2019 International Conference on Robotics and Automation (ICRA)
Palais des congres de Montreal, Montreal, Canada, May 20-24, 2019

DSNet: Joint Learning for Scene Segmentation and Disparity Estimation

Wujing Zhan, Xinqi Ou, Yunyi Yang and Long Chen

Abstract— **Recently, research works have attempted the joint prediction of scene semantics and optical flow estimation, which demonstrate the mutual improvement between both tasks. Besides, the depth information is also indispensable for the scene understanding, and disparity estimation is necessary for outputting dense depth maps. Such task shares a great similarity with the optical flow estimation since they can all be cast into a problem of capturing the difference at a location of two image frames. However, as far as we know, currently there are few networks for the joint learning of semantic and disparity. Moreover, since deep semantic information and disparity feature maps can learn from each other, we find it unnecessary with two independent encoding modules to separately extract semantic and disparity features. Therefore, we propose a unified multi-tasking architecture DSNet, for the simultaneous estimation of semantic and disparity information. In our model, semantic features, extracted by the encoding module ResNet from the left and right images, are used to obtain the deep disparity features via a novel matching module which performs pixel-to-pixel matching. In addition, we also use the disparity map to perform warp operation on deep features of the right image to deal with the problem of lacking of semantic labels. The effectiveness of our method is demonstrated by extensive experiments.**

I. Introduction

Deep learning based semantic segmentation and disparity estimation are playing increasingly important roles in visual scene perception and understanding. Although numerous CNN models have been proposed for scene parsing [24] [4] [43] [3] and disparity estimation [26] [28] [19] [2], it is still less explored to join these tasks in a single deep network

Typical deep multi-task learning approaches mainly focused on the final prediction level via employing the cross-modal interactions to mutually refine the tasks [17] [35]. Based on those methods, PAD-Net [36] considered introducing multi-task prediction and multi-modal distillation steps at the intermediate level of a CNN to improve the target tasks. However, RGB-D, as the input of those multi-tasks net-work, is different from disparity estimation. Disparity estimation requires matching corresponding pixels on the two images along the same scan-line, meaning that the deep convolutional neuronal network requires dealing with two images at the same time. The current works [18] [10] have opened a way to the fusion of networks for deep semantic and deep optical flow. Their common denominator is to integrate deep semantic features and deep optical flow features into each other's network with their mutual communication to promote the performance of both networks. The effectiveness of their experiments demonstrates that features of deep optical flow can indeed help semantic segmentation learning, and vice versa. In general, these two methods simply connect the features of two different networks directly, which is not sufficient to mine the common features between the two tasks, such as object scales, postures and boundaries. It is well known that the optical flow estimation and disparity estimation are twin tasks and both of them capture location differences between two image frames. The disparity estimation is only required to match the difference in the horizontal direction, and the optical flow estimation also needs to match the difference in the vertical direction. Therefore, a joint approach for both disparity estimation and semantic learning in a lightweight structure could also be promising. Motivated by this, we are boldly striding forward in this paper, *i.e.*, using deep semantic features to directly study the disparity estimation. Branches in our dual-task network shown in Fig. 1 share the same ResNet module and use a down-top architecture with lateral connections to reconstruct the disparity maps and semantic maps.

The main contributions of this paper are summarized as follows. First, we propose a lightweight network called DSNet for joint disparity estimation and scene parsing. Benefited from a series of shared convolutional encoder modules, the DSNet is efficient at generating semantic and disparity information together, surpassing previous models [18] [10] that extract convolutional features for semantic segmentation and disparity estimation separately. Second, through extensive experiments, we designed an efficient matching module for the learning of semantic and disparity information, building a bridge between the two tasks. At last, we put forward a training method to effectively leverage the annotated labels of semantic and disparity information in a single network.

II. Related Work

Scene parsing The introduction of FCN [24] has opened an era for the full development of deep semantic segmentation. Typically, it consists of three parts: a top-down convolutional pooling operation, a bottom-up deconvolution enhancement, and side-by-side lateral connections. Such network runs in an end-to-end processing fashion and has achieved state-of-the-art performance at the time. In order to generate denser semantic segmentation maps, the DeepLab series [3] proposed an atrous convolutional network, which reduces the resolution while increasing the size of the pooling window to enable denser semantics. In order to capture the contextual information at multiple scales, embedding the spatial pyramid pooling module or several parallel atrous convolution with dierent rates is widely used by [43] [3] [5].

The authors are with School of Data and Computer Science, Sun Yat-sen University, Guangzhou, Guangdong, P.R.China. The corresponding author is L. Chen (chenl46@mail.sysu.edu.cn))

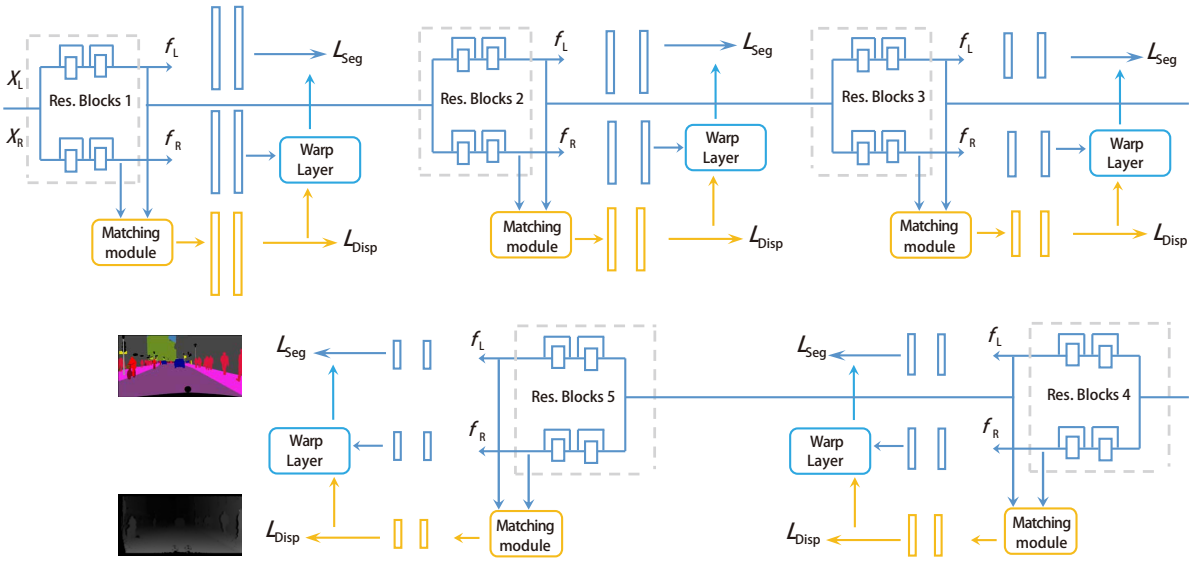

Fig. 1: An overview of the structure of DSNet.

Disparity Estimation Currently, deep learning approaches also become popular in disparity estimation research works. For instance, Mayer *et al*. proposed a network [26] for disparity estimation, it has achieved more favorable results. Based on this, [28] adopts the cascaded residual learning scheme, which uses a deeper network for the disparity estimation with a large displacement. According to the output of subnetwork, images are warped as inputs of the next shallower network for residual learning. GC-CNN [19] combines the 2D deep features corresponding to the left and right images into 3D features, and then uses the 3D convolution network to learn the disparity estimation. In order to restore the 3D tensor to the corresponding 2D tensor, Kendall *et al*. proposed to obtain the estimation result by the pixel similarity classification. Based on this, Chang *et al*. proposed a pyramid stereo matching network [2] with 3D CNN and multiple hourglass modules, and put forward a smooth disparity regression function to overcome discontinuous estimation results due to classification. Chen *et al*. proposed a cost estimation method [6] for full density disparity estimation, which uses the semi global matching for computing the similarity of image patches and takes advantages of slanted plane smoothing method [8] for smoothness. Another efficient research is from perception system [21], such as [9] which transforms a 3D lidar point cloud into a 2-D dense depth map.

On the basis of foregoing researches, multiple visual tasks are also possible to be accomplished in an integrated network. Song *et al*. proposed EdgeStereo [31] to support the joint learning of scene matching and edge detection. By adopting the context pyramid to handle ill-posed regions and the residual pyramid to replace cascaded refinement structure, EdgeStereo has achieved state-of-the-art performance on the KITTI dataset [27]. Xu *et al*. proposed a PAD-Net [36], which embeds feature of four tasks into a single network by a multi-modal distillation and has achieved state-of-the-art results on both the depth estimation and the scene parsing tasks. For stereo matching and scene parsing, Cheng *et al*. proposed the Segflow [10] to join two independent networks, *i.e*., the FlowNet and the FCN, to simultaneously estimate the optical flow and semantic segmentation in videos. The convolutional features of those two networks are combined together and the two task branches gradually deconvolve the mixed features to obtain the final semantic graph. In another work [18], Jin *et al*. transferred the features of a flow anticipating network into the scene parsing network by one residual block. Their networks can be utilized to predict both dense flow and semantic information for unobserved future video frames, which is demonstrated beneficial for robot systems like autonomous vehicles to make motion plans or decisions. Yang *et al*. proposed SegStereo [38] which conducts semantic feature embedding and regularizes semantic cues as the loss term to improve learning disparity. However, different from the above methods, we propose a lightweight structure in which two tasks share a same backbone weight, and we also design a new matching modules for extracting fuse features for scene parsing and disparity estimation.

III. Predicting Scene Semantic and Disparity

In this section, we first introduce the joint architecture of our DSNet. Then we reveal that the variation of semantic features between the left and right images can be captured by a series of matching operations. Thus the fuse of semantic features can be used for disparity estimation. Furthermore, we propose an effective training method to promote the performance of our proposed network, which is employing more effective optimization functions and warping operation for disparity and scene parsing.

A. An overview of DSNet

Inspired by the effectiveness of fully convolutional network(FCN) [24] in image segmentation and the deep structure in image classification [16] [32], we apply the semantic encoder as a dominating foundation for the overall network. From Fig. 1, we can see that the two branches, scene parsing and disparity estimation, share the same ResNet blocks. Given left-image X_L and right-image X_R as input, the output of each block is semantic features f_L and f_R. We design a matching module to capture the location features between f_L and f_R for generating disparity maps. The details of the matching module will be introduced in next section. As for the loss layer of scene parsing, we take the imbalance of pixel quantity between different classes into account. For example, the number of pixels belonging to the road class is far more than those belonging to the class of pedestrian or vehicle. Here we define a weight based on the ratio of the pixel number of one class $|Y_C|$ to the whole image area $|Y|$, *i.e.*, $\frac{|Y_C|}{|Y|}$. All the weights are sorted in a list of an ascending order. To deal with the data imbalance problem between different classes in the training procedure, we switch the weights for their corresponding classes by reversing the list, which means small classes are assigned with large weights while big classes are less weighted. So the new weight for class C is denoted as $\frac{|\bar{Y}_C|}{|Y|}$. A pixel-wise cross-entropy loss with the softmax function E is employed for the output from each deconvolution layer l of the semantic network. With that, the total loss function can be interpreted as:

$$L_{seg}(Y_{gt}) = -\sum_{l}^{5}\sum_{i,j} W_{ij} \log E(Y_{gt}^{l}(i,j); Y^{l}(i,j)) \quad (1)$$

subject to

$$W_{ij} = \frac{|\bar{Y}^{l}(i,j)|}{|Y_{gt}|}, \quad (2)$$

where $Y(i,j)$ indicates the label of pixel with image coordinates (i,j) and $|Y_{gt}|$ denotes the total pixel number of the given image.

In order to learn the features, five convolutional modules of ResNet, *i.e.*, conv1, res2, res3, res4 and res5, are connected to the horizontal correlation maps. To simplify the implementation, we maintain the original decoding way of DispNet [26] in our network. The loss function between the model output V^l and the ground truth V_{gt}^l is formulated as below:

$$L_{disp}^{l} = \begin{cases} \frac{1}{|V^l|}|V^l - V_{gt}^l| & \text{if } |V^l - V_{gt}^l| >= 1 \\ 0.5\frac{1}{|V^l|}(V^l - V_{gt}^l)^2 & \text{otherwise} \end{cases} \quad (3)$$

with the total disparity loss as

$$L_{disp} = \sum_{l}^{5} L_{disp}^{l}. \quad (4)$$

Eq. 3 is a common mixed loss function that is widely used for border detection of objects [14] [29]. When the error is large, it can be rapidly decreased. When the error is small, it decreases relatively slowly.

Due to the setup of the datasets, for each pair of images, the semantic information is usually provided for only one image (*i.e.*, the left one). To further promote the performance of the semantic network, we put forward an advanced version of training scheme by employing a warping operation. The core idea of warping is to horizontally translate the convolved features of the right image according to the disparity value and fuse them into the feature maps of the left image.

Here we define the warping machanism with $\mathbf{D}_{r\to l}$ as the estimated disparity map attained from the network N_{disp}. Then we use $\mathbf{D}_{r\to l}$ to project a location $\mathbf{p}$ in the right feature map $\mathbf{f}_r$ back to the location $\mathbf{p}+\Delta\mathbf{p}$ in the left feature map $\mathbf{f}_l$, where $\Delta\mathbf{p} = \mathbf{D}_{r\to l}(\mathbf{p})$. As the predicted disparity value $\Delta\mathbf{p}$ is basically a fractional number, the feature warping is conducted by bilinear interpolation, which can be formulated as

$$\mathbf{f}_l^c(\mathbf{p}) = \sum_{\mathbf{q}\in\mathcal{N}} T(\mathbf{q}, \mathbf{p}+\Delta\mathbf{p})\mathbf{f}_r^c(\mathbf{q}), \quad (5)$$

where c is the channel index of the feature map $\mathbf{f}$, term $\mathbf{q}$ denotes location of a pixel in the corresponding neighborhood $\mathcal{N}$ upon the right feature map and T is the bilinear interpolation kernel. Through the above warping operation, the right semantic feature map can be indirectly summed up with the left one. To balance the influence of left and right semantic features, we empirically set the weight for the original left feature map and the warped one as 0.7 and 0.3, respectively.

The training process can be divided into three steps: (1) Firstly, we train the semantic network, which extracts auxiliary features for subsequent disparity estimation. (2) Then we fix the weights of the semantic network and train the disparity decoder independently. (3) Finally, we conduct joint learning of two tasks by employing a match module and warping operation. As demonstrated by the experimental results, such a training method achieves the best performance.

B. Matching module

Concatenation is a common way to fuse information in the feature maps and we also take the more complex operation, namely correlation [12], into consideration. In this process, given the left and right feature maps, the size of the output of correlation layer is the same as that of input feature maps. Here we use a rectangular patch K_{x1} centered at pixel $x1$ in the left map as the convolution kernel in stead of a filter of adaptive weights to convolve the right map but only horizontally. Consequently, the convolution between a single patch K_{x1} and the right map result in a vector of size w, where w is the width of the right map. The size of patches is set to a reasonable range since the scope of disparity is finite. The matching operation which adopts correlation1D from the DispNetC [26] employs a convolution-like fashion to conduct comparison between f_L and f_R patch-wisely. Note that although the mathematical interpretation of correlation resembles a typical convolution layer, here it convolves a feature map with another one rather than a filter of adaptive weights. We adopt the operation of correlation1D from the

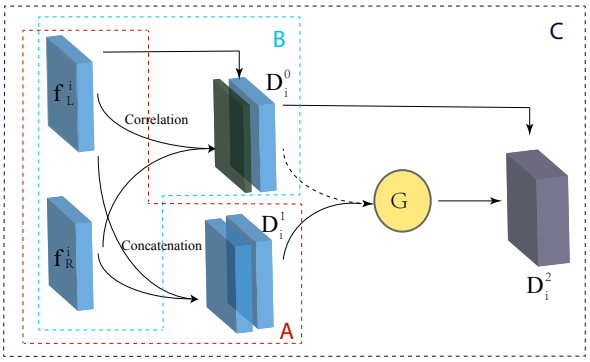

Fig. 2: Module A, B, C are three distinct matching modules. Module A uses concatenation to fuse the information from left and right feature maps. Module B uses correlation to achieve it, and module C combines module A and B with the help of attention mechanism.

DispNetC [26] for disparity estimation, which only convolves in a horizontal direction to capture differences.

More specifically, we construct three modules applying these two ways of fusing information. As shown in Fig. 2, Module A simply fuses information by concatenating feature maps f_L^i and f_R^i and extracts features continuously from the concatenated one D_i^0. Note that i means the i-th matching module. Module B combines f_L^i and the correlation between f_L^i and f_R^i to obtain a fused feature, which is denoted by

$$D_i^1 \leftarrow f_L^i + \left(f_L^i \otimes f_R^i\right), \tag{6}$$

where $\otimes$ refers to correlation operation. Moreover, in module C, we combine module A and B by introducing attention mechanism [36] which has been proven to be effective. Although both module A and module B aim to fuse information, the result feature of these two modules is expected to be as useful as possible. Therefore, based on attention mechanism, we produce an attention map G_i to make a selection:

$$G_i \leftarrow \sigma\left(W_i \otimes D_i^1\right), \tag{7}$$

where W_i refers to a convolution kernel and σ refers to a sigmoid function. Thus, module C can be denoted as following:

$$D_i^2 \leftarrow D_i^1 + G_i \odot D_i^0, \tag{8}$$

where $\odot$ refers to element-wise multiplication. D_i^0 and D_i^1 are the results of module A and module B respectively. In addition to patch-level matching achieved by module A, the utility of module C is inspired by [6] also takes advantages of pixel-level matching.

IV. Experiments

Our experiments are evaluated on two datasets: the Cityscapes [11] for scene parsing and the KITTI 2015 [27] for disparity estimation. We test with the validation set of Cityscapes [11] for our scene parsing. For our disparity estimation on the KITTI 2015 [27], we mainly pretrain the model on CityScapes dataset and then fine tune on KITTI Stereo dataset, using 75% of the train set of KITTI as our train set while testing on the remaining samples. Compared with the baseline models of scene parsing and disparity estimation, the experimental results demonstrate the effectiveness of our simple integrated dual-task network.

A. Dataset and Evaluation Metrics

The dataset of Cityscapes consists of images collected from 50 different cities in Europe, with fine annotations for 5000 images and 20000 additional images, which are only with coarse annotations. Images are captured with urban street scenes and their pixels are categorized into 19 testing classes. In comparison, the KITTI 2015 dataset contains 389 image pairs for real-world driving scenes. The constitution of KITTI 2015 is identical to Cityscapes, which also covers 19 object classes. Both datasets provide separate image groups for training and test purpose. To evaluate the segmentation performance of our proposed architecture, we resort to the standard Jaccard Index, known as the intersection-over-union (IOU) metric. To evaluate the disparity estimation in Cityscapes, we take the average endpoint error (EPE) introduced in the work [40] as a metric. On KITTI 2015, we use the D1-all error (*i.e.*, the percentage of disparity outliers in the first frame) instead of EPE as the metric, because only the D1-all error measurement is reported by the KITTI evaluation server.

B. Network Implementation and Training

To train the proposed network, we employ a data augmentation approach to expand the training dataset. In order to generate more images, we apply a scaling operation on the original dataset with a factor from the range of $[0.5, 2]$. Additionally, we conduct image rotation with a small angle interval of $[-10°, 10°]$. Furthermore, we adjust the color, brightness and exposure of images with a random factor in a small range of $[0.8, 1.2]$.

Pre-training for segmentation task For optimizing the network, we use the SGD optimizer with a batch size of 4. During training, the learning rate starts from 1e-3 and follows a polynomial decay, where the current learning rate is equaled to the original number multiplied by $(1 - \frac{iter}{maxiter})^{power}$. We set $power$ to 0.9. Total epoch on Cityscapes is set to 40.

Fine-tuning for disparity task To train the disparity branch, we freeze the encoding part of segmentation and only update the weights of its decoding part. We follow the same training hyperparameters as DispNetC [26]. To learn the features, five convolutional modules of ResNet, *i.e.*, conv1, res2, res3, res4 ,and res5, are connected to the horizontal correlation maps, which are obtained from patches with a size of 40, 40, 20, 20 and 10 pixels, respectively .

Fine-tuning for all tasks In the joint learning process, we fine-tune the two tasks and train all the network parameters, which allows the mutual enhancement of both branches. The

TABLE I: The baseline of scene parsing in the val set of Cityscapes.

Model	mIOU (%)
FCN + ResNet50	62.2
FCN + ResNet101	70.7
FCN + ResNet50 + module C + Warp	66.8
FCN + ResNet101 + module C + Warp	75.1

TABLE II: The baseline of disprity estimation in the val set of Cityscapes.

Model	EPE
FCN + ResNet50 + module A	3.05
FCN + ResNet50 + module B	2.91
FCN + ResNet50 + module C	2.80
FCN + ResNet101 + module A	2.88
FCN + ResNet101 + module B	2.83
FCN + ResNet101 + module C	2.76
FCN + ResNet50 + model C + Warp	2.05
FCN + ResNet101 + model C + Warp	1.83

learning rate starts at 1e-5 and follows a polynomial decay. At the same time, the warping operation also enables training semantics for the right images. we empirically set the weight for the original left feature map and the warped one as 0.7 and 0.3, respectively.

C. Ablation learning

To conduct the diagnostic experiments, we compare the results of training with different matching modules or init weights from different scene parsing model. The results of the origin backbone training with the scene parsing is shown in Table I. We tested those methods on the val set. The evaluation results are reported in Table II under the EPE indicator, we can see that matching module B is better than module A in all ResNet50/101 backbones. This is mainly because as the number of layers of the convolutional network increases, the distinction between the semantic features of the left and right graphs begins to decrease, and the distinction between the high-level semantic features can be improved by correlation operations, which is used for improving the efficiency of disparity estimation. However, module A is more focused on long-distance information because the disparity value is smaller and the depth is farther. Therefore, by combining the module A and module B into module C to increase the location diversity of features, our method has achieved the best results. By adding the wrapping operation and matching module, we get our final model. From Table I, we can see that DSNet has improved by at least 4% of mIOU in each backbone compared to training the scene parsing network alone. DSNet has achieved the best results on ResNet-101. For jointing learning the two tasks in all networks, the disparity estimation is up to 2.05/1.83 when using the ResNet-50 and ResNet-101 respectively.

TABLE III: Comparison with multi-task models on scene parsing.

Model	mIOU (%)
competition model without optical flow [18]	62.6
competition model with optical flow [18]	66.1
our DSNet without disparity	62.2
our DSNet with disparity	66.8

D. Comparison with other models

Scene parsing To fully demonstrate the advantages of our model on the estimation of semantic maps, we compare our model with other methods on Cityscapes. To the best of our knowledge, [18] is only one jointing learning stereo matching and scene parsing which evaluate results on Cityscapes, which transfers features of a flow anticipating network into the scene parsing network by using one residual block, the backbone of [18] is ResNet-50. Evaluation results are reported in Table III. We can see that in terms of the measurement of mIOU, our DSNet combined with baseline with an improvement of up to 4.6%. In terms of multi-tasking approaches, our single-network DSNet also achieves a considerable improvement over other dual-network architecture [18] by using a matching module C and warping method. In competition, the method achieves an mIOU value of 66.1% on Cityscapes with a gap of up to 3.5% lower than ours. In order to compare with more methods of scene parsing, we use ResNet-101 as the basis of our model, and the results are shown in the Table IV. It can be seen that our method is sufficiently competitive compared to other methods. As the comparing method only provides the scene parsing on Cityscapes, we do not quantitatively evaluate the scene parsing performance on KITTI. As for qualitative results, we provide examples from the Cityscapes, which are shown in Fig. 3.

Disparity estimation From the Table V, our model also performs better than our baseline DispNetS on the KITTI 2015 dataset by a gain of up to 1.67%. The reason is that ground truth values for disparity maps in the KITTI 2015 are very sparse, which can be seen in Fig. 4 that our model can still learn better disparity maps with the help of dense semantic features. By leveraging those methods, in comparison with other methods, EdgeStereo [31] use a context information into disparity branch and followed by a compact residual pyramid for cascaded refinement, make an effective way. Due to the limitations of the graphics card memory, we can't add more adjustments, there are still more possibilities for our approach, such as extending a small network to allow the network to output higher resolutions to enhance the semantic segmentation effect while this small network can also fit the smaller displacement in a cascaded residual refinement or add a shallow network to extract more

TABLE IV: Comparison with other models on Cityscapes, those method train with fine annotations.

Model	mIOU(%)
CRF-RNN [44]	62.5
FCN [24]	65.3
DPN [23]	66.8
LRR [13]	69.7
DeepLabv2-CRF [22]	70.4
Piecewise [1]	71.6
Our DSNet	75.1
RefineNet [41]	77.3
TuSimple [34]	77.6
SAC-multiple [42]	78.1
PSPNet [43]	78.4
DeepLabv3+ [5]	79.6

TABLE V: Comparison with models on disprity estimation on KITTI 2015.

Model	D1-all(%)
EdgeStereo [31]	2.16
SegStereo [38]	2.25
Our DSNet	2.67
Displets v2 [15]	3.43
PBCP [30]	3.61
MC-CNN-acrt [39]	3.89
chen17 [7]	4.07
PRSM [33]	4.27
DispNetC [26]	4.34
Content-CNN [25]	4.54
SPS-St [37]	5.31
OSF [27]	5.79
MDP [20]	6.74

robust features, which widely used in [28] [31] [38]. As for qualitative results, we provide examples from the KITTI, which are shown in Fig. 4.

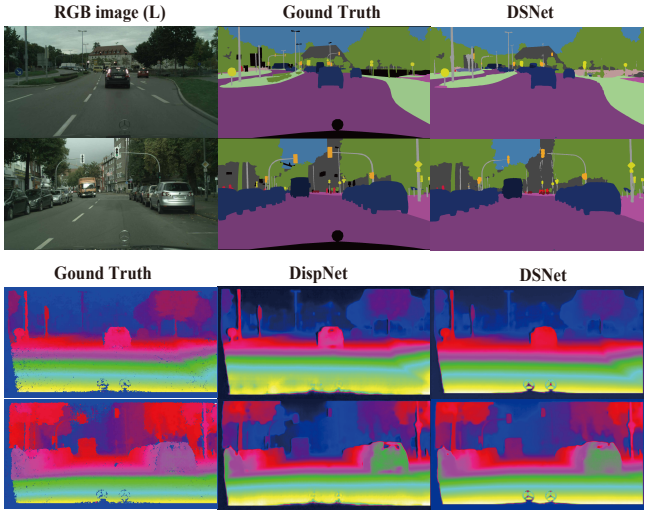

Fig. 3: Qualitative test results on the Cityscapes. On the top left we display two samples which are the left images in their corresponding image pairs. In the following of first row, we respectively show the ground truth and predicted semantic segmentation by our DSNet. In the second row, we show the ground truth for disparity estimation and the estimation results by DispNetC [26] and DSNet.

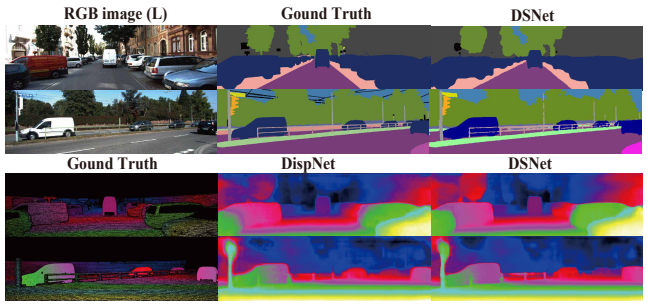

Fig. 4: Qualitative test results on KITTI 2015 benchmark. On the top left we display two samples which are the left images in their corresponding image pairs. In the following of first row, we respectively show the ground truth and semantic segmentation by our DSNet. In the second row, we show the (sparse) ground truth for disparity estimation and the estimation results by DispNetS [26] and DSNet.

V. Conclusions

In this paper, we proposed an integrated network named DSNet for simultaneously dense semantic segmentation and disparity estimation. The network uses a series of matching modules to extract features of deep semantics and deep disparities, greatly improving the computational efficiency as well as reducing the memory consumption. We also propose a training approach for the multi-task architecture, by applying an alternative training fashion and combining the warp operation and learning of the two sub-tasks are mutually beneficial. Furthermore, a pixel-level cost weighting function is introduced for training semantic segmentation. In order to obtain finer disparities, a method with a smoothed mixture of L1- and L2-loss is introduced. Extensive experimental results prove that semantic features can be used for disparity estimation. At the same time, our integrated network DSNet provides a performance in consistency with other dual-network architectures and improves the effectiveness for each subtask.

Acknowledgement

This work was supported in part by the National Key R&D Program of China under Grant 2018YFB1305002 and the National Natural Science Foundation of China under Grant 61773414.

REFERENCES

[1] A. Arnab and P. H. S. Torr, "Pixelwise instance segmentation with a dynamically instantiated network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 879–888.

[2] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5410–5418.

[3] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[4] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *international conference on learning representations*, 2015.

[5] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation." *arXiv: Computer Vision and Pattern Recognition*, 2018.

[6] L. Chen, L. Fan, J. Chen, D. Cao, and F. Wang, "A full density stereo matching system based on the combination of cnns and slanted-planes," *IEEE Transactions on Systems, Man, and Cybernetics*, pp. 1–12, 2017.

[7] ——, "A full density stereo matching system based on the combination of cnns and slanted-planes," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2017.

[8] L. Chen, L. Fan, G. Xie, K. Huang, and A. Nuchter, "Moving-object detection from consecutive stereo pairs using slanted plane smoothing," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 11, pp. 3093–3102, 2017.

[9] L. Chen, Y. He, J. Chen, Q. Li, and Q. Zou, "Transforming a 3-d lidar point cloud into a 2-d dense depth map through a parameter self-adaptive framework," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 1, pp. 165–176, 2017.

[10] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang, "Segflow: Joint learning for video object segmentation and optical flow," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 686–695.

[11] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] A. Dosovitskiy, P. Fischery, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. V. Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," *international conference on computer vision*, pp. 2758–2766, 2015.

[13] G. Ghiasi and C. C. Fowlkes, "Laplacian pyramid reconstruction and refinement for semantic segmentation," 2016.

[14] R. B. Girshick, "Fast r-cnn," *international conference on computer vision*, pp. 1440–1448, 2015.

[15] F. Guney and A. Geiger, "Displets: Resolving stereo ambiguities using object knowledge," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4165–4175.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] O. H. Jafari, O. Groth, A. Kirillov, M. Y. Yang, and C. Rother, "Analyzing modular cnn architectures for joint depth prediction and semantic segmentation," *international conference on robotics and automation*, pp. 4620–4627, 2017.

[18] X. Jin, H. Xiao, X. Shen, J. Yang, Z. Lin, Y. Chen, Z. Jie, J. Feng, and S. Yan, "Predicting scene parsing and motion dynamics in the future," in *Advances in Neural Information Processing Systems*, 2017, pp. 6918–6927.

[19] A. Kendall, H. Martirosyan, S. Dasgupta, and P. Henry, "End-to-end learning of geometry and context for deep stereo regression," *international conference on computer vision*, pp. 66–75, 2017.

[20] A. Li, D. Chen, Y. Liu, and Z. Yuan, "Coordinating multiple disparity proposals for stereo computation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4022–4030.

[21] Q. Li, L. Chen, M. Li, S. Shaw, and A. Nuchter, "A sensor-fusion drivable-region and lane-detection system for autonomous vehicle navigation in challenging road scenarios," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 2, pp. 540–555, 2014.

[22] I. K. K. M. A. L. Y. Liang-Chieh Chen, George Papandreou, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[23] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," pp. 1377–1385, 2015.

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *computer vision and pattern recognition*, pp. 3431–3440, 2015.

[25] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5695–5703.

[26] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "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*, 2016, pp. 4040–4048.

[27] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[28] J. Pang, W. Sun, J. Ren, C. Yang, and Q. Yan, "Cascade residual learning: A two-stage convolutional neural network for stereo matching," in *International Conf. on Computer Vision-Workshop on Geometry Meets Deep Learning (ICCVW 2017)*, vol. 3, no. 9, 2017.

[29] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[30] A. Seki and M. Pollefeys, "Patch based confidence prediction for dense disparity map." in *BMVC*, vol. 2, no. 3, 2016, p. 4.

[31] X. Song, X. Zhao, H. Hu, and L. Fang, "Edgestereo: A context integrated residual pyramid network for stereo matching," *arXiv preprint arXiv:1803.05196*, 2018.

[32] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[33] C. Vogel, K. Schindler, and S. Roth, "3d scene flow estimation with a piecewise rigid scene model," *International Journal of Computer Vision*, vol. 115, no. 1, pp. 1–28, 2015.

[34] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *IEEE Winter Conference on Applications of Computer Vision*, 2018, pp. 1451–1460.

[35] P. Wang, X. Shen, Z. Lin, S. D. Cohen, B. L. Price, and A. L. Yuille, "Towards unified depth and semantic prediction from a single image," pp. 2800–2809, 2015.

[36] D. Xu, W. Ouyang, X. Wang, and N. Sebe, "Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing," *computer vision and pattern recognition*, pp. 675–684, 2018.

[37] K. Yamaguchi, D. McAllester, and R. Urtasun, "Efficient joint segmentation, occlusion labeling, stereo and flow estimation," in *European Conference on Computer Vision*. Springer, 2014, pp. 756–771.

[38] G. Yang, H. Zhao, J. Shi, Z. Deng, and J. Jia, "Segstereo: Exploiting semantic information for disparity estimation," *arXiv preprint arXiv:1807.11699*, 2018.

[39] J. Zbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1592–1599.

[40] ——, "Stereo matching by training a convolutional neural network to compare image patches," *Journal of Machine Learning Research*, vol. 17, no. 1-32, p. 2, 2016.

[41] R. Zhang, S. Tang, M. Lin, J. Li, and S. Yan, "Global-residual and local-boundary refinement networks for rectifying scene parsing predictions," in *Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017, pp. 3427–3433.

[42] R. Zhang, S. Tang, Y. Zhang, J. Li, and S. Yan, "Scale-adaptive convolutions for scene parsing," in *IEEE International Conference on Computer Vision*, 2017, pp. 2050–2058.

[43] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.

[44] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," pp. 1529–1537, 2015.