One-stage deep stereo network

Ziming Liu, Ezio Malis, Philippe Martinet

ONE-STAGE DEEP STEREO NETWORK

Ziming Liu [1,2]*, Ezio Malis* [1]*, Philippe Martinet* [1]

[1] Inria, Sophia Antipolis, France [2] Université Côte d'Azur, Nice, France
{firstname.lastname}@inria.fr

ABSTRACT

Stereo matching is one of the low-level visual perception tasks. Currently, two-stage 2D-3D networks and three-stage recurrent networks dominate deep stereo matching. These methods build a cost volume with low-resolution stereo feature maps, which splits the network into a feature net and a matching net. However, the 2D feature map is not uncontrollable, and the low-resolution feature map has lost important matching information. To overcome these problems, we propose the first one-stage 2D-3D deep stereo network, named StereoOne. It has an efficient module that builds a cost volume at image resolution in real-time. The feature extraction and matching are learned in a single 3D network. According to the experiments, the new network can easily surpass the 2D-3D network baseline and it can achieve competitive performance with the state-of-the-art.

Index Terms— Stereo Matching, Depth Estimation, Deep Learning

1. INTRODUCTION & RELATED WORKS

As one of the most important computer vision tasks, stereo matching has experienced several stages of development, including traditional algorithms (e.g., SGM [1]), early convolutional networks (e.g., MC-CNN [2]), two-stage 2D-3D stereo networks (e.g., PSMNet [3]), and three-stage recurrent stereo networks [4, 5]. In the age of deep learning techniques, state-of-the-art stereo matching methods have been dominated by deep stereo matching networks. Among them, the most successful solutions are the two-stage 2D-3D CNN methods [3, 6]. Even the newest state-of-the-art recurrent stereo networks [5] highly rely on the two-stage stereo networks [3].

The two-stage methods were proposed in [7] and [3]. The original two-stage stereo networks consist of a 2D feature extraction network and a 3D matching network. Since then, many methods have been proposed based on this two-stage architecture, such as [6] and [8].

More recently, three-stage recurrent stereo networks have achieved state-of-the-art accuracy performance [5, 4, 9]. These methods perform well on high-resolution stereo images, but their inference speed is affected by the time-consuming design of recurrent GRU units. Recurrent stereo networks are also based on the two-stage stereo network [5].

The previous methods build the cost volume on the low-resolution feature maps. The important matching information has been lost because the stereo matching is performed at a low-resolution. This is an primary problem for the 2D-3D networks [3] or recurrent networks [9]. Besides the spatial resolution, the disparity resolution in cost volume also affects the stereo matching.

In addition, the stereo 2D feature maps are not optimal in the previous methods [3, 10]. The 2D-3D network methods optimize the feature network and the matching network by minimizing the loss between the ground truth disparity and the prediction. The optimization goal is to minimizing the matching cost, instead of extracting high-quality feature maps. There are some methods [10] have shown that the extracted 2D feature maps are not suitable to build a single peak cost volume. They propose a constraint loss after the 2D feature network to reduce this effect.

Instead, if the stereo cost volume is built at the raw image resolution and the feature extraction and the feature matching are learned in one 3D network, the matching information lost can be reduced and the conflict of the optimal goal between the feature network and the matching network can be avoided.

In this paper, we propose the first one-stage 3D stereo network, namely StereoOne. StereoOne build the stereo cost volume on the raw stereo images with a new image volume module. We propose an efficient and real-time volume generation methods which is much faster than the previous methods [3, 9]. Furthermore, we introduce general 3D network [11, 12] to learn the feature extraction and matching.

Furthermore, a disparity dense-sparse network is introduced to maintain a high-resolution disparity in the cost volume. On the one hand, the disparity-dense network is low-cost for high-resolution disparity. On another hand, the dense-sparse design is easier to process different disparity range scales of different samples.

This paper is organized as follows: In Section 1, we introduce the background and motivation for our proposed one-stage stereo matching network. In Section 2, we describe the details of the proposed StereoOne. In Section 3, we present our experimental results on popular stereo matching benchmarks. Finally, we conclude our work and discuss future research directions in Section 4.

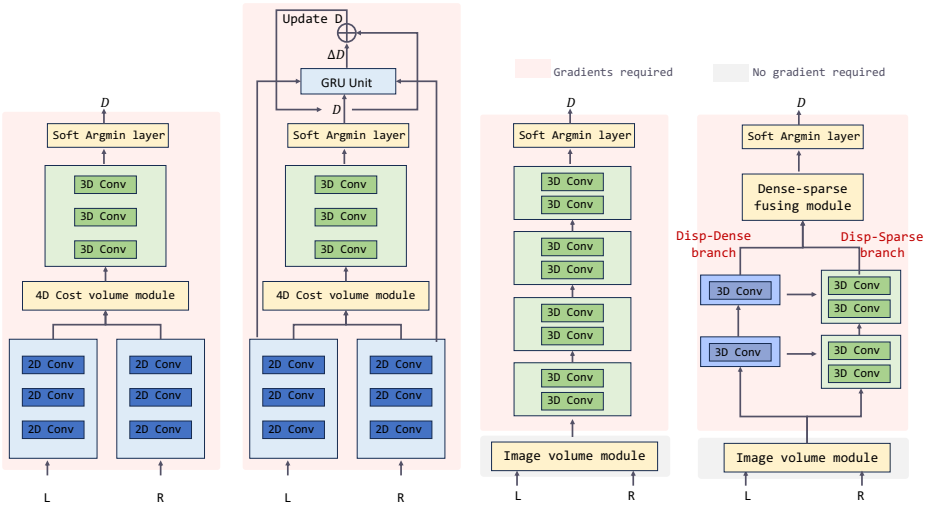

Fig. 1: The structure of two-stage 2D-3D stereo networks [7, 3], three-stage recurrent stereo network [4, 9, 5], one-stage 3D stereo network (single branch) and one-stage 3D stereo network (dense-sparse).

2. METHOD

In this section, we describe the details of the proposed 3D stereo network, namely StereoOne. The structure comparison is shown in Fig. 1.

2.1. An Efficient Image Volume Module

As the cost volume $\mathcal{V}$ is built by enumerating all aligned stereo images of different disparities.

$$\mathcal{V} = ([\mathbf{L}, f(\mathbf{R})_{\mathbf{P}_x+d}]_c, ...,)_{d=0}^{D-1} \tag{1}$$

Where $[L,R]_c$ is the concatenating of left image L and right image R on the channel dimension. $\mathbf{P}_x$ is the x direction image coordinate. $f(.)_{\mathbf{P}_x+d}$ is the d pixels image horizontal shift.

The previous methods use two ways to build cost volume: image warping [9, 10] or looping-index [3], as Eq. 2.

$$\begin{aligned} &Warping: \ \mathcal{V} = [\mathcal{L}, \ \mathcal{W}_{3D}(R)_{(\mathcal{P}_x+d)_{d=0}^{D-1}, \mathcal{P}_y}]_c \\ &Looping-index: \ \mathcal{V} = ([\mathbf{L}, \underline{\mathbf{R}[:, \ 0:W-d]}]_c \ for \ d \in [0, D)) \end{aligned} \tag{2}$$

Where $\mathcal{W}_{3D}$ is a 3D image warping. $\mathbf{R}[:, \ 0:W-d]$ indexes the aligned part of the $\mathbf{R}$. Usually, They can not realize real-time inference. Therefore, we propose an efficient volume generation method (EffiVolume).

Before all algorithms launch, we first compute a index matrix $\mathbf{I}_{D\times W}$ (Eq. 3) which indexes the disparity and image horizontal dimension for right image. $\mathbf{I}_{D\times W}$ is only computed once.

$$\begin{aligned} \mathbf{I} = &((0,1,...,w)_{w=0}^{W-1}, ...)_{\times D} \\ &- ((0,...)_{\times W}, ..., (d,...)_{\times W})_{d=0}^{D-1} \end{aligned} \tag{3}$$

Then, the stereo images $\mathbf{I}$ and $\mathbf{R}$ and the index matrix $\mathbf{I}$ are expanded to the size $C\times D\times H\times W$, denoted as $\mathcal{L}, \mathcal{R}, \mathcal{I}$.

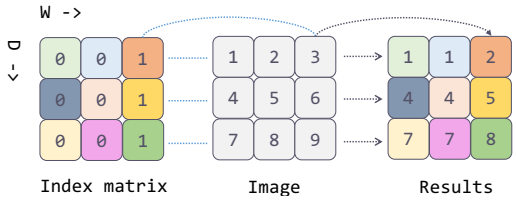

Fig. 2: The efficient image volume index on disparity and image width dimension.

With the index tensors, we can easily obtain the aligned right image $f(\mathcal{R})$ in one step (e.g. using *torch.gather* in pytorch). This process is shown in Fig. 2.

Finally, the raw cost volume is generated by concatenating them $\mathcal{V} = [\mathcal{L}, \mathcal{R}]_c$. Our method avoid the time-consuming loop operation, and be much faster than the previous on both CPU and GPU devices.

2.2. 3D Network

2.2.1. Overall

In StereoOne, the raw cost volume is built on raw stereo images. StereoOne network introduces a general 3D network to process this image volume. In this way, the 3D network not only learns the stereo features but also learns the feature matching in one network. StereoOne is based on a general 3D network, which makes it to be flexible and easier to be deployed. Specifically, the 3D network is an encoder-decoder structure. For the encoder part, the general 3D encoder networks [13, 12] can be used. For the decoder part, we designed a 3D feature pyramid network (FPN) based on 2D FPN [14]. In the 3D FPN, feature maps of four stages are mapped to the same channel size with FPN lateral connection layers, which is a 3D convolution layer ($kernel = (1,1,1), stride = (1,1,1)$). Then, these feature maps are summed from the top

Method	KITTI2012(3pixel Error)			KITTI2015
	Out-Noc%	Out-All/%	Avg-All/px	D1-fg
PSMNet 2018	8.36	10.18	1.6	4.62
ACVNet 2022	7.03	8.67	1.5	3.07
AcfNet 2020	6.93	8.52	1.9	3.80
CoEX 2021	6.83	8.63	1.4	3.41
SegStereo 2018	6.35	8.06	1.3	4.07
CREStereo 2022	6.27	7.27	1.4	2.86
GANet 2019	6.22	7.92	1.3	3.46
HITNet 2021	5.91	7.54	1.2	3.20
LEAStereo 2020	5.35	6.50	1.2	2.91
CFNet 2023	5.96	7.29	1.3	3.56
CroCo-Stereo 2023	-	-	-	2.65
Ours 2023	**4.99**	**6.50**	**1.2**	**2.62**

Table 1: The error results on the online benchmark KITTI2012, KITTI2015. The error metrics are described in benchmarks.

to the bottom. Each feature map is upsampled $\times 2$ size. Finally, the bottom feature are mapped to channel $= 1$.

Same as the previous, with the learned cost volume, a soft Argmin disparity prediction layer is used following the previous stereo networks [7, 3].

Furthermore, because the different images have different disparity ranges, we propose a disparity dense-sparse encoder network and a dense-sparse fusing module to learn a better stereo cost volume.

2.2.2. Disparity Dense-sparse Encoder

Disp-Dense Branch: The dense disparity branch maintains a high disparity resolution, it focuses on learning the matching information. It has a shallow feature dimension to reduce the computation cost, usually $1/8$ of the sparse disparity branch.

Disp-Sparse Branch: In contrast, the sparse branch focuses on learning the image's spatial information and extracting better features. Each pixel on the predicted disparity map is not independent, the spatial context information is important for the disparity prediction, especially the homogeneous area [15]. Therefore, the sparse disparity branch keeps a higher channel dimension but a low disparity resolution to learn abundant image spatial information.

The Feature Connection From the Dense to the Sparse: To better fuse dense-sparse information, the dense feature will be fused into the sparse feature at the stage $1, 2, 3$ of the encoder (four stages structure). We only use single-direction connection according to the previous deep learning networks [12, 14]. The single-direction connection has a similar performance as the bi-direction connection but has a lower computation cost.

The Dense-Sparse Fusing Module

To fuse the two volumes of the disparity dense and sparse branches, we explore different ways to fuse dense and sparse cost volumes. (i) Concatenate them across the disparity dimension. Then use a linear layer to transform the disparity resolution to the original. (ii) Add up the two cost volumes. Firstly transform them into the original disparity resolution, then add up two volumes. (iii) Concatenate them across the channel dimension. Firstly up-sample them to the original disparity resolution, then concatenate them in the channel, and finally transform it with a linear layer. With the experiments, the third module has the best performance.

3. EXPERIMENTS

3.1. The Comparison with the State-of-the-art Methods

Method	Error(EPE)	Speed(FPS)	FLOPS
PSMNet [3]	0.98	9	1.02T
Cascade-PSM [6]	0.93	23	1.37T
Cascade-Gwc [6]	0.81	19	1.49T
AcfNet [10]	0.87	9	1.02T
CREStereo [9]	0.78	5	2.27T
† Ours(ResNet18 [11, 12])	**0.73**	12	1.48T
CoEX [16]	1.14	81	0.04T
StereoNet [17]	1.29	65	0.11T
CGIStereo [18]	1.51	48	0.06T
† Ours(ResNet8 [11, 12])	1.22	50	0.09T
⋆ Ours(MobileNetv2 [19])	**0.89**	28	0.05T

Table 2: The comparison with other methods on Scene Flow data test set. The model code is from the official release, all experiments use the same optimizer. Device: Nvidia A40 GPU. †: disparity dense-sparse network. ⋆: single branch.

We compare the StereoOne with the other methods on three benchmarks: the large-scale SceneFlow data, KITTI2012, 2015 Stereo data. Firstly, as the Tab 2 shows, the StereoOne achieves the lowest error compared with most recent methods.

Then we report the results on KITTI 2012, 2015 stereo data as Tab. 1. It suggests that StereoOne can achieve competitive results with state-of-the-art methods on real-world data.

3.2. The Ablation Studies

3.2.1. *Image Volume Module*

Device	Method	time/ms	memory/M
A40GPU	Warping	323	14,154
A40GPU	Looping	421	4,938
A40GPU	Ours	18	7,244
2080TiGPU	Warping	678	9,422
2080TiGPU	Looping	1318	4,814
2080TiGPU	Ours	20	7,118
CPU	Warping	2181	5,068
CPU	Looping	221	5,060
CPU	Ours	204	5,066

Table 3: The performance of different volume generation methods. GPU capability: A40(8.6), 2080Ti(7.5), CPU: AMD EPYC 7413 24-Core.

As shown in Tab. 3, we compare three methods. The proposed efficient volume method achieves much fastest inference speed and keep a low memory cost. Compared with the simple looping-index operation or 3D image-warping operation, the proposed module is $\times 66$ and $\times 34$ faster than them on 2080Ti GPU. The results suggest that our method has robust performance on different devices.

3.2.2. *The Disparity Distribution*

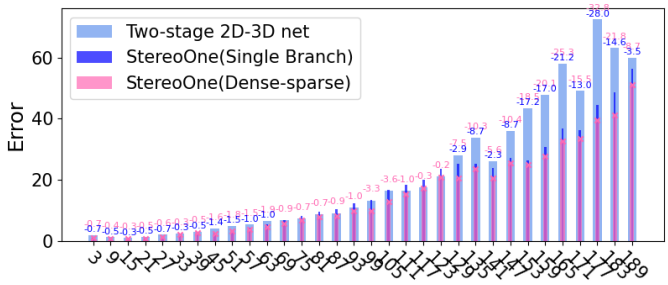

Fig. 3: The results on different disparity distributions on SceneFlow dataset. The two-stage network is PSMNet [3]. Every pixel error is used to measure the predictions.

Further, we compare the results on different disparity distributions for the one-stage 3D and two-stage 2D-3D stereo networks. Fig. 3 suggests that StereoOne with disparity dense-sparse branches can solve the problem of different disparity distribution as we claim, especially for high disparities.

3.2.3. *Disparity Resolution*

In this part, we explore different disparity resolutions for the dense and sparse branches. We use $6, 12, 24$ for the sparse branch, $48, 96, 192$ for the dense branch. The error results are shown in Fig. 4. Overall, the result is better with high disparity resolution. This demonstrates the disparity-dense and sparse structure. The result of 192 also suggests that the dense/sparse ratio should not be too large.

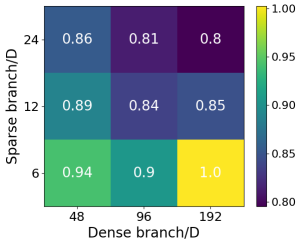

Fig. 4: The predicted disparity EPE errors with different disparity resolution $/D$.

Fusion	EPE	3PE	D1
D-cat	0.932	0.046	0.044
Add	0.867	0.046	0.043
C-cat	0.752	0.044	0.042

Table 4: The disparity error results of different fusion methods for the dense and sparse costs volumes. D-cat: concatenate the volumes in disparity dimension. Add: add up the volumes. C-cat: concatenate the volumes in channel dimension.

3.2.4. *Dense-Sparse Fusing Module*

To fuse the dense and sparse cost volume efficiently, we conduct experiments of three kinds of modules as described in Sec. 2.2.2. The results as recorded in Tab. 4. The Channel-concatenating has the lowest errors on every-pixel-error, 3pixels-error, and D1 [1] error metrics.

3.3. Details and Datasets

To evaluate the performance of StereoOne, we use the large-scale dataset Scene Flow [20] and KITTI 2012/2015 benchmarks [21]. In the experiments on Scene Flow datasets, we use Lion optimizer with 1e-4 learning rate, and the optimizer uses gradient clip norm with L2 norm, the max norm is 35 for 48k iterations. For the experiments on KITTI, a AdamW optimizer with learning rate 2e-4 is used for 20k iterations. The gradient clip norm is also used. The batch size is 8 on two GPUs for both of them.

4. CONCLUSION

By reviewing the previous stereo networks, we analyzed the limitations of the previous. To realize the excellent disparity prediction, we propose the first one-stage 3D stereo network. In this paper, we use the simple ResNet [13] structure, more advanced network structures can be explored in the future.

Acknowledgments

This work was funded by 3IA Côte d'Azur. The experiments use OPAL computing cluster of INRIA and UCA.

[1] D1: Percentage of stereo disparity outliers in first frame

5. REFERENCES

[1] Heiko Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *CVPR*. IEEE, 2005, vol. 2, pp. 807–814.

[2] Jure Zbontar, Yann LeCun, et al., "Stereo matching by training a convolutional neural network to compare image patches.," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2287–2318, 2016.

[3] Jia-Ren Chang and Yong-Sheng Chen, "Pyramid stereo matching network," in *CVPR*. IEEE, 2018, pp. 5410–5418.

[4] Lahav Lipson, Zachary Teed, and Jia Deng, "Raft-stereo: Multilevel recurrent field transforms for stereo matching," in *International Conference on 3D Vision*, 2021.

[5] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang, "Iterative geometry encoding volume for stereo matching," in *CVPR*, 2023.

[6] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *CVPR*, 2020, pp. 2495–2504.

[7] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry, "End-to-end learning of geometry and context for deep stereo regression," in *ICCV*. IEEE, 2017, pp. 66–75.

[8] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li, "Group-wise correlation stereo network," in *CVPR*, 2019, pp. 3273–3282.

[9] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu, "Practical stereo matching via cascaded recurrent network with adaptive correlation," in *CVPR*, 2022, pp. 16263–16272.

[10] Youmin Zhang, Yimin Chen, Xiao Bai, Suihanjin Yu, Kun Yu, Zhiwei Li, and Kuiyuan Yang, "Adaptive unimodal cost volume filtering for deep stereo matching," in *AAAI*, 2020, pp. 12926–12934.

[11] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017, pp. 6299–6308.

[12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He, "Slowfast networks for video recognition," in *ICCV*. IEEE, 2019, pp. 6202–6211.

[13] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017, pp. 6299–6308.

[14] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017, pp. 2117–2125.

[15] S. Mahdi H. Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yağız Aksoy, "Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging," 2021.

[16] Antyanta Bangunharcana, Jae Won Cho, Seokju Lee, In So Kweon, Kyung-Soo Kim, and Soohyun Kim, "Correlate-and-excite: Real-time stereo matching via guided cost volume excitation," in *IROS*. IEEE, 2021, pp. 3542–3548.

[17] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi, "Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction," in *ECCV*, 2018, pp. 573–590.

[18] Gangwei Xu, Huan Zhou, and Xin Yang, "Cgi-stereo: Accurate and real-time stereo matching via context and geometry interaction," *arXiv preprint arXiv:2301.02789*, 2023.

[19] Okan Köpüklü, Neslihan Kose, Ahmet Gunduz, and Gerhard Rigoll, "Resource efficient 3d convolutional neural networks," in *ICCV Workshop*. IEEE, 2019, pp. 1910–1919.

[20] N. Mayer, E. Ilg, P. Häusser, 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 *CVPR*, 2016.

[21] Moritz Menze and Andreas Geiger, "Object scene flow for autonomous vehicles," in *CVPR*, 2015, pp. 3061–3070.