Deeply-fused Attentive Network for Stereo Matching

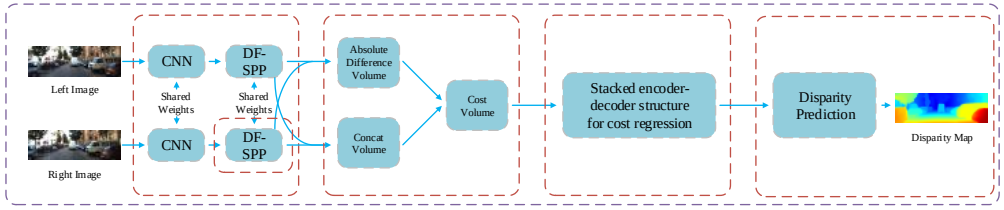
Zuliu Yang[1], Xindong Ai[1], Weida Yang[1], Yong Zhao[1*] Qifei Dai[2] and Fuchi Li[2]
[1]Shenzhen Graduate School of Peking University (PKUSZ), [2]Shenzhen Apical Technology Co.

ABSTRACT

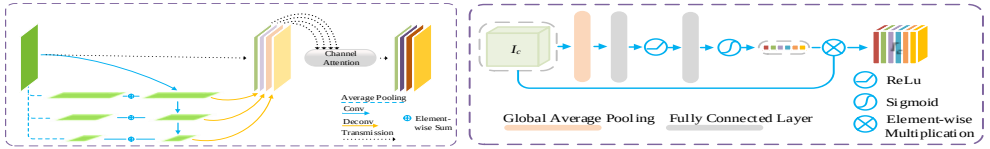
In this paper, we propose a novel learning-based network for stereo matching called DF-Net, which makes three main contributions that are experimentally shown to have practical merit. Firstly, we further increase the accuracy by using the deeply fused spatial pyramid pooling (DF-SPP) module, which can acquire the continuous multi-scale context information in both parallel and cascade manners. Secondly, we introduce channel attention block to dynamically boost the informative features. Finally, we propose a stacked encoder-decoder structure with 3D attention gate for cost regularization. More precisely, the module fuses the coding features to their next encoder-decoder structure under the supervision of attention gate with long-range skip connection, and thus exploit deep and hierarchical context information for disparity prediction. The performance on SceneFlow and KITTI datasets shows that our model is able to generate better results against several state-of-the-art algorithms.

NETWORK ARCHITECTURE

In this section, we describe each component of our network. The network consists of four parts: unary shared feature extraction module, 4D cost volume construction, cost volume regularization and disparity prediction. The whole framework of our network is given as following.

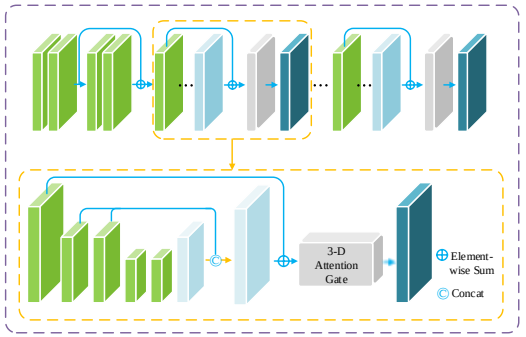


1. Unary shared feature extraction: we redesign the previous Spatial Pyramid Pooling (SPP) architecture[1]. Our approach aims to feed the next-scale features with the former parallel extracted pyramidal features in cascade. This design can bring us two benefits without introducing other complex structures: denser and continuous pyramidal features, and larger receptive field. Moreover, we exploit the SE-Channel Attention[2] module to each up-sampled features before concatenation.



2. Cost volume: we firstly adopt the concatenated 4D cost volume. We further extend the cost volume through calculating the horizontal absolute difference between two input features along disparity dimension, inspired from the sum of absolute difference (SAD) method of traditional algorithm. We then directly concatenate those two different cost volume together as our final cost volume.

3. Stacked encoder-decoder structure with attention gate: In this module, we will use a 3D encoder-decoder structure with skip-connection, besides, we introduce the 3D attention gate module to the encoder-decoder structure before the feature be transmitted to next stage, we stack three this structure in cascade.



4. Disparity regression: to enable stable disparity map estimating, we use a soft argmin function to regress a smooth disparity estimation which was first advanced in GC-Net[3]. Specifically, two 3D convolutions and up-sample operation are employed to the output of every encoder-decoder with new one channel 4D cost output of initial input size. We then squeeze the 4D cost to 3D cost and convert it to probability volume with the softmax operation. The estimated disparity can be calculated by following formula, as

$$\hat{d} = \sum_{d=0}^{D_{max}} d * \sigma(-c_d)$$

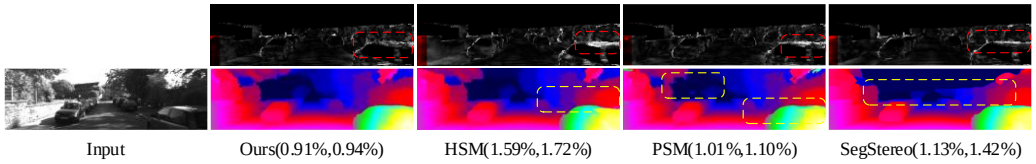
RESULTS

We empirically evaluate our proposed network on three stereo datasets: Scene Flow[4], KITTI 2012[5], and KITTI 2015[6].

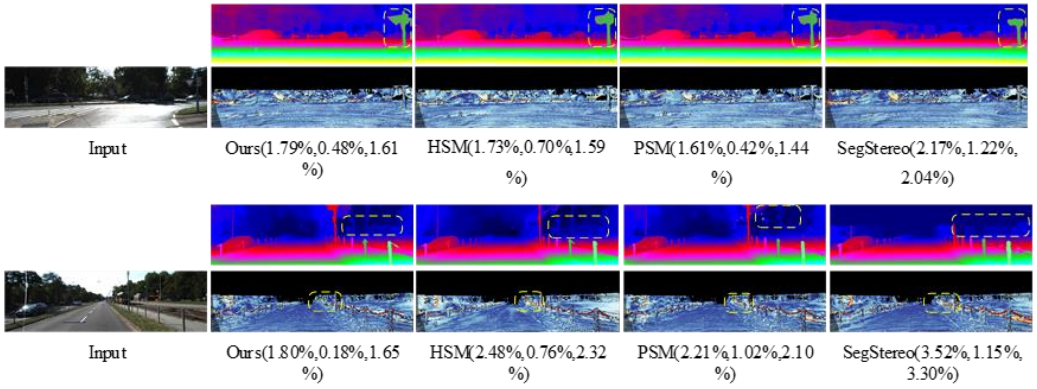
1. Ablation study is tested on SceneFlow dataset, the step-by-step reductions in error rate brought by each component are illustrated in following Table.

Network setting								SceneFlow			
Feature extraction			Cost volume		Cost volume regularization						
SPP	DF-SPP	Attention	Concat	Subtraction	basic	Stacked ED	3D-AG	>1px	>2px	>3px	EPE
√			√		√			9.971	5.650	4.231	0.963
	√		√		√			9.717	5.565	4.191	0.958
	√	√	√		√			9.556	5.459	4.113	0.932
	√	√	√	√	√			8.876	5.058	3.790	0.857
	√	√	√			√		8.634	4.646	3.369	0.799
	√	√	√	√		√		8.108	4.421	3.230	0.757
	√	√	√	√		√	√	7.994	4.368	3.191	0.756

2. The following two figures and tables show the performance comparisons of our network with other stereo networks on KITTI 2012 and KITTI 2015 dataset.



KITTI 2012	>3 px				>5 px			
	Out-Noc	Out-All	Avg-Noc	Avg-All	Out-Noc	Out-All	Avg-Noc	Avg-All
GCNet	1.77%	2.30%	0.6 px	0.7 px	1.12%	1.46%	0.6 px	0.7 px
HSM-1.5x	1.53%	1.99%	0.5 px	0.6 px	0.87%	1.16%	0.5 px	0.6 px
PSM	1.49%	1.89%	0.5 px	0.6 px	0.90%	1.15%	0.5 px	0.6 px
CFP-Net	1.41%	1.83%	0.5 px	0.5 px	0.85%	1.10%	0.5 px	0.5 px
EdgeStereo	1.46%	1.83%	0.4 px	0.5 px	0.83%	1.04%	0.4 px	0.5 px
SegStereo	1.68%	2.03%	0.5 px	0.6 px	1.00%	1.21%	0.5 px	0.6 px
Ours	1.41	1.82	0.5px	0.5px	0.85%	1.10%	0.5 px	0.5 px



KITTI 2015	All Pixels			Noc-Occluded			Runtime(s)
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all	
GCNet [18]	2.21	6.16	2.87	2.02	5.58	2.61	0.900
HSM-1.5x []	1.95	3.93	2.28	1.76	3.55	2.06	0.085
PSM	1.86	4.62	2.32	1.71	4.31	2.14	0.410
CFP-Net	1.90	4.39	2.31	1.73	3.92	2.09	0.900
EdgeStereo	1.87	3.61	2.16	1.72	3.41	2.00	0.700
SegStereo	1.88	4.07	2.25	1.76	3.70	2.08	0.600
Ours	1.78	4.03	2.15	1.62	3.65	1.95	0.700

CONCLUSION

In this work, we propose a novel end-to-end network for stereo matching, named DF-Net. Our aim in this work is firstly to extract continuous pyramidal features with DF-SPP and attention block, then we form the extended cost volume and finally we use the SEDA to regularize cost volume and produce the disparity map. The ablation study indicates that the proposed structures can effectively learn the informative features, and lead to better disparity map prediction. Furthermore, we show the advantages of DF-Net compared with several state-of-the-art networks on KITTI 2012 and KITTI 2015 datasets. In the future work, we are interested in introducing a more effective network work which can trade off speed and accuracy.

REFERENCE

[1] Nguyen, Tien Phuoc, and Jae Wook Jeon. "Wide context learning network for stereo matching." Signal Processing-image Communication (2019): 263-273.

[2] Hu, J. , Shen, L. , Albanie, S. , Sun, G. , & Wu, E. . (2017). Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, PP(99).

[3] Kendall, Alex, et al. "End-to-end learning of geometry and context for deep stereo regression." Proceedings of the IEEE International Conference on Computer Vision. 2017.

[4] Mayer, Nikolaus, et al. "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[5] Geiger, Andreas, Philip Lenz, and Raquel Urtasun. "Are we ready for autonomous driving? the kitti vision benchmark suite." 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012.

[6] Menze, Moritz, and Andreas Geiger. "Object scene flow for autonomous vehicles." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.