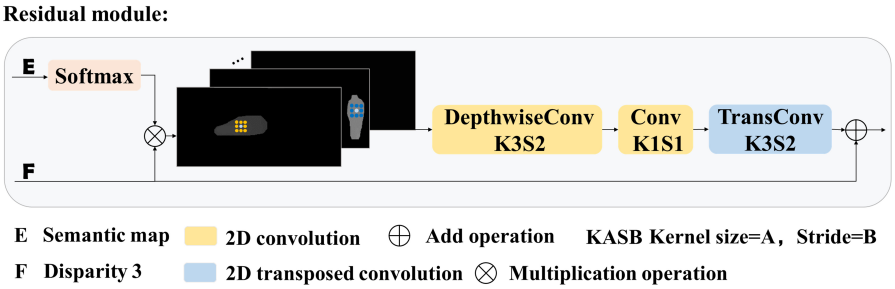

Fig. 2. Structure of the residual module.

shares some shallow layers with the disparity branch and has independent high-level layers so that the unique characteristics of semantics can be extracted.

To embed semantic guidance into disparity computing process, three novel modules are added to the baseline network, i.e., residual module, confidence module, and loss module. The residual module takes $disp3$ as initial input, which is divided into multiple channels according to the semantic categories and further optimized by category-dependent convolution. The confidence module is combined with the $disp1$'s cost volume. It takes the semantic and disparity features from left and right images and produces the consistency confidence between the semantic and disparity correlation. This confidence will help better regress initial disparities. The loss module takes the final prediction $disp4$ and semantics as input and is responsible for embedding the semantics boundary and inner smooth constraints on supervisory loss.

3.2 Semantic Branch

Semantic segmentation and disparity networks share the shallow layers. The independent semantic branch employs two more residual blocks with 256 channels for deeper feature extraction. Similar to disparity branch, pyramid pooling module is also used for acquisition of local and global semantic features. Finally, the semantic results are obtained through a classification layer. Such structure has two advantages: (1) the disparity and semantics share the shallow layers making the network more efficient on learning the common features of two tasks. (2) The structure of semantic branch is similar to the disparity branch, so the whole network structure is more unified.

3.3 Residual Module

The motivation of this module is from the observation that the disparities within an object or category are mostly smooth. However, for different categories the degree of smooth may not be the same. For example, the surface of the road is generally flat with smooth, while the surface of other categories like trees may

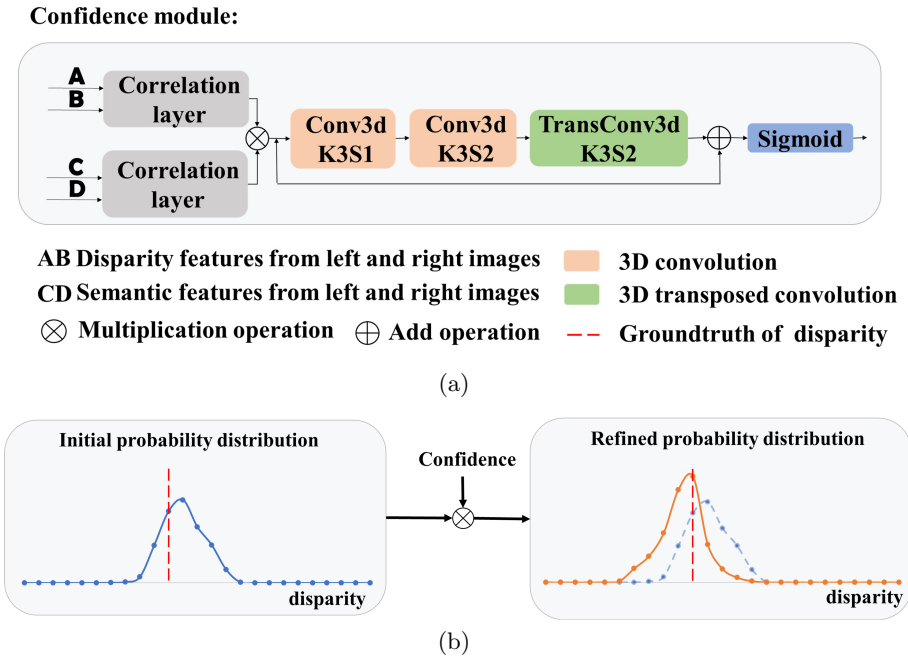

Fig. 3. Illustration of (a) network structure and (b) effects of the confidence module.

be much more uneven. Therefore, carrying out different convolution operations according to their semantic categories may help better learn the continuity of disparity for each category.

The residual module takes the semantic probability map and $disp3$ as input. $disp3$ with size $H \times W$ is multiplied with the semantic probability map with size $H \times W \times C$, where C is the number of classes. Since the semantic probabilities are between 0 and 1 representing the possibilities of each category, the operation results in category-wise raw disparity map with size $H \times W \times C$ for optimization. In other words, each channel of the map is the raw disparity under a certain category, as illustrated in Fig. 2.

Depthwise convolution is then performed for each channel. A pointwise convolution is followed to integrate all category channels. Finally a transposed convolution is used to compute the disparity residual.

3.4 Confidence Module

Given the disparity, we can obtain the corresponding pixel pairs between two images. The semantic correlation can be considered as a kind of constraint. That is, as for one pair, if these two pixels don't belong to the same category, then they are unlikely to be the correct matching points. So we propose a confidence module to embed this constraint. The module computes the consistency between

the correlation on disparity and semantics and takes it as the confidence of $disp1$'s cost volume. Only when both of disparity and semantic correlations are high, the confidence value of the corresponding disparity candidate is high.

The network structure of confidence module is shown in Fig. 3(a). The semantic features as well as the disparity features with size of $H/4 \times W/4$ are fed into the correlation layer respectively to compute the correlation for each candidate disparity. The operation of correlation [20] is shown in Eq. 1.

$$Correlation(x, y, d) = \frac{1}{N_c} inner < f_1(x, y), f_2(x - d, y) > \ . \tag{1}$$

where $inner <>$ denotes the inner product operation, N_c is the number of feature channels, f_1 and f_2 denote the left and right feature maps. Compared with 3D cost volume, this operation intuitively shows the matching degree of features under different disparities. We apply this operation on disparity or semantic features respectively.

The output of the correlation layers in Fig. 3(a) represent semantic and disparity correlations respectively. They are then multiplied and fed into three consecutive 3D convolution layers with a residual structure to compute the consistency between these two different correlations. Finally a sigmoid function is employed to constrain the range of the output confidence.

As shown in Fig. 1, there are three levels of regression output in PSMNet [15]. Multiplying the obtained confidence values to $disp1$'s cost volume is a good choice because it's better to improve the disparity in the early stage. The effect of the confidence module on the disparity regression is illustrated in Fig. 3(b). The initial cost volume or the probability distribution along disparity dimension of a pixel may be inaccurate, with the probability peak drifted from the ground truth. After the correction from the confidence, the distribution can be adjusted and closer to the ground truth.

3.5 Loss Module

Loss for Single Tasks. For disparity estimation, we mainly use *smooth* L_1 loss since it is insensitive to outliers and the gradient change is relatively small.

$$L_{disp} = \frac{1}{N} \sum_{i=1}^{N} smooth(d_i, d_i^*) \ . \tag{2}$$

In Eq. 2, d_i and d_i^* are the predicted disparity and corresponding groundtruth respectively, N is the number of valid pixels. During training, we supervise $disp1$, $disp2$, $disp4$, and output $disp4$ during testing, as shown in Fig. 1.

As for semantic loss L_{sem}, we adopt the cross-entropy loss to train the semantic branch as shown in following:

$$L_{sem} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y(i, c) log(p(i, c)) \ . \tag{3}$$

(a) (b) (c)

Fig. 4. Illustration of layout similarity of semantics (b) and disparity image (c) given an example scene from KITTI [31] (a). Although they have similar smooth layout, there are some inconsistent areas between the two images. For Example, the red boundary in (b) and yellow boundary in (c) do not appear in the counterpart image.

where p is the predicted semantic probability, y is the corresponding groundtruth, C is the number of classes and N is the number of valid pixels.

Semantic Guided Disparity Loss. Semantic and disparity map are to some extent similar in geometric layouts. Therefore, the semantic ground truth can be utilized to guide the disparity map with a reasonable smooth constraint.

(1) Guided by Semantic Boundary

For most of the semantic foreground, such as vehicles, pedestrians or other objects, if semantic boundaries exist, so does disparity map. But it doesn't hold everywhere. For example, for parts of the background areas, such as "road", "sidewalk" or "parking", there are no obvious boundaries between two adjacent categories in disparity map, as shown in Fig. 4(b). The semantic boundary between the road and parking marked by red dashed lines doesn't appear in the disparity map. Therefore, we only supervise the disparity boundaries which has strong co-appearing relationships with semantics according to their categories.

We tend to punish on the pixels whose positions in ground truth semantics are boundary areas while in the predicted disparity map are not. The loss function motivated by [14] is constructed in Eq. 4, where a mask m_b is used to handle the valid area.

$$L_{bdry} = \frac{1}{N}\left(\sum_{i,j}|\varphi_x^2\left(sem_{i,j,m_b}\right)|e^{-|\varphi_x^2(d_{i,j,m_b})|} + \sum_{i,j}|\varphi_y^2\left(sem_{i,j,m_b}\right)|e^{-|\varphi_y^2(d_{i,j,m_b})|}\right). \tag{4}$$

In Eq. 4, sem_{i,j,m_b} and d_{i,j,m_b} are the semantic groundtruth and predicted disparity at $p(i,j)$ with $m_b(i,j)$=1 respectively, φ_x^2 and φ_y^2 are the second-order gradient along the horizontal and vertical direction respectively, N is the number of valid pixels whose $m_b(i,j) = 1$.

Taking images in KITTI 2015 dataset as examples, the mask can be defined as:

$$m_b\left(i,j\right) = \begin{cases} 1, & p\left(i,j\right) \notin \text{“}road\text{”}, \text{“}sidewalk\text{”}, \text{“}vegetation\text{”}, \text{“}terrain\text{”}, \\ 0, & otherwise. \end{cases} \tag{5}$$

(2) Guided by Semantic Smooth

Smooth constraint in ground truth semantics can be another powerful supervision for disparity prediction. However there are also inconsistent areas where the semantics are smooth while the disparity are not. As shown in Fig. 4(c), the object boundary represented by yellow dashed line doesn't exist in the semantic map.

Our solution is a threshold based method. When the gradient of disparity is greater than a threshold λ, it indicates that there exist true boundaries so we don't enforce the semantic smooth constraint on it. In Eq. 6, we mostly punish on the pixels which are smooth in the semantic map while unsmooth in the disparity map, with the mask m_s to define the valid area.

$$L_{sm} = \frac{1}{N}\left(\sum_{i,j}|\varphi_x^2\left(d_{i,j,m_s}\right)|e^{-|\varphi_x^2(sem_{i,j,m_s})|} + \sum_{i,j}|\varphi_y^2\left(d_{i,j,m_s}\right)|e^{-|\varphi_y^2(sem_{i,j,m_s})|}\right) . \tag{6}$$

$$m_s\left(i,j\right) = \begin{cases} 1, & gradient\ of\ disparity\ map\ at\ p\left(i,j\right)\ less\ than\ \lambda, \\ 0, & otherwise. \end{cases} \tag{7}$$

Finally, integrating all of the above loss together, the total loss in our model is:

$$L = L_{disp} + w_{sem} * L_{sem} + w_{bdry} * L_{bdry} + w_{sm} * L_{sm} \ . \tag{8}$$

with

$$L_{disp} = w_{disp1} * L_{disp1} + w_{disp2} * L_{disp2} + w_{disp4} * L_{disp4} \ . \tag{9}$$

4 Experiments and Analysis

Our SGNet model is implemented by Pytorch [32] and we adopt Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) to optimize the model. The learning rate is first set to 0.001 then set to 0.0001 in the later stage. The batch size is 2 in training and validation process due to the limited GPU resources. Input images are randomly cropped to size 160×320 for Virtual KITTI and 256×512 for the other datasets, and the maximum disparity is set to 192. Some parameters in loss function are set according to [15] as $w_{disp1} = 0.5$, $w_{disp2} = 0.7$, $w_{disp4} = 1$, $w_{sem} = 1$, and w_{bdry}, w_{sm}, λ are determined experimentally.

During training or testing, following datasets are used:

(1) Scene Flow [21]

Scene Flow [21] is a synthetic dataset consisting of 35454 training images and 4370 testing images with dense ground truth disparity maps. In our experiments, this dataset is only used for pre-training. Since it doesn't contain semantic labels, only the stereo matching branch is pre-trained.

(2) KITTI stereo 2015 & 2012 [31, 33]

Table 1. Results of different combining operations in confidence module. "Disp-cor" and "Sem-cor" denotes the correlation of disparity features and semantic features respectively.

Combination mode	$3px$ (%)	EPE (pixel)
Disp-cor + Seg-cor	1.362	0.6269
Disp-cor × Seg-cor	**1.299**	**0.6198**
Disp-cor × Disp-cor × Seg-cor	1.319	0.6212
Disp-cor × Seg-cor × Seg-cor	1.309	0.6238

Table 2. Results of different setting in loss module on KITTI 2015 validation set.

w_{bdry}	w_{sm}	λ	$3px$ (%)	EPE (pixel)
0.5	0.5	2	1.330	0.6274
0.5	0.5	3	**1.299**	**0.6198**
0.5	0.5	4	1.326	0.6226
0.7	0.5	3	1.336	0.6252
0.5	0.7	3	1.336	0.6235

KITTI [31, 33] are the datasets with real-world street views which use lidar to obtain sparse groundtruth disparity. KITTI 2015 [31] includes 200 training images with semantic labels and 200 testing images. In the ablation experiments, 160 images are taken as the training set, and the remaining 40 images are taken as the validation set. KITTI 2012 [33] includes 194 training images without semantic labels and 195 testing images.

(3) Virtual KITTI [34]

Virtual KITTI 2 [34] is a dataset of virtual urban scenes that contains dense depth maps and semantic labels. The "15-deg-left" subsequence in sequence 2 is sampled as the validation set with 233 images, and the same subsequences in the remaining sequences are used as the training set which includes 1893 images.

In the main ablation experiments, our model is pretrained on the Scene Flow dataset then finetuned on KITTI 2015. Since both of KITTI 2015 and 2012 are the real-world datasets with urban scenes, when submitting to the benchmark, we train the pretrained model on mixed KITTI 2015 and KITTI 2012 datasets for 500 epochs to learn the generalization of features, and finally finetune it only on KITTI 2015 or KITTI 2012 dataset for another 200 epochs. During the training, only KITTI 2015 dataset provides the semantic labels.

Averaged end-point-error (EPE) and percentage of outliers with error more than k-pixel or 5% disparity (kpx) are used as performance metrics for all of the following experiments.

Table 3. Ablation experiments of different modules on KITTI 2015 and Virtual KITTI dataset. "Baseline" refers to the disparity branch PSMNet [15], "C", "R" and "L" denote the confidence module, the residual module and the semantic guided loss module.

Model	Confidence module	Residual module	Loss module	$3px$ (%)	EPE (pixel)	Run time (s)
Scene Flow + KITTI 2015						
Baseline				1.415	0.6341	0.671
Baseline-C	✓			1.371	0.6275	–
Baseline-R		✓		1.368	0.6253	–
Baseline-CR	✓	✓		1.328	0.6203	–
Baseline-CRL	✓	✓	✓	**1.299**	**0.6198**	0.674
Virtual KITTI						
Baseline				4.108	0.6237	–
Baseline-CRL	✓	✓	✓	**3.874**	**0.5892**	–

4.1 Ablation Studies

In this ablation studies, our model is first pre-trained on the Scene Flow dataset for 15 epochs with learning rate of 0.001, then fine-tuned on KITTI 2015 training set with learning rate of 0.001 in the first 600 epochs and 0.0001 in the next 100 epochs. As for Virtual KITTI, our model is trained from scratch with learning rate of 0.001 in the first 200 epochs and 0.0001 in the next 100 epochs.

Parameter Selection. We conduct experiments on different settings or hyper parameters for the modules to determine the best configuration.

As for the confidence module, we test the way of combination between disparity and semantic correlation, as shown in Table 1. Compared with addition operation, the model with multiplication reduces the $3px$ metric from 1.362% to 1.299%. Considering the possible different weights ratio of disparity and semantics, we also test the operation of multiplying disparity or semantic correlation one more time, which only results in worse performance. So we just select a single multiplication as the correlation combination.

As for the loss, some varying choices of the weight w_{bdry}, w_{sm} and the threshold λ are shown in Table 2. When setting w_{bdry}= w_{sm}=0.5 and varying the value of λ, the results show $\lambda = 3$ is a good choice to identify the real boundaries. Given $\lambda = 3$, change w_{bdry} and w_{sm}, the best performance is obtained when w_{bdry}= w_{sm}=0.5. Therefore we set w_{bdry}= w_{sm}=0.5 and $\lambda = 3$ for all of the following experiments.

Ablation of Modules. On the baseline disparity branch, we further evaluate the effectiveness of different modules by adding them separately as shown in Table 3. When only confidence or residual module is added, $3px$ error and EPE are reduced about 0.045% and 0.007 pixels respectively. Adding both modules can further decrease the error. On both KITTI 2015 and Virtual KITTI dataset, the best performance is achieved when all of the three modules are added. We also test the inference time of "Baseline" and "Baseline-CRL" on one Nvidia

Table 4. Comparison on KITTI 2015 benchmark.

Model	All pixels			Non-occluded pixels		
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all
Models without semantics						
DispNetC [21]	4.32	4.41	4.34	4.11	3.72	4.05
MC-CNN-acrt [3]	2.89	8.88	3.89	2.48	7.64	3.33
GC-Net [24]	2.21	6.16	2.87	2.02	5.58	2.61
CRL [22]	2.48	3.59	2.67	2.32	3.12	2.45
PSMNet [15]	1.86	4.62	2.32	1.71	4.31	2.14
GwcNet-g [35]	1.74	3.93	2.11	1.61	3.49	1.92
EdgeStereo-V2 [27]	1.84	**3.30**	2.08	1.69	**2.94**	1.89
AANet+ [23]	1.65	3.96	2.03	1.49	3.66	1.85
Models with semantics						
SegStereo [12]	1.88	4.07	2.25	1.76	3.70	2.08
SSPCVNet [14]	1.75	3.89	2.11	1.61	3.40	1.91
SGNet(ours)	**1.63**	3.76	**1.99**	**1.46**	3.40	**1.78**

TITAN 1080TI. Although more modules together with a semantic branch are added, the inference time of our SGNet is still very close to the baseline.

In addition, the semantic performance of "Baseline-CRL" is evaluated with $mIoU = 48.12\%$ and $mAcc = 55.25\%$ in validation set.

4.2 Comparing with Other Methods

Results on KITTI 2015 Benchmark. We submit the best model trained on all KITTI 2015 training set to the online benchmark. The results are shown in Table 4, where "All pixels" and "Non-occluded pixels" separately represent the different range of pixels for evaluation, "D1" is the percentage of outliers with error more than 3-pixel or 5% disparity ($3px$), "bg", "fg" and "all" denote the estimated area over background, foreground and all area, respectively. State-of-the-art models with or without considering semantics are listed for comparison.

As shown in Table 4, our model achieves the lowest error in most important "D1-all" metrics on both "All pixels" and "Non-occluded pixels", surpassing the other non-semantic or semantic guided methods in the list by a notable margin. In particular, it improved about 0.3% on "D1-all" comparing with its baseline PSMNet [15], which demonstrates the effectiveness of our semantic guided policy.

Some qualitative comparisons on error map with PSMNet [15] and SegStereo [12] are shown in Fig. 5. Generally, our model produces smoother predictions with lower error across the entire image. Some noticeable improvements can be found in the area bounded by the red boxes.

Results on KITTI 2012 Benchmark. We also fine-tune the model on KITTI 2012 dataset. Since no semantic labels are provided in KITTI 2012, the semantic branch is only trained by images on KITTI 2015. We then submit the prediction

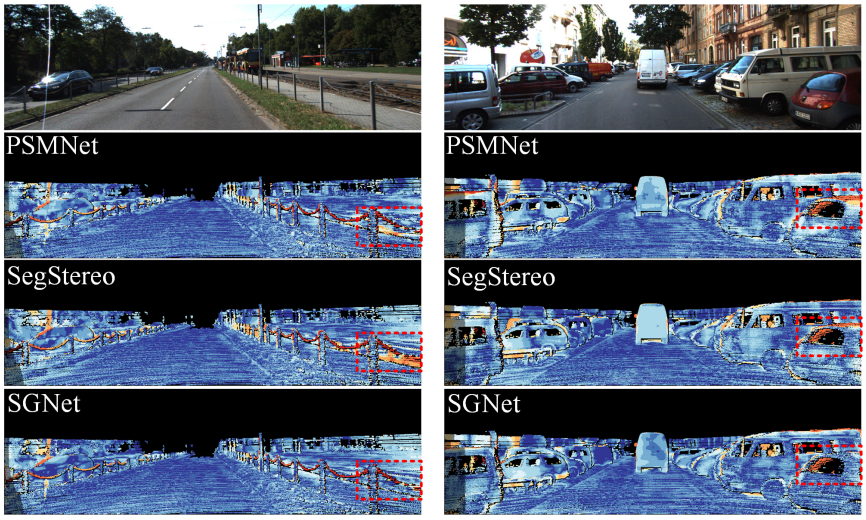

Fig. 5. Qualitative error maps on KITTI 2015 test set. The deeper blue color means the lower error, while the deeper red means the higher error. More noticeable differences can be observed in the area inside red box.

results to KITTI 2012 benchmark for evaluation. The results are shown in Table 5, where "Noc" and "All" represent the percentage of erroneous pixels only in non-occluded areas or in total, respectively.

Again, our model outperforms the baseline PSMNet [15] almost in all metrics except for $5px$ error on "All" pixels, where equal values are obtained. As expected, our model also performs better than most of the other non-semantics or semantics based method. Comparing with EdgeStereo-V2 [27], which embeds edge features and corresponding loss into the model, our method obtains better value on most non-occluded areas. We believe the performance of our model can be further improved if semantics ground truth are available in the training data.

Some qualitative comparisons with PSMNet [15] are shown in Fig. 6. Thanks to the effective guidance of semantics, our model can eliminate some holes inside objects and make the region more smoothing, as shown in the area inside the white boxes in Fig. 6.

5 Conclusion

Semantics as additional scene clues can provide valuable information for better stereo matching. In this paper we propose a semantic guided stereo matching network which optimizes the disparity computation from three semantics-related perspectives. In the confidence module, we employ the consistency between correlations on disparity and semantic features to adjust the cost volume. Within the residual module, semantics based depthwise convolution operation is presented

Table 5. Comparison on KITTI 2012 benchmark.

Model	2*px*		3*px*		4*px*		5*px*	
	Noc	All	Noc	All	Noc	All	Noc	All
Models without semantics								
DispNetC [21]	7.38	8.11	4.11	4.65	2.77	3.20	2.05	2.39
MC-CNN-acrt [3]	3.90	5.45	2.43	3.63	1.90	2.85	1.64	2.39
GC-Net [24]	2.71	3.46	1.77	2.30	1.36	1.77	1.12	1.46
PSMNet [15]	2.44	3.01	1.49	1.89	1.12	1.42	0.90	1.15
AANet+ [23]	2.30	2.96	1.55	2.04	1.20	1.58	0.98	1.30
EdgeStereo-V2 [27]	2.32	**2.88**	1.46	**1.83**	1.07	**1.34**	**0.83**	**1.04**
Models with semantics								
SegStereo [12]	2.66	3.19	1.68	2.03	1.25	1.52	1.00	1.21
SSPCVNet [14]	2.47	3.09	1.47	1.90	1.08	1.41	0.87	1.14
SGNet(ours)	**2.22**	2.89	**1.38**	1.85	**1.05**	1.40	0.86	1.15

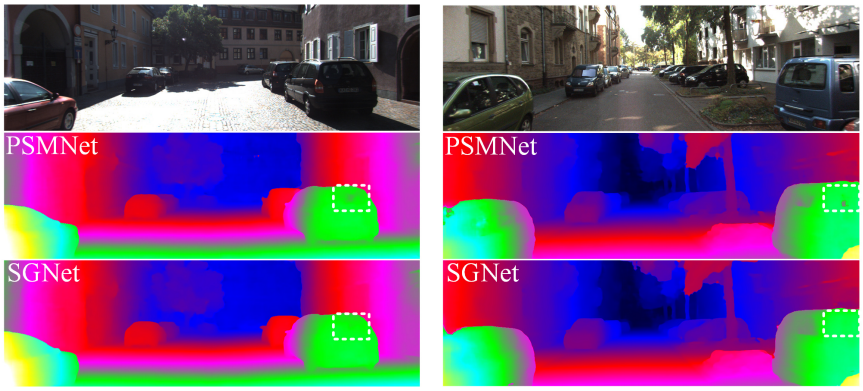

Fig. 6. The qualitative results on KITTI 2012 test set.

to obtain category-dependent disparity residual in order to refine the initial disparity. An improved loss function module based on the similarity between semantic and disparity map is also presented to guide the smooth of disparity outputs. The entire model can be trained end-to-end and run with similar computing time with the baseline. Experiments on various KITTI dataset and benchmarks are carried out and state-of-the-art performances are achieved, which demonstrate the success of our semantic guide policy.

Funding: The work is supported by NSFC-Zhejiang Joint Fund for the Integration of Industrialization and Informatization under grant No. U1709214.

References

1. Schmid, K., Tomic, T., Ruess, F., Hirschmüller, H., Suppa, M.: Stereo vision based indoor/outdoor navigation for flying robots. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE (2013) 3955–3962
2. Chen, C., Seff, A., Kornhauser, A., Xiao, J.: Deepdriving: Learning affordance for direct perception in autonomous driving. In: Proceedings of the IEEE International Conference on Computer Vision. IEEE (2015) 2722–2730
3. Žbontar, J., LeCun, Y.: Stereo matching by training a convolutional neural network to compare image patches. The journal of machine learning research **17** (2016) 2287–2318
4. Shaked, A., Wolf, L.: Improved stereo matching with constant highway networks and reflective confidence learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE (2017) 4641–4650
5. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE (2015) 3431–3440
6. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE (2017) 2881–2890
7. Cheng, J., Tsai, Y.H., Wang, S., Yang, M.H.: Segflow: Joint learning for video object segmentation and optical flow. In: Proceedings of the IEEE International Conference on Computer Vision. IEEE (2017) 686–695
8. Zhang, Z., Cui, Z., Xu, C., Jie, Z., Li, X., Yang, J.: Joint task-recursive learning for semantic segmentation and depth estimation. In: Proceedings of the European Conference on Computer Vision. (2018) 235–251
9. Jiao, J., Cao, Y., Song, Y., Lau, R.: Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In: Proceedings of the European Conference on Computer Vision. (2018) 53–69
10. Chen, P.Y., Liu, A.H., Liu, Y.C., Wang, Y.C.F.: Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE (2019) 2624–2632
11. Zou, N., Xiang, Z., Chen, Y., Chen, S., Qiao, C.: Simultaneous semantic segmentation and depth completion with constraint of boundary. Sensors **20** (2020) 635
12. Yang, G., Zhao, H., Shi, J., Deng, Z., Jia, J.: Segstereo: Exploiting semantic information for disparity estimation. In: Proceedings of the European Conference on Computer Vision. (2018) 636–651
13. Zhang, J., Skinner, K.A., Vasudevan, R., Johnson-Roberson, M.: Dispsegnet: Leveraging semantics for end-to-end learning of disparity estimation from stereo imagery. IEEE Robotics and Automation Letters **4** IEEE (2019) 1162–1169
14. Wu, Z., Wu, X., Zhang, X., Wang, S., Ju, L.: Semantic stereo matching with pyramid cost volumes. In: Proceedings of the IEEE International Conference on Computer Vision. IEEE (2019) 7484–7493
15. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE (2018) 5410–5418

16. Scharstein, D., Szeliski, R., Zabih, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision **47** (2001) 7–42
17. Kanade, T., Kano, H., Kimura, S., Yoshida, A., Oda, K.: Development of a video-rate stereo machine. In: Proceedings 1995 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human Robot Interaction and Cooperative Robots. Volume 3., IEEE (1995) 95–100
18. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence **23** IEEE (2001) 1222–1239
19. Klaus, A., Sormann, M., Karner, K.: Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In: 18th International Conference on Pattern Recognition (ICPR'06). Volume 3., IEEE (2006) 15–18
20. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. IEEE (2015) 2758–2766
21. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE (2016) 4040–4048
22. Pang, J., Sun, W., Ren, J.S., Yang, C., Yan, Q.: Cascade residual learning: A two-stage convolutional neural network for stereo matching. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. IEEE (2017) 887–895
23. Xu, H., Zhang, J.: Aanet: Adaptive aggregation network for efficient stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE (2020) 1959–1968
24. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A.: End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the IEEE International Conference on Computer Vision. IEEE (2017) 66–75
25. Yu, L., Wang, Y., Wu, Y., Jia, Y.: Deep stereo matching with explicit cost aggregation sub-architecture. In: Thirty-Second AAAI Conference on Artificial Intelligence. (2018)
26. Song, X., Zhao, X., Hu, H., Fang, L.: Edgestereo: A context integrated residual pyramid network for stereo matching. In: Proceedings of the Asian Conference on Computer Vision, Springer (2018) 20–35
27. Song, X., Zhao, X., Fang, L., Hu, H., Yu, Y.: Edgestereo: An effective multi-task learning network for stereo matching and edge detection. International Journal of Computer Vision (2020) 1–21
28. Zhang, F., Prisacariu, V., Yang, R., Torr, P.H.: Ga-net: Guided aggregation net for end-to-end stereo matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE (2019) 185–194
29. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence **30** IEEE (2008) 328–341
30. Yang, G., Manela, J., Happold, M., Ramanan, D.: Hierarchical deep stereo matching on high-resolution images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE (2019) 5515–5524

31. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Conference on Computer Vision and Pattern Recognition. (2015)
32. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems. (2019) 8024–8035
33. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition. (2012)
34. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE (2016)
35. Guo, X., Yang, K., Yang, W., Wang, X., Li, H.: Group-wise correlation stereo network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE (2019) 3273–3282