J. Shanghai Jiao Tong Univ. (Sci.), 2022
https://doi.org/10.1007/s12204-022-2435-4

Multilevel Disparity Reconstruction Network for Real-Time Stereo Matching

LIU Zhuoran (刘卓然), *ZHAO Xu** (赵 旭)
(School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: Recently, stereo matching algorithms based on end-to-end convolutional neural networks achieve excellent performance far exceeding traditional algorithms. Current state-of-the-art stereo matching networks mostly rely on full cost volume and 3D convolutions to regress dense disparity maps. These modules are computationally complex and high consumption of memory, and difficult to deploy in real-time applications. To overcome this problem, we propose multilevel disparity reconstruction network, MDRNet, a lightweight stereo matching network without any 3D convolutions. We use stacked residual pyramids to gradually reconstruct disparity maps from low-level resolution to full-level resolution, replacing common 3D computation and optimization convolutions. Our approach achieves a competitive performance compared with other algorithms on stereo benchmarks and real-time inference at 30 frames per second with 4×10^4 resolutions.
Key words: stereo matching, disparity reconstruction, real-time, stacked residual pyramid
CLC number: TP 183 **Document code:** A

0 Introduction

Depth estimation and stereo matching from binocular image pairs as one of the most fundamental computer vision tasks[1], have a range of practical applications including autonomous driving[2], 3D reconstruction, and robotics[3]. Given a stereo image pair, the key of stereo matching is to compute the horizontal offset called disparity d between a pixel (x, y) on the left image and its corresponding pixel $(x-d, y)$ on the right image. Then the depth z of pixel (x, y) can be calculated by

$$z = \frac{fB}{d}, \tag{1}$$

where f is the binocular camera's focal length, and B is the distance between binocular camera centers.

Traditional stereo matching methods commonly follow a four-step optimization pipeline[4]: matching cost computation, cost aggregation, optimization, and disparity refinement. Almost of them are roughly classified as global and local methods. The global methods commonly solve the optimization problem by minimizing global objective functions including data and smooth terms[5-6]. The local methods focus on neighbor features[7-8] and get faster than global methods[9]. Although traditional methods have significant progress, they still perform poorly in difficult scenarios of stereo matching such as large no texture areas and thin structures.

Recently, based on end-to-end convolutional neural networks (CNNs), stereo matching algorithms make remarkable progress and achieve better performance than traditional algorithms. DispNetC[10] is the first end-to-end CNNs for stereo matching by training to measure the similarity between left and right pixels. GC-Net[11] first uses 3D convolutions to compute and optimize disparity maps from 4D cost volume aggregated by left and right image features. On this basis, PSMNet[12] adds pyramid pooling module and stacked hourglass 3D CNN to extend the context information support and further improve the accuracy. GA-Net[13] uses two guided aggregation layers assisting 3D CNN, reduces computational complexity of the whole network, and achieves state-of-the-art performance on KITTI stereo benchmarks[2,14].

Although recent state-of-the-art works get amazing results on real-world datasets, rare of them can make inference less than 100 ms and require relatively few computational resources. 3D convolutions provide significant improvement in the ability of model to measure disparity features and strong regularization, with a large cost of inference time and memory consumption. Another recent work, AANet[15], proposes deformable convolution and local and cross-scale cost aggregation to completely replace 3D convolutions, but deformable

Received: 2020-12-18 **Accepted:** 2021-03-19
∗**E-mail:** zhaoxu@sjtu.edu.cn

convolutions modules also require high performance hardware's assist.

To this end, we propose a multilevel disparity reconstruction network (MDRNet), consisting of encoder-decoder 2D convolutions and stacked residual pyramids (SRPs) without 3D convolutions. Instead of computing and optimizing disparity maps using 3D CNN modules from 4D cost volume, we design multilevel stacked 2D convolutions trained to reconstruct dense disparity from lower resolution disparity hypothesis and current resolution feature maps. And multiple residual modules and spatial pyramid pooling module can assist to expand the receptive field and make effective utilization of multi-scale context information. In this way, disparity feature maps in the process of forward propagation gradually restore details from low resolution to high resolution. Through cross-datasets training and test, we achieve competitive performance while real-time running and adaptability in different scenarios of virtual and real-world data.

1 Depth Estimation Networks

Before end-to-end stereo matching algorithms, CNNs are introduced to replace steps in traditional methods. Zbontar and Lecun[16] introduced a deep learning network to measure matching cost between left and right image patches. And to improve matching accuracy, typical post-processing functions including semi-global matching cost aggregation are necessary. Luo et al.[17] introduced a notable network to regard matching cost over all possible disparities to multi-label classification by faster Siamese network. Chen et al.[18] introduced an embedding network to collect multi-scale matching cost calculation features. Gidaris and Komodakis[19] proposed a three-stage model to detect and refine disparity predictions instead of hand-crafted disparity refinement. Shaked and Wolf[20] proposed a network for further refinement of disparity maps by pooling global information from cost volume to disparity confidence scores. Based on these works, tasks in traditional four-step methods are gradually replaced by convolution for more effective features aggregation and refinement.

Recent end-to-end depth estimation networks have been proposed to fuse these steps and compute the whole dense disparity map without post-processing. Aggregating context information is essential for stereo matching in no texture areas and thin structures. The encoder-decoder architecture and spatial pyramid pooling are two common methods to aggregate global and local context information. Encoder-decoder architecture can integrate multi-scale feature information via top-down and bottom-up convolutions and skip connections. The first aggregating coarse-to-fine prediction network is a fully convolutional network (FCN)[21], which remarkably improves segmentation results. Then U-net[22] was proposed to aggregate coarse-to-fine feature maps instead of coarse-to-fine predictions, and achieves excellent results in biomedical images segmentation task.

ParseNet[23] first introduces pyramid pooling based on the thesis that empirical receptive field is inadequate compared with theoretical receptive field in deep learning networks. Pyramid pooling can enlarge the empirical receptive field and extract information in the whole image level to improve network performance. To collect effectively multi-scale contextual information, PSPNet[24] presents pyramid pooling for multi-scale feature maps embedding.

Spatial pyramid pooling has been used in optical flow task. SPYNet[25] introduces a coarse-to-fine approach via image pyramid pooling to estimate optical flow. PWC-Net[26] also uses feature pyramids to improve optical flow estimation and enlarge receptive field of network. For stereo matching, PSMNet[12] introduces spatial pyramid pooling in encoder-decoder architecture and exploits global context information at the whole image level. The model consists of spatial pyramid pooling and stacked hourglass module for effective context information and cost volume regularization.

Recent works have noticed the drawbacks of 3D convolutions and methods are proposed to replace 3D convolutions without losing the ability of feature aggregation. AANet[16] introduces a special attention module via deformable convolution and intra-scale and cross scale aggregation modules for pyramid levels. More recently, HITNet[27] introduces an efficient disparity propagation stage making use of slanted windows with learned descriptors for computing initialization disparity map to high resolution matches.

In this work, we propose a lightweight stereo matching network without any 3D convolutions. We use encoder-decoder architecture and spatial pyramid pooling to integrate multi-scale feature maps and SRPs for gradually reconstructing multilevel disparity maps.

2 Multilevel Disparity Reconstruction Network

The key of current problems is how to replace 3D convolutions, which have unparalleled performance advantages in disparity computing, feature aggregation and disparity refinement. In fact, based on experience of traditional four-step method, cost volume adds disparity dimension by connecting left feature maps and shifting right feature maps, and 3D convolutions are used to learn the ability of simultaneous disparity computation and refinement. So an effective attempt is to separate disparity computation and refine two modules, to reduce the dimensions of feature information processed in the later part of network. In this way, we design MDRNet, consisting of feature extraction module,

disparity computation module, and multilevel disparity refinement modules.

The overall architecture of MDRNet is shown in Fig. 1. We use a share-weights encoder to extract descriptors for coarse-to-fine features, and also share-weights decoder to efficiently extract multi-scale details of information via skip connecting same resolution feature maps. Encoder architecture consists of convolutional residual blocks with 3×3 convolution layers to advance feature channels and stride 2 in every second layer to gradually reduce the resolution of feature maps, with leaky ReLUs as non-linearities. Decoder architecture is almost a symmetric structure with replacing down samplings with up samplings. In addition, skip connections between same feature resolution layers and spatial pyramid pooling between encoder and decoder architecture, can expand receptive fields with low computation consumption and avoid loss of detail information of multi-scale feature maps. After each of up sampling block outputs in decoder modules, we get four-level resolution feature maps extracted on the left and right images, and recorded as

$$F = \{F_1, F_2, F_3, F_4\}, \tag{2}$$

$$F_i = \{\boldsymbol{F}_i^{\mathrm{L}}, \boldsymbol{F}_i^{\mathrm{R}}, \}, \quad i \in \{1, 2, 3, 4\}, \tag{3}$$

where, F is the multi-scale feature maps set; $\boldsymbol{F}_i^{\mathrm{L}}$ and $\boldsymbol{F}_i^{\mathrm{R}}$ are feature maps; subscript i is different resolution number from high resolution to low resolution; superscripts L, R are from left and right images.

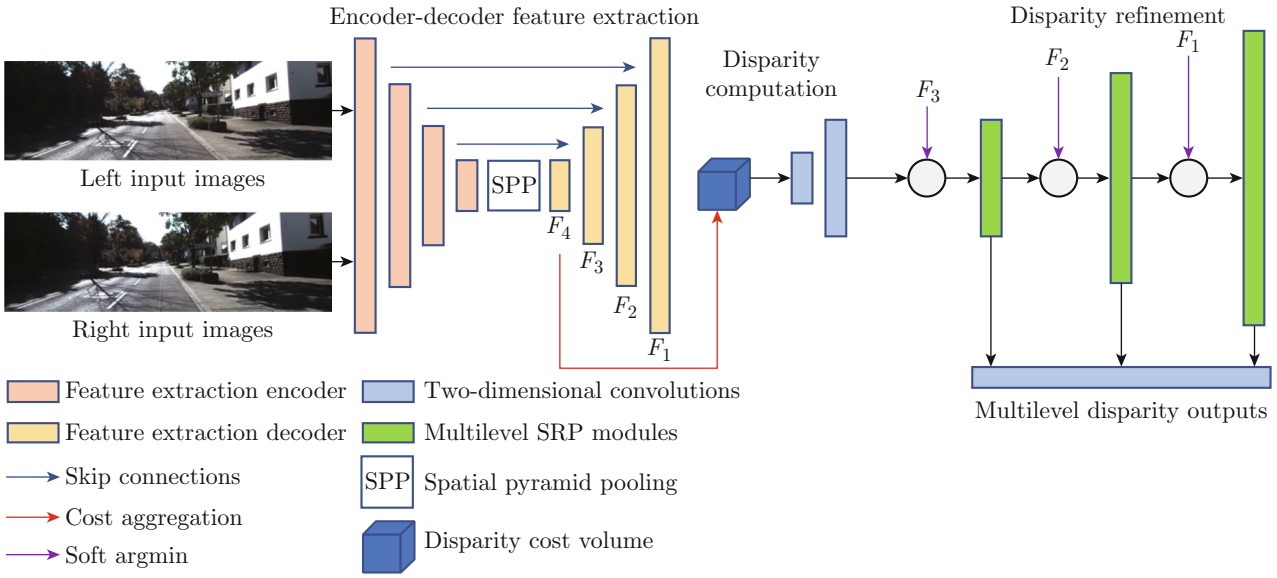


Fig. 1 Architecture overview of proposed MDRNet

In regard to disparity computation and refinement, our strategy is computing roughly disparity feature maps at the lowest resolution from $\{\boldsymbol{F}_4^{\mathrm{L}}, \boldsymbol{F}_4^{\mathrm{R}}\}$, then reconstructing the details of disparity features during the process of convolution and up sampling to the highest resolution, with high-level disparity hypothesis and F_i. Concrete methods are building an initial 4D cost volume by generating shifted $\boldsymbol{F}_i^{\mathrm{R}}$:

$$\boldsymbol{C}_i(C, d, y, x) = \\ \mathrm{concat}\left(\boldsymbol{F}_i^{\mathrm{L}}(C, y, x) - \boldsymbol{F}_i^{\mathrm{R}}(C, y, x + d), d\right), \tag{4}$$

where, C is the feature dimension; x is the horizontal dimension; y is the vertical dimension; d is the disparity dimension from 0 to max disparity D; $\boldsymbol{C}_i$ are generated for multi-level cost volume; $\mathrm{concat}(\cdot, \cdot)$ is a connection vector operation in dimension D.

At the lowest resolution, we transform 4D cost volume $\boldsymbol{C}_4\left(32, \frac{D}{64}, \frac{H}{64}, \frac{W}{64}\right)$ to 3D cost volume $\boldsymbol{C}_4'\left(32 \times \frac{D}{64}, \frac{H}{64}, \frac{W}{64}\right)$ and connect $\{\boldsymbol{F}_4^{\mathrm{L}}, \boldsymbol{F}_4^{\mathrm{R}}\}$ then compute initialization disparity feature maps by a 2D convolution layers with 3×3 kernel and another with 1×1 kernel changing channels to $\left(16, \frac{H}{64}, \frac{W}{64}\right)$, where H and W are the height and width of original image resolution.

For other cost volume levels, disparity feature maps D_i are generated by

$$\boldsymbol{D}_i = \underset{d \in [0, D]}{\mathrm{argmin}}\, \boldsymbol{C}_i. \tag{5}$$

By now, we have computed disparity feature maps $\boldsymbol{D}_i = \left(16, \frac{H}{2^{i+2}}, \frac{W}{2^{i+2}}\right)$, $i \in \{1, 2, 3, 4\}$. For the disparity refinement, we design SRP shown in Fig. 2, which consists of multi residual modules and spatial pyramid pooling followed. The input of SRP is $\left(64, \frac{H}{2^{i+2}}, \frac{W}{2^{i+2}}\right)$ connection of upsampling $\boldsymbol{D}_{i-1}\boldsymbol{D}_i$,

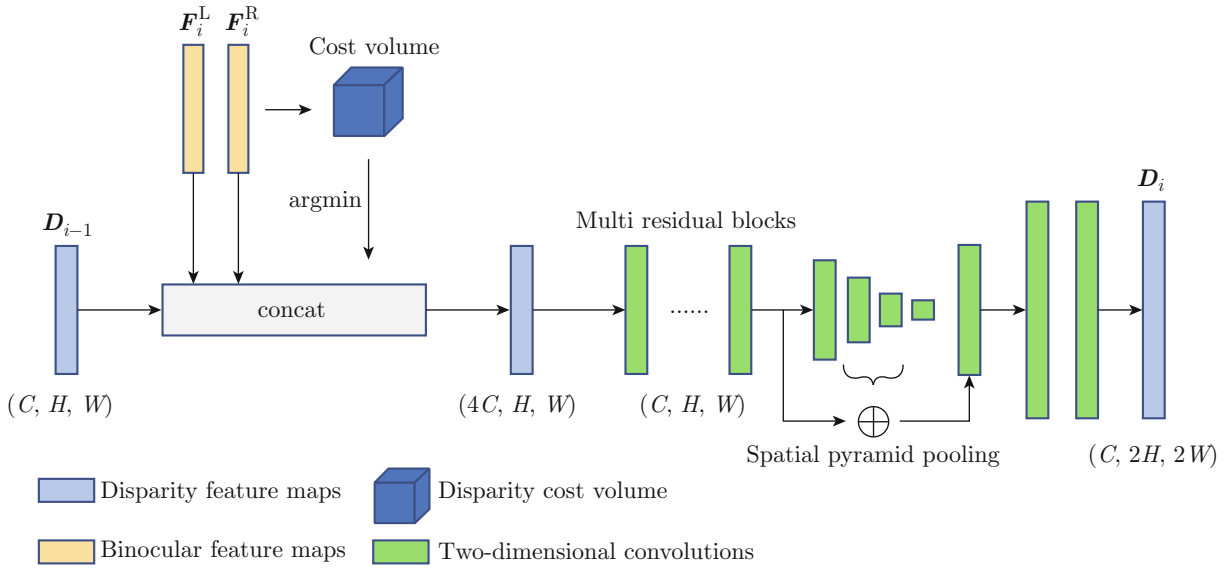


Fig. 2 SRP module

$\boldsymbol{F}_i^{\mathrm{L}}$ and $\boldsymbol{F}_i^{\mathrm{R}}$ to fuse disparity calculation at lower level and detail information at high level. The output of SRP is $\left(16, \frac{H}{2^{i+1}}, \frac{W}{2^{i+1}}\right)$ through up sampling and 1×1 convolution changing channels for next SRP or final disparity map regression.

In order to achieve faster training and better performance, we design multi-level loss functions. Two convolution layers with 3×3 and 1×1 kernel after each output of SRPs get several disparity map predictions under different scales. For every scale, we subsample the ground truth to the same resolution and compute smooth L1 loss:

$$L_{\mathrm{s}} = \frac{1}{N}\sum_{i=0}^{N}\mathrm{smooth}(||\hat{d} - d_{\mathrm{gt}}||), \tag{6}$$

$$\mathrm{smooth}(x) = \begin{cases} x - 0.5, & |x| > 1 \\ \frac{1}{2}x^2, & |x| \leqslant 1 \end{cases}, \tag{7}$$

where, N is the number of pixels; $\hat{d}$ is the disparity estimation result; d_{gt} is the disparity ground truth.

In addition, to make up for the lack of local correlation in disparity dimension, we add disparity gradient loss:

$$L_{\mathrm{g}} = \frac{1}{N}\sum_{i=0}^{N}(||\hat{g_x} - g_{\mathrm{gt}_x}|| + ||\hat{g_y} - g_{\mathrm{gt}_y}||), \tag{8}$$

where, $\hat{g_x}$ and $\hat{g_y}$ are x, y gradients of disparity estimation results; g_{gt_x} and g_{gt_y} are x, y gradients of the same level ground truth.

The whole loss function for training is with the weights of hyper parameters α and β:

$$f_{\mathrm{L}} = \sum_{i=0}^{S}\alpha(L_{\mathrm{s}} + \beta L_{\mathrm{g}}), \tag{9}$$

where S is the number of SRP modules.

3 Experiments

In training experiments, according to the characteristics of different scenes, we combine cross-data sets based on publicly available datasets for good adaptability in virtual and real-world environment. General scene cross-data sets consist of Middlebury[10], ETH3D[28] and Sceneflow[29]. Middlebury contains 33 training stereo pairs. ETH3D contains 27 low-resolution stereo pairs in real-world scene with sparsely labels. Sceneflow contains around 3×10^4 stereo pairs generated in virtual 3D scenes with high quality disparity labels. Driving scene cross-data sets consist of KITTI[2], HR[30] and Driving Stereo[31] demo images. KITTI-12 contains 194 stereo pairs and KITTI-15 contains 200 stereo pairs with 4×10^4 resolutions in real-world driving scene. HR contains 780 training pairs generated in virtual driving scenes with high resolution. Driving Stereo demo contains 300 stereo pairs in real-world driving with various weather and light environment.

We train the MDRNet based on Pytorch and Adam optimizer with a learning rate of 0.001 and betas parameters of (0.9, 0.999). Setting the batch size to 32 on a machine with 4 GeForce RTX 2080 Ti GPUs, we pre-train MDRNet in general scene cross-data sets with 10 epochs and finetune in driving scene cross-data sets with another 10 epochs for the last two epochs shrinking learning rate.

In cross-data sets' pretreatment, we increase proportion of images from small datasets in the train images list by copying themselves and randomly changing the image brightness to achieve the effect of data augmentation. And for high resolution datasets, we use down-sampling properly to reduce their resolution which assists the max-disparity fixed in training MDRNet to search procedure even faster. Other typical data augmentation methods including random asymmetric adjustments for perturbing stereo pairs contrast and replacing little random areas in right image with patches are used to enhance the robustness of the network.

As listed in Table 1, performances of MDRNet with different SRP settings in training based on general scene cross-data sets indicate that the spatial pyramid pooling assists the network in working better. And within a certain range of numbers of residual in SRP it can improve network performance. Considering that when the number of residual blocks sets to 8, the performance is not significantly improved but the consumption of calculation is increased, so 6 residual blocks are the right network structure.

Table 1 MDRNet evaluation with different SRP settings

Residual numbers in single SRP	Spatial pyramid pooling	End point error in general scene cross-data sets/pixel
3	–	2.63
3	√	2.34
6	–	1.82
6	√	1.71
8	√	1.71

As shown in Table 2, experiments with various combinations of loss weights for multilevel disparity estimation maps (α_3 for lower resolution SRP output and α_1 for higher resolution SRP output) indicate that the best set of multilevel loss weight is $\alpha_3 = 0.4$, $\alpha_2 = 0.6$, $\alpha_1 = 1.0$ and $\beta = 0.5$, which achieves 3.89% 3-pixel-error in driving scene cross-data sets after fine tuning.

Table 2 Different weight values for multilevel f_{L} on validation errors

α_3	α_2	α_1	β	3-pixel-error in driving scene cross-data sets/%
0	0	1.0	0	5.65
0.2	0.3	1.0	0.3	4.57
0.4	0.6	1.0	0.5	3.89
0.6	0.8	1.0	0.7	4.16
0.7	0.9	1.0	1.0	4.02

Experiments shown in Table 3 compare MDRNet with four representative stereo networks in parameters and memory consumption with GPUs. Our method shows much fewer parameters and less computational cost, which is very helpful to deploy the network in practical application. This indicates that replacing 3D disparity computation and optimization with SRPs to reconstruct multilevel disparity maps is a valuable way.

Table 3 Comparison of parameters and memory consumption with four stereo networks

Method	Parameters	Memory/GB
StereoNet[32]	6.20×10^5	1.41
GC-Net[11]	2.85×10^6	11.52
PSMNet[12]	5.22×10^6	4.08
GA-Net[13]	4.60×10^6	6.23
MDRNet (ours)	3.60×10^5	0.63

And multilevel disparity maps from the lowest resolution to the highest resolution computed by SRP outputs are shown in Fig. 3. Details of disparity map are restored in the process of gradual reconstruction.

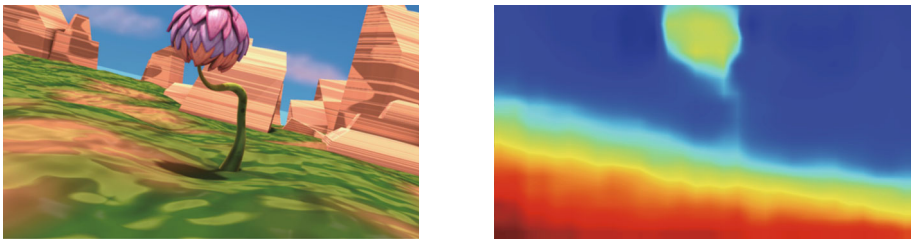
In the whole test, we randomly select 500 stereo pairs from Sceneflow datasets and 100 stereo pairs from the whole Middlebury and ETH3D copied in image lists with randomly changing the image brightness and contrast to aggregate cross-data test sets in general scene. Similarly we aggregate cross-data test sets in driving scene from extracting randomly 50 stereo pairs, each of KITTI-15, KITTI-12, HR and Driving Stereo demo data with a little artificially added data interference.

Generating cross-data sets and testing performance of MDRNet compared with other representative stereo networks are a random test used to measure the generalization ability for different environment. We replicate these representative open source works and compare MDRNet with them on the same generated test sets. Such experiments have been done 20 times and the final test errors are the average of all experimental results, which is shown in Table 4 for end point error in general scene cross-data test sets and in Table 5 for 3-pixel-error in driving scene cross-data test sets.

Experiments show that our method MDRNet, achieves competitive performance and faster inference compared with other representative stereo networks. We notice that although our method is not as good

Table 4 Performance compared with representative stereo networks in general datasets

Method	Time/s	End point error in general scene cross-data test sets/pixel
StereoNet[32]	0.015	1.92
DispNetC[10]	0.06	2.22
GC-Net[11]	0.88	3.07
PSMNet[12]	0.4	1.65
GA-Net[13]	0.84	1.43
MDRNet (ours)	0.03	1.82



(a) Original left image from sceneflow (b) Output of low level SRP module

(c) Output of middle level SRP module (d) Output of high level SRP module

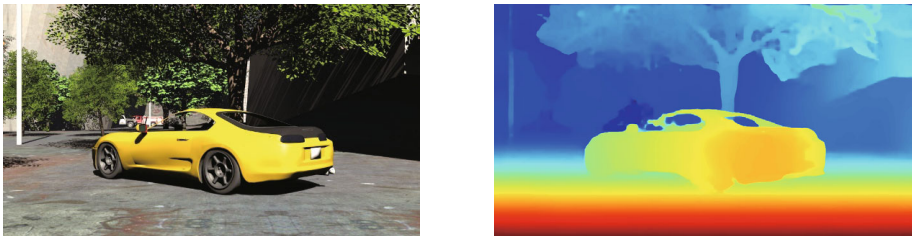
Fig. 3 Multilevel disparity maps

Table 5 Performance compared with representative stereo networks in driving datasets

Method	Time/s	3-pixel-error in driving scene cross-data test sets/%
StereoNet[32]	0.015	5.94
DispNetC[10]	0.06	5.44
GC-Net[11]	0.9	3.93
PSMNet[12]	0.42	3.6
GA-Net-deep[13]	1.6	2.92
MADNet[33]	0.02	5.74
HD^3[34]	0.14	3.17
MDRNet (ours)	0.032	4.34

as some of the state-of-the-art methods based on 3D convolutions by now, MDRNet performs better than other real-time 3D convolution stereo matching networks, with the least parameters and the lowest calculation consumption. And replacing 3D convolutions with light-weight 2D convolution modules is practicable, which is highly significant for real-time depth estimation applications with low computing resources.

For visualization experiments, we select some binocular image pairs with thin structure objects and large textureless areas as test images. Experimental visualization results are shown in Fig. 4. In Fig. 4(a), the outline of the sports car and the shape of the shell are almost clear, which shows that the algorithm has a strong adaptability in the thin structure objects. In Fig. 4(b), our work can estimate complete disparity information in large textureless areas such as the back of the chair and bench without any small connected areas



(a) Original virtual driving image and disparity pseudo-color map

(b) Original flying things image and disparity pseudo-color map

Fig. 4 Illustration visualization results on general datasets

and incomplete edge, which indicates that the algorithm has good estimation performance for the textureless areas in the image. This shows that SRP modules have excellent fusion and disparity calculation abilities for the feature information of different scales.

Figure 5 shows the performance of MDRNet with the large-resolution inputs from Middlebury dataset. For most of pixels in left image, algorithm can get correct and clear disparity estimation results. It should be noted that in the nearest artificial flower of the scene, due to the high resolution of images, disparity estimation results of this part exceed the maximum disparity value set during the network training, so disparity estimation errors occur. How to reduce the influence of fixed disparity search range on algorithm generalization is one of the problems we need to solve in the future work.

Fig. 5 Illustration visualization results on Middlebury datasets

4 Conclusion

In this paper, we propose a lightweight stereo matching network consisting of encoder-decoder feature extraction layers and SRPs without any 3D convolutions. To replace commonly disparity computation and refinement by 3D convolutions and full 4D cost volume, we design the disparity initialization modules and multi-level disparity refinement models. Extensive experiments based on cross-datasets validate its competitive and real-time performance, and good adaptability in virtual and real-world data. In future work, on the premise of not significantly increasing the consumption of network inference, we intend to add cross-scale between high resolution and low resolution to improve algorithm performance. Another significant direction is trying to deploy the algorithm to embedded devices for actual scenarios application. Efficient and practical depth estimation algorithms will be the important direction in stereo matching networks' developing based on CNNs.

References

[1] SCHARSTEIN D, SZELISKI R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms [J]. *International Journal of Computer Vision*, 2002, **47**(1/2/3): 7-42.

[2] MENZE M, GEIGER A. Object scene flow for autonomous vehicles [C]//*2015 IEEE Conference on Computer Vision and Pattern Recognition*. Boston: IEEE 2015: 3061-3070.

[3] SCHMID K, TOMIC T, RUESS F, et al. Stereo vision based indoor/outdoor navigation for flying robots [C]//*2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Tokyo: IEEE, 2013: 3955-3962.

[4] ZHANG L, SEITZ S M. Estimating optimal parameters for MRF stereo from a single image pair [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, **29**(2): 331-342.

[5] SUN J, ZHENG N N, SHUM H Y. Stereo matching using belief propagation [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003, **25**(7): 787-800.

[6] KOLMOGOROV V, ZABIH R. Computing visual correspondence with occlusions using graph cuts [C]//*Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*. Vancouver: IEEE, 2001, **2**: 508-515.

[7] YOON K J, KWEON I S. Adaptive support-weight approach for correspondence search [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006, **28**(4): 650-656.

[8] HOSNI A, RHEMANN C, BLEYER M, et al. Fast cost-volume filtering for visual correspondence and beyond [J].*IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013,**35**(2): 504-511.

[9] MIN D, LU J, DO M N. A revisit to cost aggregation in stereo matching: How far can we reduce its computational redundancy? [C]//*2011 International Conference on Computer Vision*. Barcelona: IEEE, 2011: 1567-1574.

[10] MAYER N, ILG E, HÄUSSER P, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation [C]//*2016 IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas: IEEE, 2016: 4040-4048.

[11] KENDALL A, MARTIROSYAN H, DASGUPTA S, et al. End-to-end learning of geometry and context for deep stereo regression [C]//*2017 IEEE International Conference on Computer Vision*. Venice: IEEE, 2017: 66-75.

[12] CHANG J R, CHEN Y S. Pyramid stereo matching network [C]//*2018 IEEE/CVF Conference on*

Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 5410-5418.

[13] ZHANG F, PRISACARIU V, YANG R, et al. Ga-net: Guided aggregation net for end-to-end stereo matching [C]//*2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019: 185-194.

[14] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite [C]//*2012 IEEE Conference on Computer Vision and Pattern Recognition*. Providence: IEEE, 2012: 3354-3361.

[15] XU H, ZHANG J. AANet: Adaptive aggregation network for efficient stereo matching [C]//*2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle: IEEE, 2020: 1956-1965.

[16] ŽBONTAR J, LECUN Y. Computing the stereo matching cost with a convolutional neural network [C]//*2015 Conference on Computer Vision and Pattern Recognition*. Boston: IEEE, 2015: 1592-1599.

[17] LUO W, SCHWING A G, URTASUN R. Efficient deep learning for stereo matching [C]//*2016 Conference on Computer Vision and Pattern Recognition*. Las Vegas: IEEE, 2016: 5695-5703.

[18] CHEN Z, SUN X, WANG L, et al. A deep visual correspondence embedding model for stereo matching costs [C]//*2015 IEEE International Conference on Computer Vision*. Santiago: IEEE, 2015: 972-980.

[19] GIDARIS S, KOMODAKIS N. Detect, replace, refine: Deep structured prediction for pixel wise labeling [C]//*2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu: IEEE, 2017: 7187-7196.

[20] SHAKED A, WOLF L. Improved stereo matching with constant highway networks and reflective confidence learning [C]//*2017 Conference on Computer Vision and Pattern Recognition*. Honolulu: IEEE, 2017: 6901-6910.

[21] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation [C]//*2015 IEEE Conference on Computer Vision and Pattern Recognition*. Boston: IEEE, 2015: 3431-3440.

[22] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation [M]//Medical image computing and computer-assisted intervention - MICCAI 2015. Cham: Springer, 2015: 234-241.

[23] LIU W, RABINOVICH A, BERG A C. Parsenet: Looking wider to see better [EB/OL]. (2015-06-15). https://arxiv.org/abs/1506.04579.

[24] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network [C]//*2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu: IEEE, 2017: 6230-6239.

[25] RANJAN A, BLACK M J. Optical flow estimation using a spatial pyramid network [C]//*2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu: IEEE, 2017: 2720-2729.

[26] SUN D, YANG X, LIU M Y, et al. PWC-net: CNNs for optical flow using pyramid, warping, and cost volume [C]//*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City: IEEE, 2018: 8934-8943.

[27] TANKOVICH V, HÄNE C, ZHANG Y, et al. HITNet: Hierarchical iterative tile refinement network for real-time stereo matching [C]//*2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville: IEEE, 2021: 14357-14367.

[28] SCHARSTEIN D, HIRSCHMÜLLER H, KITAJIMA Y, et al. High-resolution stereo datasets with subpixel-accurate ground truth [M]//Pattern recognition. Cham: Springer, 2014: 31-42.

[29] SCHÖPS T, SCHÖNBERGER J L, GALLIANI S, et al. A multi-view stereo benchmark with high-resolution images and multi-camera videos [C]//*2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu: IEEE, 2017: 2538-2547.

[30] YANG G, MANELA J, HAPPOLD M, et al. Hierarchical deep stereo matching on high-resolution images [C]//*2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019: 5510-5519.

[31] YANG G, SONG X, HUANG C, et al. DrivingStereo: A large-scale dataset for stereo matching in autonomous driving scenarios [C]//*2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019: 899-908.

[32] KHAMIS S, FANELLO S, RHEMANN C, et al. StereoNet: Guided hierarchical refinement for real-time edge-aware depth prediction [M]//Computer vision - ECCV 2018. Cham: Springer, 2018: 596-613.

[33] TONIONI A, TOSI F, POGGI M, et al. Real-time self-adaptive deep stereo [C]//*2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019: 195-204.

[34] YIN Z, DARRELL T, YU F. Hierarchical discrete distribution decomposition for match density estimation [C]// *2019 IEEE Conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019: 6037-6046.