Exploring Stereovision-Based 3-D Scene Reconstruction for Augmented Reality

Guang-Yu Nie
Beijing Institute of Technology, China

Yun Liu
Nankai University, China

Cong Wang
China Electronics Standardization Institute, China

Yue Liu*
Beijing Institute of Technology, China
AICFVE of Beijing Film Academy, China

Yongtian Wang
Beijing Institute of Technology, China
AICFVE of Beijing Film Academy, China

Abstract

Three-dimensional (3-D) scene reconstruction is one of the key techniques in Augmented Reality (AR), which is related to the integration of image processing and display systems of complex information. Stereo matching is a computer vision based approach for 3-D scene reconstruction. In this paper, we explore an improved stereo matching network, **SLED-Net**, in which a Single Long Encoder-Decoder is proposed to replace the stacked hourglass network in PSM-Net for better contextual information learning. We compare SLED-Net to state-of-the-art methods recently published, and demonstrate its superior performance on Scene Flow and KITTI2015 test sets.

Index Terms: Computing methodologies—Scene understanding; Computing methodologies—Reconstruction; Computing methodologies—Mixed/augmented reality

1 Introduction

3-D scene reconstruction is one of the most critical techniques in AR, which has been developed with substantial effort and can be conducted by either traditional surveying or novel 3-D modeling systems. Traditional remotely sensed reconstruction techniques include two major methods, i.e., airborne image photogrammetry and LiDAR (Light Detection And Ranging), but these techniques suffer from such problems as time consuming, high financial and equipment costs, difficulty of cloud point processing, and so on. To overcome the disadvantages of traditional methods, stereo matching has been introduced into 3-D modeling systems in recent years. PSM-Net [1] is one of the state-of-the-art stereo matching methods, and contains stacked hourglass networks for cost volume regularization. However, such networks pay less attention to the local appearance and are thus not suitable for stereo matching. In this paper, we propose a *Single Long Encoder-Decoder network* (SLED-Net) to replace the stacked hourglass network used in PSM-Net [1]. The experimental results demonstrate the effectiveness of our new module in stereo matching.

2 Single Long Encoder-Decoder

The full framework of SLED-Net is illustrated in Fig. 2. We design a Single Long Encoder-Decoder (SLED, with light orange color) to replace the stacked hourglass networks in PSM-Net [1]. The SLED is set up as follows: **Encoder:** Convolutional and average pooling layers are alternatively stacked to process the feature maps down to a resolution with $\frac{1}{32}\times$ scale from that with $\frac{1}{4}\times$ scale. In details, as shown in Fig. 2, the encoder totally consists of eight residual blocks, and in each block an element-wise sum is applied to fuse the features generated by a convolution operation and the features generated by the previous block, and then the fused features are feed into the subsequent block. **Decoder:** After reaching the lowest resolution, the network begins the top-down sequence of upsampling and combination of features across different scales. To bring together information across two adjacent resolutions, we upsample the features with lower resolution by trilinear interpolation. Then, we adopt an element-wise sum to combine the upsampled features and the larger features from the encoder with skip layers, and an operation closely following sum operation to fuse the combined features through a residual block which contains a 3×3 atrous convolution with dilation of 2.

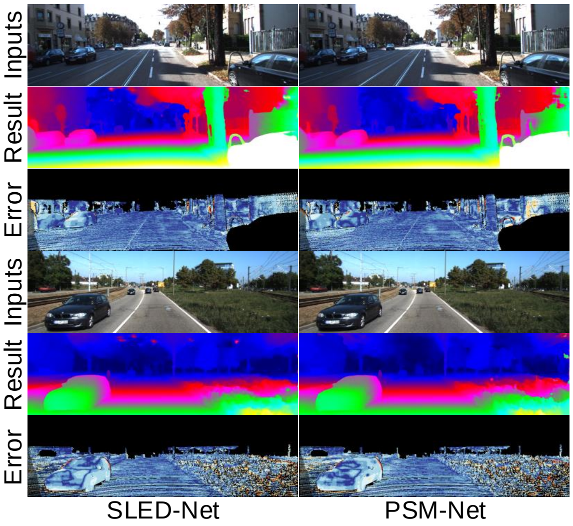

Figure 1: Results of our model and PSM-Net [1] in KITTI 2015 dataset

Table 1: Performance comparison on Scene Flow Validation set

Mod.	EPE	Mod.	EPE	Mod.	EPE	Mod.	EPE
SLED-Net	0.699	PSM-Net [1]	1.09	CRL [4]	1.32	iResNet [2]	1.40

Mod.: model; **EPE**: Average disparity/end-point-error.

To enhance the performance of SLED-Net, the initial cost volume is used as an intermediate supervision [3].

3 Implementation Details

Datasets We adopt two publicly available datasets for training and testing: The Scene Flow datasets and KITTI2015 dataset

Training The training of SLED-Net follows the process described in PSM-Net [1]. We implement SLED-Net using PyTorch

[*] e-mail: liuyue@bit.edu.cn

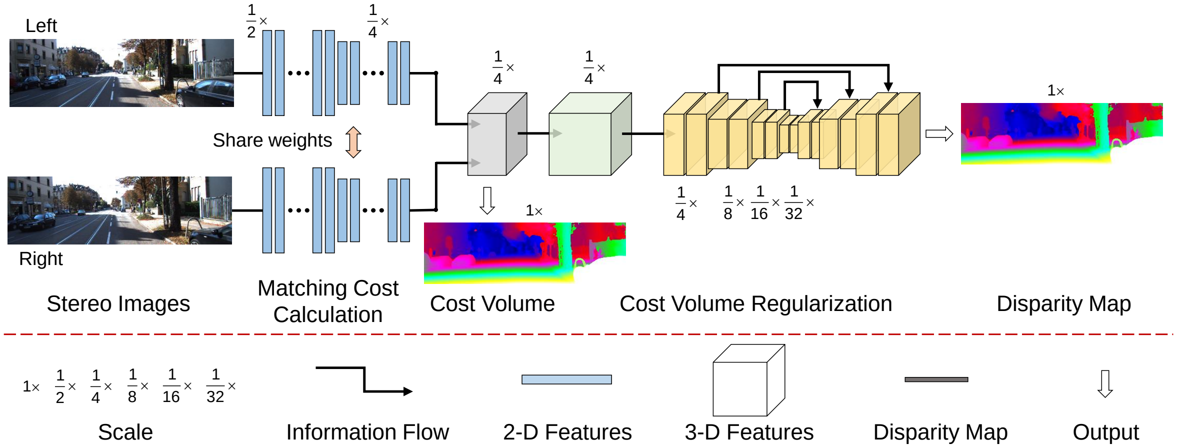
Left
Right
Share weights
$\frac{1}{2}\times$
$\frac{1}{4}\times$
$\frac{1}{4}\times$
$\frac{1}{4}\times$
1×
$\frac{1}{4}\times$ $\frac{1}{8}\times$ $\frac{1}{16}\times$ $\frac{1}{32}\times$
1×
Stereo Images
Matching Cost Calculation
Cost Volume
Cost Volume Regularization
Disparity Map
1× $\frac{1}{2}\times$ $\frac{1}{4}\times$ $\frac{1}{8}\times$ $\frac{1}{16}\times$ $\frac{1}{32}\times$
Scale
Information Flow
2-D Features
3-D Features
Disparity Map
Output

Figure 2: The diagrammatic sketch of SLED-Net. It is based on PSM-Net [1], but replaces the stacked hourglass networks with SLED (Blocks with Light Orange color). A pair of stereo images (i.e., Left, Right) passes through the network to generate the disparity prediction (i.e., Disparity Map).

Table 2: KITTI2015 Results

Mod.	All (%)			Noc (%)		
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all
SLED-Net	**1.85**	4.15	**2.23**	**1.70**	3.68	**2.02**
PSM-Net [1]	1.86	4.62	2.32	1.71	4.31	2.14
iResNet [2]	2.25	**3.40**	2.44	2.07	**2.76**	2.19
CRL [4]	2.48	3.59	2.67	2.32	3.12	2.45

"**All/Noc**" : In total/non-occluded regions; "**D1-bg/fg/all**": Three-point error in background/foreground/all regions.

Table 3: Ablation Study for Cost Volume Regularization

Mod.	Scene Flow				KITTI2015	Para.
	$> 1px$	$> 3px$	$> 5px$	EPE	D1-all (%)	
SCC-Net	12.268	5.213	3.884	1.273	2.142	3.84M
1 HG	9.145	4.045	2.940	0.939	1.995	4.06M
2 HGs	8.369	3.360	2.341	0.816	1.800	4.64M
3 HGs	8.335	3.312	2.303	0.787	1.754	5.22M
SLED-Net	7.044	3.039	2.098	0.699	1.728	4.86M

$>$***t*px**: EPE larger than **t** pixels; **Para.**: number of parameters.

and conduct experiments on four NVIDIA TITAN Xp GPUs. We first train SLED-Net on Scene Flow datasets with a fixed learning rate of 0.001 for 20 epochs, then we fine-tune the network on KITTI2015 dataset with stepped learning rates of 0.001 for 600 epochs and 0.0001 for another 400 epochs.

4 Experimental Results

We compare SLED-Net with three state-of-the-art approaches. The evaluation results on the Scene Flow test set and KITTI2015 test set are summarized in Tab. 1 and Tab. 2, respectively. SLED-Net achieves the superior performance on both two datasets. Compared with PSM-Net [1], although the number of parameters in SLED-Net reduces by 0.36M, SLED-Net surpasses PSM-Net by 35.9% in terms of end-point-error on Scene Flow dataset, and by 3.9% in terms of the overall three-pixel-error on KITTI 2015 dataset, respectively. We display some examples from KITTI2015 dataset in Fig. 1.

5 Ablation Study

To explore the reasons why SLED is better than stacked hourglass networks in stereo matching, we design some ablation studies: 1) We compare PSM-Net [1] with a single hourglass network (**1HG**) to a baseline network with four Simply Cascaded Convolution layers (**SCC-Net**). In details, the encoder of the single hourglass network in 1HG contains four 3×3 convolutional layers with stride of 2, SCC-Net is generated through replacing the single hourglass network in 1HG with a plain network consisting of four 3×3 convolution layers with stride of 1. As shown in Tab. 3, the performance has 26.2% increase in Scene Flow test set and 6.9% increase in KITTI2015 test set when using a single encoder-decoder network but with approximately the same number of layers and parameters. 2) We conduct experiments to demonstrate the stacked design by increasing the stacked number of hourglass network from 1 to 3, which corresponds to **1HG, 2HGs**, and **3HGs** in Tab. 3. During the training the weights of all loss functions are set to 1, which is different from such setting in PSM-Net [1]. Results show that the pixel-wise estimation can be better refined with the number of staked hourglass networks increasing. 3) We explore whether **SLED** is better than the stacked design (**3 HGs**) in stereo matching or not. As shown in Tab. 3, compared with 3HGs, the performance of SLED-Net has 11.2% increase in Scene Flow test set and 1.5% increase in KITTI2015 test set, which demonstrates that the change in network architecture does play the core role, and the stacked hourglass network in PSM-Net [1] limits the ability of the regularization for cost volume.

6 Conclusion

In this paper, we present a single long encoder-decoder to replace the stacked hourglass networks in PSM-Net [1] when regularizing the cost volume in stereo matching. The cost volume is encoded by several simply cascaded convolutional and pooling layers and then decoded by consecutive trilinear interpolation and 1×1 convolution operations to refine the features with low resolutions. The experimental results show that SLED-Net achieves superior performance on both Scene Flow datasets and KITTI2015 benchmark, which has been demonstrated that SLED is better than stacked design, and SLED-Net enables to provide accurate results for 3-D scene reconstruction in AR.

Acknowledgments

This work has been supported by the National Key Research and Development Program of China (No. 2016YFB0401202) and the National Natural Science Foundation of China (Grant No.61731003).

References

[1] J.-R. Chang and Y.-S. Chen. Pyramid stereo matching network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5410–5418, 2018. 1, 2

[2] Z. Liang, Y. Feng, Y. G. H. L. W. Chen, and L. Q. L. Z. J. Zhang. Learning for disparity estimation through feature constancy. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2811–2820, 2018. 1, 2

[3] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *Eur. Conf. Comput. Vis.*, pp. 483–499, 2016. 1

[4] J. Pang, W. Sun, J. S. Ren, C. Yang, and Q. Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *IEEE Conf. Comput. Vis. Pattern Recog.*, vol. 7, 2017. 1, 2