Improved-PSMNet for Deep Stereo Disparity Estimation

Ganlin Zhang, Xinyu Shen, Yunying Zhu, Haokai Pang
ETH Zurich
{zhangga, xishen, yunzhu, hapang}@*student.ethz.ch*

***Abstract*—Depth estimation problem for stereo images is explored by plenty of state-of-the-art deep learning methods in recent years. One of the most typical end-to-end models is called Pyramid Stereo Matching Network (PSMNet). To further integrate semantic cues in the model and enhance feature representation, we propose an Improved-PSMNet by modifying PSMNet in three aspects, which can increase the accuracy and promote efficiency of the original model. Firstly, a group-wise correlation measuring similarities of the extracted features is complementary to the concatenation of features in PSMNet. Then semantic segmentation information predicted by an advanced segmentation model is embedded into cost volume as a semantic part. Finally, we construct a dilated residual cost filtering to find the correct minimum solution of cost volume. The experiments are conducted on SceneFlow and KITTI2015 datasets and the performance of the PSMNet model is used as the benchmark. According to the results of our experiments, our Improved-PSMNet outperforms the baseline model on both datasets.**

Source code can be found here: Github: Improved-PSMNet-for-Deep-Stereo-Disparity-Estimation

I. Introduction

Depth estimation for stereo vision is crucial in many computer vision applications, such as autonomous driving, object detection, augmented realities [1]. In the rectified stereo setting, depth estimation can be treated as finding the disparity (the horizontal displacement between the left and right images) for each pixel in the reference(left) image.

The traditional pipeline for stereo matching involves the finding of corresponding points based on matching cost, cost aggregation, disparity optimization and post-processing [2]. However, finding stereo disparity remains a challenging problem due to the complex environment of the real world. It is difficult to find accurate corresponding points in ill-posed regions such as highly reflective surfaces, repetitive textures, and occlusion areas [3].

Recently, end-to-end deep learning models have been applied to disparity estimation with significant success [4][5][6][7][8]. PSMNet [9] constructs the milestone of stereo based on spatial pyramid pooling (SPP) and 3D encoder-decoder modules: the SPP module forms concatenation-based feature volume by aggregating context in different scales and locations, the 3D encoder-decoder module further regularizes the cost volume through repeated top-down/bottom-up processes to improve the utilization of global context information. Although it achieves higher accuracy and efficiency compared to the previous GC-Net [10], PSMNet still has several drawbacks. PSMNet directly concatenate the left and the right image features, which leads to the lack of similarity information in the cost volume. The full correlation constructs a powerful way of incorporating feature similarities, but it loses much information since it produces only a single-channel correlation map for each disparity level [11]. We propose the combination of concatenation and group-wise correlation to solve the above problems. The extracted features are split into groups along the channel dimension, and the correlation maps are calculated among every group over all disparity levels to obtain multi-matching cost proposals. Group-wise correlation represents feature similarities efficiently without the loss of information, and direct concatenation preserves the original extracted features to guide the subsequent process [11].

In this work, we also integrate semantic segmentation information to stereo matching partially inspired by the work of SSPVC-Net [5] and SegStereo [12]. Semantic segmentation captures different objects and their boundaries in images and relates with disparity maps in spatial and intensity information. Especially, an accurate semantic segmentation can rectify the disparity values along the object boundaries, which are usually more prone to accumulate error in stereo matching [13]. Thus, semantic information predicted by the state-of-the-art segmentation network can be embedded into the cost volume to improve the performance efficiently without complicated joint training and loss function in our work.

PSMNet usually finds the solution with several local minima when argmin on the cost volume, especially for the homogeneous or repeating texture [14]. By transforming the cost filtering as a deep learning process with a multi-dilation unit and residual connection, we attempt to find the correct local minima. Besides, our dilated residual cost filtering tackles the excessive parameters problem which PSMNet suffers from.

Our main contributions to Improved-PSMNet are listed below:

- We combine the group-wise correlation and concatenation to construct cost volumes to obtain a powerful feature representation.
- We integrate the semantic information to the cost volume to improve performance on the boundary and discontinuous part.
- We propose a 3D multi-dilated residual module in cost filtering instead of the 3D encoder-decoder module (3D stacked hourglass CNN) to give better results.
- We achieve better performance and use fewer parameters compared to PSMNet in Scene Flow and KITTI 2015 dataset.

II. Models and Methods

The architecture of the proposed Improved-PSMNet is shown in Figure 1. We can see that new cost volumes are built to incorporate semantic information and group-wise

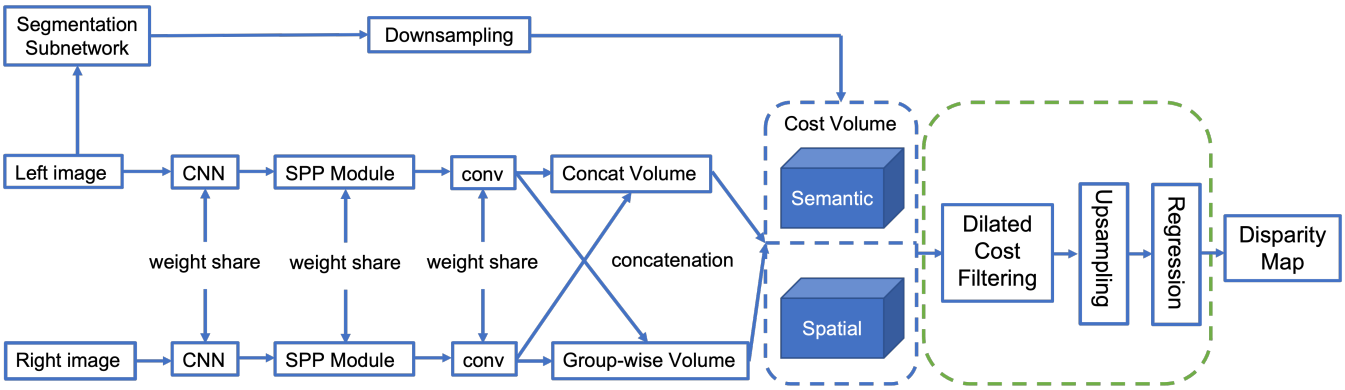

Fig. 1. Architecture overview of Improved-PSMNet. The left and right image are fed into the PSMNet feature extraction backbone, and the spatial cost volume is formed by concatenation and group-wise correlation of features. The left image is also fed into segmentation subnetwork to achieve semantic information to embed the cost volume, which is the input to 3D dilated residual cost filtering and disparity regression.

correlation. In addition, dilated residual cost filtering module is added for cost-volume aggregation and regularization.

A. Network Architecture

For feature extraction, we adopt the ResNet-like [15] network used in PSMNet with the half dilation settings and spatial pyramid pooling module. The cost volume consists of a semantic and a spatial part. The spatial part combines a concatenation volume and a group-wise correlation volume. The concatenation volume is the same as PSMNet and details of the proposed group-wise correlation volume will be described in Section B. The semantic volume is the downsampled output of the semantic segmentation subnetwork. All these cost volumes are then fed into a 3D dilated residual cost filtering module for aggregation and regularization. Finally, a disparity regression is applied to estimate the continuous disparity map. The group-wise correlation, semantic segmentation, and dilated residual module are elaborated in the following sections.

B. Group-wise Correlation

The left and right features are correlated or concatenated to form the cost volume in previous works [10] [16]. However, both correlation volume and concatenation volume have disadvantages.Although the full correlation measures feature similarities efficiently, it leads to loss of much information because it produces only a single-channel correlation map for each disparity level. The concatenation volume cannot represent feature similarities, therefore, much more parameters are required to extract similarity information in the 3D aggregation process. We combine the group-wise correlation and concatenation as spatial cost volume to tackle the above drawbacks.

The milestone of the group-wise correlation is computing the correlation of every split feature group. N_c channels of features can be divided into N_g groups evenly and each group will contain $\frac{N_c}{N_g}$ features. For example, the g^{th} group f_l^g, f_r^gcontains $(g.\frac{N_c}{N_g}, g.\frac{N_c}{N_g}+1, ..., \frac{N_c}{N_g}(g+1)-1)$th channels of the original feature f_l, f_r. The group-wise correlation is calculated as below:

$$C_{gwc}(d,x,y,g) = \frac{1}{N_c/N_g}\langle f_l^g(x,y), f_r^g(x-d,y)\rangle, \quad (1)$$

where $\langle\cdot,\cdot\rangle$ is the inner product [11].

The group-wise correlation successfully utilizes the ability of traditional correlation matching costs and achieves powerful similarity-measure features for the 3D aggregation network, which alleviates the parameter demand. Results of GwcNet [11] justify the fact that group-wise correlation requires fewer parameters and the group-wise correlation volume and the concatenation volume are complementary to each other.

C. Semantic Segmentation

Compared to disparity estimation, semantic segmentation is a high-level classification task where each pixel in the image is assigned to a class. The discontinuous boundary information of semantic segmentation can be leveraged to assist the stereo matching. We use semantic cues to help predict and rectify the final disparity map. As a result, we first incorporate the cues by embedding semantic features.

For the semantic branch, the semantic segmentation subnetwork follows the encoder-decoder structure, where encoders are usually modified directly from classification networks, and decoders consist of final convolutions and upsampling. The pre-trained ResNet50-Dilated and Pyramid Pooling Module (PPM) [17][18] are implemented as encoder-decoder [19] part of the segmentation subnetwork of our Improved-PSMNet.

To form the single semantic cost volume, we use the left image as input of the segmentation subnetwork. The use of semantic cost volume aims to capture context cues in a simple manner and learn the semantic segmentation features. By concatenating each semantic feature across each disparity level, and packing them into a 4D volume, we obtain a semantic cost volume with the size of $C \times \frac{1}{4}H \times \frac{1}{4}W \times \frac{1}{4}D$, where C is the number of channels, W and H are the width and height of original images respectively, and D is the maximum disparity. The semantic cost volume occupies the same size as spatial cost volume.

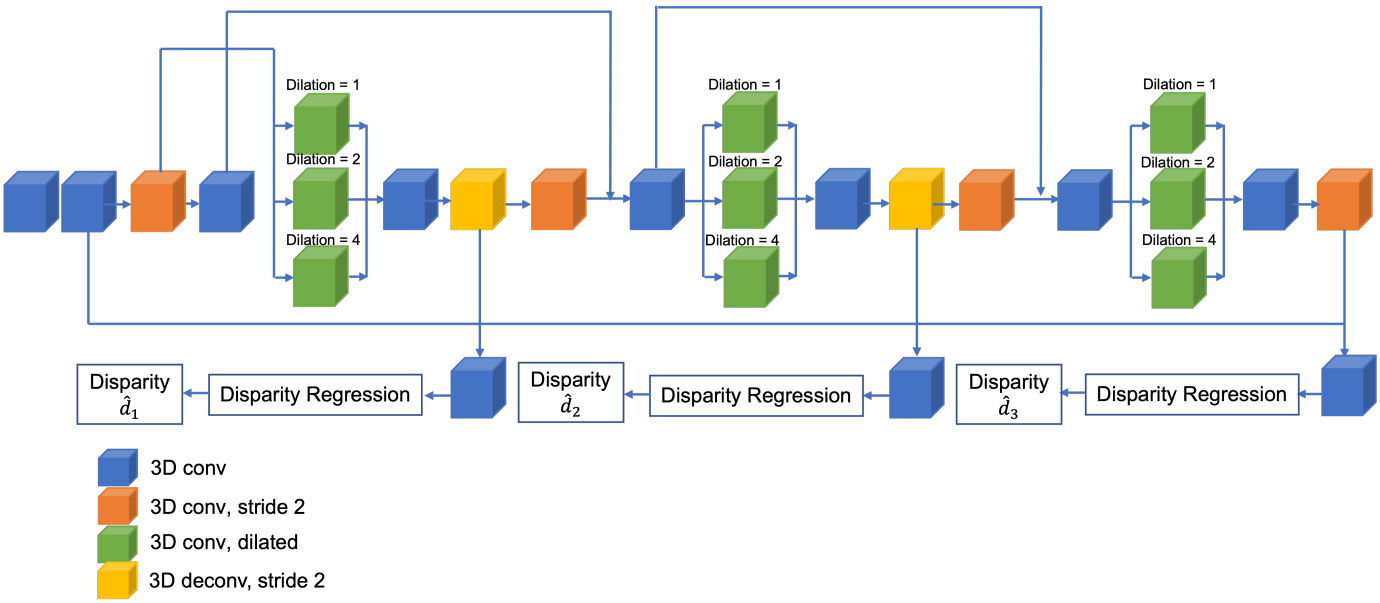

Fig. 2. Architecture overview of Dilated Residual Cost Filtering in 3D CNN which is the green dash line in Figure 1

D. Dilated Residual Cost Filtering

Although the direct argmin on the cost volume should theoretically lead to the correct local minima, it has been proved that it is common for the solution to have several local minima [20] [21]. Regions with homogeneous or repeating textures in the image are prone to stick in this situation.

We attempt not only to solve the above problem by applying 3D convolution into the cost filtering process but also to reduce the resolution, which can reduce the number of parameters needed without any loss in our model [22] [23]. After processing the cost volume with a single 3D convolution ($3 \times 3 \times 3$) along the width, height, and depth dimension as the same in PSMNet, we apply a convolution with stride 2 to reduce the resolution, then three convolutions with dilated 1, 2 and 4 in parallel to collect the contextual information in different receptive fields [14]. A convolution on the concatenation of multi-dilated convolution filters aims to combine the information fetched from various receptive fields. Similar to PSMNet, we use a residual connection to improve the quality of disparity prediction. The architecture for the whole procedure is shown in Figure 2, and the disparity prediction and loss function will be described in section E.

E. Disparity Regression and Loss Function

We adopt disparity regression [24] to estimate continuous disparity maps. The predicted disparity $\hat{d}$ is calculated as the sum of each disparity d weighted by its probability, as

$$\hat{d} = \sum_{d=0}^{D_{max}} d \times softmax(-c_d) \tag{2}$$

As shown in Figure 2, there would be three disparity outputs from our Improved-PSMNet. We use the same loss function as PSMNet:

$$\begin{aligned} &Loss(d, \hat{d}_1, \hat{d}_2, \hat{d}_3) \\ &\quad = 0.5L(d, \hat{d}_1) + 0.7L(d, \hat{d}_2) + L(d, \hat{d}_3) \end{aligned} \tag{3}$$

in which $L(\cdot, \cdot)$ is the loss function defined as

$$L(d, \hat{d}) = \frac{1}{N} \sum_{i=1}^{N} smooth_{L_1}(d_i - \hat{d}_i), \tag{4}$$

in which

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2, & \text{if } \|x\|<1 \\ \|x\| - 0.5, & \text{otherwise} \end{cases} \tag{5}$$

III. Experiments and Results

A. Datasets

We used two different dataset to train and test our model.

SceneFlow [25]:

This is a synthetic dataset consisting of Flyingthings3D, Driving, and Monkaa, which provide 35,454 training and 4,370 testing images of size 960×540 with accurate ground-truth disparity maps. This dataset is challenging due to the presence of occlusions, thin structures, and large disparities.

KITTI2015 [26]:

This is a real-world dataset that contains 200 training images with sparse ground-truth disparities acquired using a LiDAR sensor from both static and dynamic outdoor scenes, and 200 test images without ground-truth disparities.

Due to the limitation of time and computation resources, we trained both the baseline model PSMNet and our model Improved-PSMNet on only one-third of the original SceneFlow dataset (driving sub-dataset, which is similar to KITTI's environment) and tested on the KITTI2015 dataset.

TABLE I
ABLATION STUDY ON SCENEFLOW AND KITTI 2015 DATASETS

Network setting			SceneFlow (driving)	KITTI2015		# Parameters
Semantic Segmentation	Group-wise Corr	Dilated ResNet	End Point Error	3-Pixel Error Finetuned(%)	3-Pixel Error	
			3.917	1.971	0.341	5.226M
✓			3.614	1.841	0.235	5.227M
	✓		3.422	1.793	0.224	5.223M
		✓	3.472	1.839	0.227	4.339M
✓		✓	3.469	1.923	0.206	4.340M
	✓	✓	3.216	**1.754**	0.200	**4.336M**
✓	✓		3.217	1.887	0.224	5.224M
✓	✓	✓	**3.167**	1.762	**0.155**	4.337M

B. Metrics

We use two types of error function to measure the quality of disparity estimation.

End Point Error:

This measures the average distance between ground-truth disparity and the output of the network. Math representation is same as equation (4).

3-Pixel Error:

This metrics gives the proportion of pixels whose difference between estimated disparity and ground-truth is larger than 3 and larger than 5% of ground-truth.

$$\text{3-Pixel Error} = \frac{\|B\|}{\|A\|} \tag{6}$$

in which

$A = \{\text{pixel} \mid \text{pixel with valid ground-truth disparity}\}$,

$B = \{x | x \in A,\ gt(x) - estimate(x) > \max\{3, 0.05gt(x)\}\}$

C. Implementation

The Improved-PSMNet architecture was implemented using PyTorch. All models were end-to-end trained with Adam (β_1 = 0.9, β_2 = 0.999) optimizer. We performed color normalization on the entire dataset for data preprocessing. During training, images were randomly cropped to size H = 256 and W = 512. The maximum disparity (D) was set to 192.

We trained our models using the SceneFlow dataset (especially the driving dataset) with a constant learning rate of 0.001 for 80 epochs. For SceneFlow, the trained model was directly used for testing.

For KITTI2015, we not only directly tested the model trained with SceneFlow dataset on KITTI2015 but also fine-tuned on the KITTI2015 training set for 300 epochs to obtain another test result. Since KITTI2015 is a real-world dataset, there are some pixels without disparity due to the physical property of LiDAR, in the fine-tuning stage, we just filter out these pixels and focus on those with disparity. The learning rate of this fine-tuning began at 0.001 for the first 200 epochs and 0.0001 for the remaining 100 epochs.

The batch size was set to 8 on four NVIDIA GeForce GTX 1080 (each of 2). We used ETHZ's Euler Cluster to train our model. Running parameters on Euler: *-n 4 -R "rusage[mem=7000,ngpus_excl_p=4]"*. The training process took about 3 hours for the SceneFlow dataset and 1.5 hours for the KITTI dataset.

D. Ablation Studies

We conducted ablation studies by breaking down different parts of our Improved-PSMNet on the SceneFlow and the KITTI2015 dataset. The importance of three key ideas in our model was evaluated: 1) group-wise correlation in cost volumes, 2) dilated residual cost filtering, and 3) semantic segmentation. The results shown in Table I justify our design choices for our model: all three parts improve the accuracy of disparity estimation when used both alone and in combination compared to the baseline PSMNet. The result of our proposed model and baseline PSMNet model is shown in Figure 3. We can see that the boundary of vehicles is more clear and smooth in our method. Also, the disparity from our method is more continuous for pixels belongs to the same plane. A detailed analysis is conducted in the next section. The result and training loss of other combination of the proposed three features can also be found in Appendix A and Appendix B.

E. Comparison with baseline PSMNet

We compare the performance of our Improved-PSMNet with PSMNet:

On SceneFlow:

According to Table I, all three methods reduce the end point error with group-wise correlation gaining the most progress. The various combination of three methods all improve the performance. The best performance is achieved to 3.167 end point error with all of our three methods, which reduces the error by 19.1% compared to the baseline model PSMNet.

On KITTI2015:

In the test on KITTI2015 dataset, we can see that if we apply the model trained only on Scene Flow directly, the model with our proposed three additional features performs best with a 0.155 3-pixel error. The relationship of results of

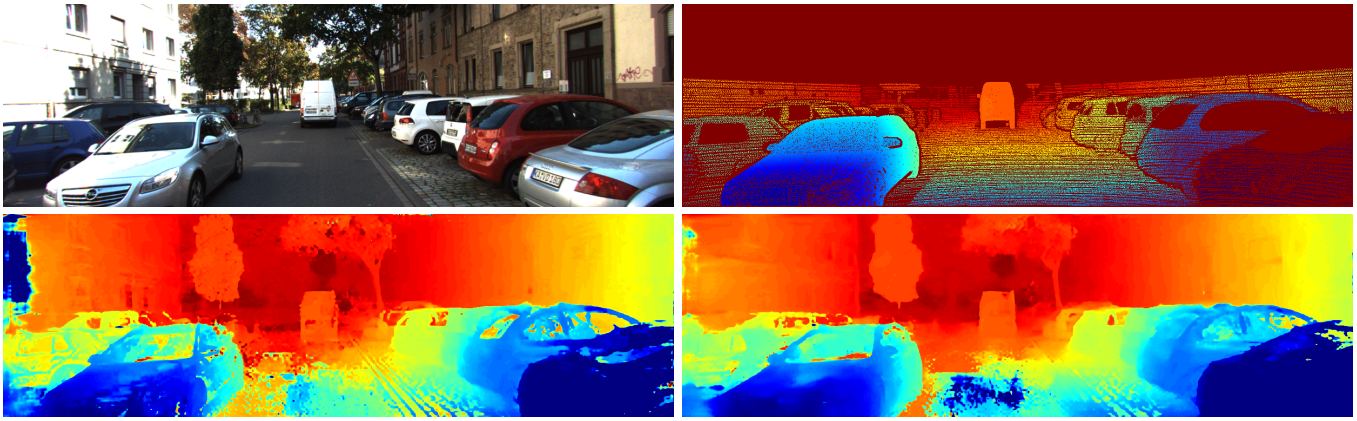

Fig. 3. Upper left: input left image from KITTI2015. Upper right: groundtruth disparity. Bottom left: output of PSMNet. Bottom right: output of our Improved-PSMNet.

every model is consistent with Scene Flow. This justifies our model is more robust to environmental and scenario change, possessing the ability to fit the general situation efficiently instead of a specific one.

After 300 epochs of fine-tuning on KITTI2015, the combination of semantic segmentation and dilated residual cost filtering makes the smallest progress. This is probably because, in KITTI2015, accurate segmentation is harder since there are more objects with unclear contours, such as trees and utility poles. Without the similarities representation by group-wise correlation, 3D aggregation blurs the boundary to some extent. However, all combined and separate methods still improve accuracy and all three methods achieve excellent performance together compared to PSMNet.

For every model implemented in ablation studies, fine-tuning reduce the 3-pixel error by 88% compared to the direct test, which verifies the generalization ability of our model and reasonability of fine-tuning. This also implies that our Improved-PSMNet can be used to estimate disparity on any stereo dataset by fine-tuning it.

Analysis of Parameters:

We also measure the number of parameters of the network. As shown in the last column of Table I, the number of parameters in our model is 17% (0.889 million) fewer than the baseline PSMNet model, which means that our model will save more computational resources in both training and evaluation. Group-wise correlation and dilated residual cost filtering both contribute to the reduction of parameters, which supports our analysis in the methods section.

IV. Discussion

We implemented all three ideas of Improved-PSMNet presented in the proposal and trained and evaluated our models on SceneFlow and KITTI2015 datasets. We analyzed every new component performance by ablation studies and provide the pipeline to apply our models to other new datasets by fine-tuning. The advantage of our model is that it not only improves the accuracy in estimating the disparity but also optimizes the parameters and complex network architecture of PSMNet.

Although Improved-PSMNet achieves expected performance in the experiments, there still exist future works and limitations due to time and computation resources:

1) Conduct the tuning process of hyperparameters of training process and model to achieve the optimal performance.
2) Include more datasets to the model training and evaluation process, for example, full SceneFlow, KITTI2012, ETH3D. That's since we only use the driving dataset from SceneFlow and KITTI2015 in the limited time.
3) Train the model jointly with the segmentation groundtruth label and integrate the segmentation part into the loss function, which compares its performance with our efficient pre-trained segmentation subnetwork.
4) Refine the estimated disparity by geometric and photometric errors.

V. Summary

To summarize, we developed Improved-PSMNet, an improved stereo disparity estimation network based on PSMNet. The correlation and segmentation information is added into cost volume to combine both semantic and spatial information in different respective fields. A dilated residual cost filtering module is applied to find the correct minima in cost volume and to decrease the number of parameters. Based on the results obtained from the experiments on SceneFlow and KITTI2015 dataset, our Improved-PSMNet outperformed the baseline model not only on accuracy but also on efficiency. In addition, the disparity maps clearly shows that our model did better on marginal and detailed areas.

REFERENCES

[1] H. Laga, L. V. Jospin, F. Boussaid, and M. Bennamoun, "A survey on deep learning techniques for stereo-based depth estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1–1, 2020. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2020.3032602

[2] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, no. 1, pp. 7–42, 2002.

[3] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. IEEE, 2005, pp. 807–814.

[4] W. Chuah, R. Tennakoon, R. Hoseinnezhad, A. Bab-Hadiashar, and D. Suter, "Adjusting bias in long range stereo matching: A semantics guided approach," 2020.

[5] Z. Wu, X. Wu, X. Zhang, S. Wang, and L. Ju, "Semantic stereo matching with pyramid cost volumes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[6] Y. Zhang, Y. Chen, X. Bai, S. Yu, K. Yu, Z. Li, and K. Yang, "Adaptive unimodal cost volume filtering for deep stereo matching," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 926–12 934.

[7] S. Tulyakov, A. Ivanov, and F. Fleuret, "Practical deep stereo (pds): Toward applications-friendly deep stereo matching," 2018.

[8] M. Wei, M. Zhu, Y. Wu, J. Sun, J. Wang, and C. Liu, "A fast stereo matching network with multi-cross attention," *Sensors*, vol. 21, no. 18, p. 6016, 2021.

[9] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[10] S. D. P. H. R. K. A. B. Alex Kendall, Hayk Martirosyan and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017.

[11] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3273–3282.

[12] G. Yang, H. Zhao, J. Shi, Z. Deng, and J. Jia, "Segstereo: Exploiting semantic information for disparity estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 636–651.

[13] X. Song, X. Zhao, L. Fang, and H. Hu, "Edgestereo: An effective multi-task learning network for stereo matching and edge detection," 2019.

[14] R. Chabra, J. Straub, C. Sweeney, R. A. Newcombe, and H. Fuchs, "Stereodrnet: Dilated residual stereo net," *CoRR*, vol. abs/1904.02251, 2019. [Online]. Available: http://arxiv.org/abs/1904.02251

[15] S. R. Kaiming He, Xiangyu Zhang and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[16] 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.

[17] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal on Computer Vision*, 2018.

[18] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[19] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[20] H. Hirschmuller, "Stereo processing by semi global matching and mutual information," in *Proceedings of the IEEE Transactions on pattern analysis and machine intelligence*, 2008, p. 30(2):328–341.

[21] S. J. L. Richard A Newcombe and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in *Proceedings of the 2011 IEEE International Conference*, 2011, p. 2320–2327.

[22] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[23] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 472–480.

[24] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 66–75.

[25] P. H. P. F. D. C. A. D. Nikolaus Mayer, Eddy Ilg and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of IEEE Confer- ence on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[26] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

APPENDIX

A. Comparison of disparity map

Fig. 4. Left input image

Fig. 5. Right input image

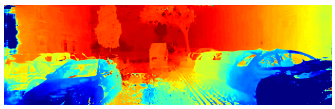

Fig. 6. Baseline

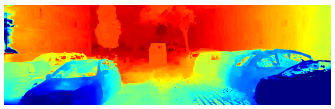

Fig. 7. Group-wise Correlation

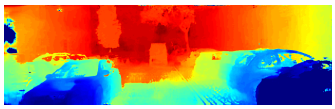

Fig. 8. Dilated Residual

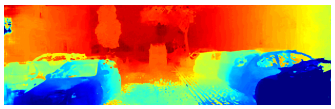

Fig. 9. Segmentation

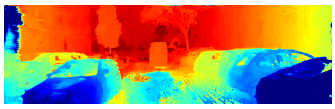

Fig. 10. Group-wise and Dilated

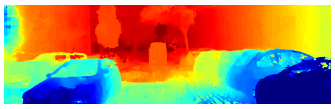

Fig. 11. Group-wise and Segmentation

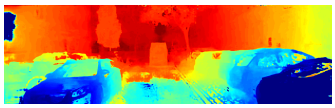

Fig. 12. Dilated and Segmentation

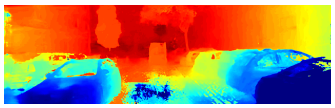

Fig. 13. Improved PSMNet

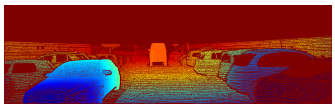

Fig. 14. Groundtruth disparity

B. Training Loss