Image and Vision Computing 147 (2024) 105074

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Hourglass cascaded recurrent stereo matching network

Tuming Yuan [a,*], Jiancheng Hu [a], ShuangJiang Ou [b], Weijia Yang [a], Yafang Hei [a]

[a] College of Applied Mathematics, Chengdu University of Information Technolog, Chengdu 610225, PR China
[b] School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, PR China

ARTICLE INFO

Keywords:
Stereo matching
Stacked hourglass network
Recurrent neural network
Attention mechanism

ABSTRACT

Stereo matching acts a crucial role in computer vision and robotics applications. An accurate cost volume and robust disparity regression method are essential for stereo matching of high accuracy. Following GCNet and PSMNet, constructing 4D cost volume and then using the soft argmin method to regress has been dominated. However, it will encounter many difficulties due to the multi-modal distribution of cost volume. One of the reasons for this multi-modal distribution is the occlusion area which not be possible to find a matching region on the reference image and rarely discussed. In this paper, we propose to use global context information could improve the performance of model in occluded regions. Recently, novel recurrent neural network regression methods are proposed, but most of them regress disparity maps from 3D cost volume. In this paper, we propose the new combinatorial paradigm that combine stacked hourglass modules and recurrent neural networks to further aggregate 4D cost volume and regress disparity respectively. The proposed method can be seamlessly integrated into most stereo matching networks, we improved the accuracy by 45% for PSMNet and 38% for GwcNet in our experiment. Experimental results on Scene Flow, KITTI2012, KITTI2015, and ETH3D datasets show our method is competitive. The code is available at: https://github.com/truman1211/HCRnet.

1. Introduction

Depth estimation is a key technique for many applications such as autonomous driving, 3D model reconstruction, and robotics [1,2]. Binocular stereo matching is a common means of depth estimation. Gave a pair of rectified stereo images (one left frame and one right frame taken by a binocular camera), the purpose is to compute the movement between two corresponding pixels (along the horizontal line) also known as disparity. The depth can be obtained by the principle of similar triangles while the disparity is calculated. Recently, convolutional neural networks (CNN) [3–5] have displayed a considerable achievement in this domain. Most of the CNN stereo models are inspired by the traditional methods which consist of four steps [6], i.e. feature extracting, matching cost, cost aggregation and disparity regression.

In the field of optical flow, RAFT [7] has been a reference to the latest method, and abundant works of optical flow [8–12] are adopting the RAFT GRU module and are making great progress. Stereo is closely related to optical flow, despite the similarities between stereo and flow, RAFT-stereo [13] has not received sufficient attention in stereo like RAFT in optical flow, and the GRU module is not widely used in stereo. Although RAFT-stereo and CREStereo introduced the GRU module in stereo, they also abandoned some advantages of the mainstream stereo baseline, such as the regularization of cost volume. In this work, we aim to explore a more appropriate way to use the GRU module that could leverage the GRU module and the mainstream baseline. We build our model based on two observations, firstly, well-known baselines such as PSMNet and GwcNet use the hourglass module to regularize the cost volume, its encoder could construct multiscale cost volume while the too small cost volume will lose a lot of detail information which means it is more suitable for handling large objects, because some properties of large objects are preserved better at low resolutions such as edge information. Secondly, the GRU regression method could cooperate Context information of Multi-level Receptive fields (CMR), large objects require more receptive fields, suggesting that different scale cost volumes could cooperate with GRU that integrated different context information. We believe that giving the GRU different context information at different stages can force it to process different information such as objects of different scales.

For the occlusion region, correlations between appearance features are not meaningful for cost volume. When local features are not suitable for occluded point, natural thinking is that non-local approach might be helpful. Traditional method [8] conducts a left–right consistency check

* Corresponding author.
E-mail address: truman1211@163.com (T. Yuan).

https://doi.org/10.1016/j.imavis.2024.105074
Received 26 December 2022; Received in revised form 21 April 2024; Accepted 11 May 2024
Available online 15 May 2024

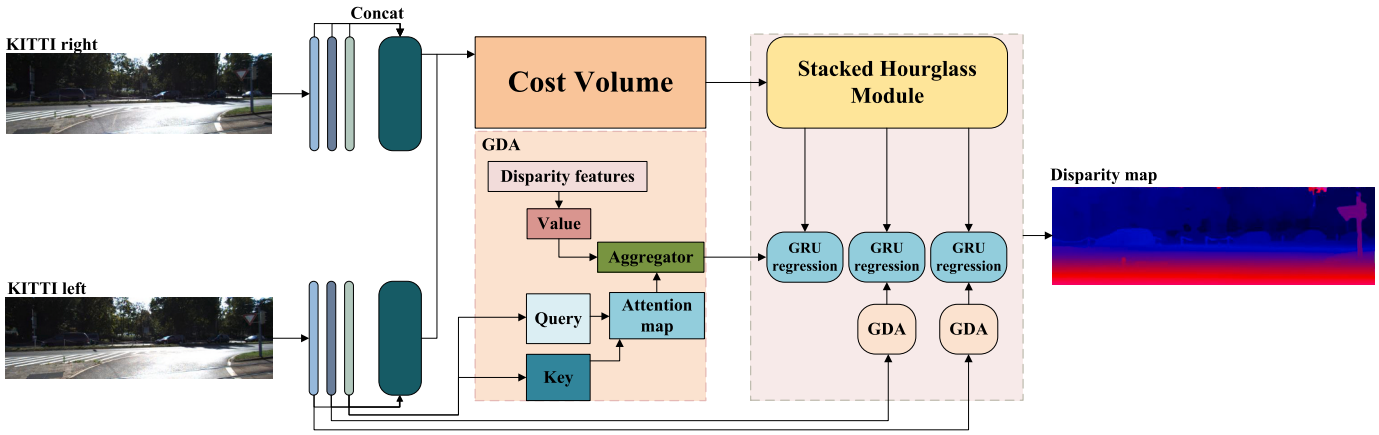

Fig. 1. Architecture overview. Our proposed network consists of four key parts: cost volume construction, stacked hourglass modules (More details are shown in Fig. 2), and multi-level GRU regression modules which embedded different context information, and global disparity aggregation (GDA) module.

to estimate the occluded regions, and then uses interpolation methods to fill in the disparity of the occluded regions as a post-processing step. On the contrary, we do not explicitly estimate an occlusion map, but we take a non-local approach to provide additional information to correctly predict disparity in occluded regions. Specifically, we introduce a Global Disparity Aggregation (GDA) module that we first compute an attention map based on the self-similarities of the right image, then use the attention map to aggregate disparity. We use the global information to augment our model to estimate disparity of occluded regions.

In this paper, we propose Hourglass Cascaded Recurrent stereo matching network (HCRNet). Our main contributions are as follows. We show that Context information of Multi-level Receptive fields is highly beneficial for stereo, furthermore, the GRU-based regression method could improve the well-known baseline such as PSMNet and GwcNet directly. We show that global disparity aggregation leads to improvement of accuracy in occluded regions, without damaging the performance in non-occluded regions. HCRNet achieves state-of-the-art performance on the Scene Flow, KITTI2012, and KITTI2015, ETH3D datasets, outperforming many previous works based on the hourglass module.

2. Related work

Recently, considerable achievements have been made by learning-based models. Following DispNet [5] which was one of the first papers to propose an end-to-end model to regress disparity, researchers devoted to effective construction and regularization of cost volume, not only that, many novel regression methods have been proposed.

Cost volume construction and aggregation are two tightly-coupled modules where aggregation is pivotal to regularize cost volume. Existing cost volume construction methods can be roughly categorized in to two types: 3D cost volume and 4D cost volume. DispNetC [14] directly measures the similarities of the left and right image features to form a 3D cost volume. Then, 2D convolutions are applied to aggregate cost volume. Such 3D cost volume demands low memory and less computation, while the encoded information is limited because of large information loss in the channel dimension. GCNet [4] concatenates each unary feature with its corresponding unary from the opposite stereo image across each disparity level to form a 4D cost volume [3,15]. Such 4D cost volume preserves abundant information about features while requires extensive 3D convolution to aggregate cost volume and learns similarity measurements from scratch. To trade off the efficiency and effectiveness, GwcNet [15] proposed group-wise correlation that left features and right features are divided into groups along the channel dimension and compute similarities among each group to obtain multiple matching cost, which are then packed to form a 4D cost volume.

Even though 4D cost volume needs more computation and parameters to aggregate cost volume the final 3D cost volume is obtained, and its effectiveness was demonstrated by extensive related work [3,15–18]. Inspired by encoder-decoder struct which are designed for dense prediction task to get around their computational burden by encoding subsampled feature maps, GCNet [4] extend it to three dimensions and apply 3D transposed convolution with stride two and a single feature output to obtain regularized cost volume. In order to learn more context information, following human pose estimation work [19], PSMNet [3] introduced stacked encoder-decoder (termed hourglass in [19]) in stereo and achieve improved performance. In recent years, stacked hourglass module has been dominated module to aggregate cost volume, while cost volume can never be perfect, even when using extensive 3D convolutions to regularize [13,17,20,21]. For example, if areas are occluded in the opposite stereo image, the cost curve will be unreliable. Unreliable cost volume could result in imprecise depth estimation when the soft argmin method was used. Most related to our work are RAFT-stereo [14] and CREStereo [22], RAFT-stereo introduced an iterative refinement method in the optical flow network RAFT [23] which iterative refinement disparity map from cost volume, following RAFT-stereo, CREStereo proposed an Adaptive Group Correlation Volume to reduce matching ambiguity results from imperfect calibration for real-world stereo cameras. Different from the general stereo network, both of them lack a process to aggregate cost volume. In this work, following RAFT-stereo, we use the GRU module to refine the disparity map iteratively and aggregate cost volume through a widely used stacked hourglass module. Compared with the soft argmin method which is a weighted average method, GRU module increase computation due to the convolution operations which are more complex, while GRU based method could make a better prediction according to the results of the previous step and consider the context information and cost volume at the same time. To ease the computational burden, we refine disparity across multi-scale cost volume, because small cost volume reduced fine-grained details, we embed context information of high-level receptive fields to corresponding GRU module to encourage model to handle large object at early stage.

To the best of our knowledge, occlusion is often discussed in traditional method. From the existing deep learning theory, the traditional approach seems to be more compatible as an unsupervised learning method. Left–right consistency checking is often used to reject outlier pixels in the occluded region, which is widely used in traditional methods [8,24]. The same idea is used in unsupervised stereo matching algorithms [25,26], where occluded regions are filtered through mutual supervision of the geometric parallax relationship between the left and right images, and then disparity filling is performed. However, filtering out occluded regions and rejecting them is inherently a very difficult

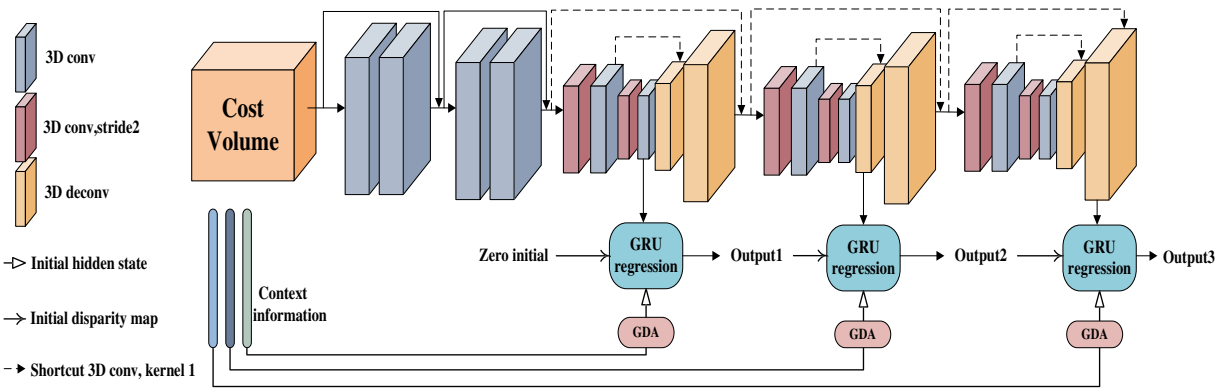

Fig. 2. The structure of our proposed stacked hourglass cascaded regression module. It consists of three hourglass modules and three GRU units (More details are shown in Fig. 3), three GDA modules (as shown in Fig. 1).

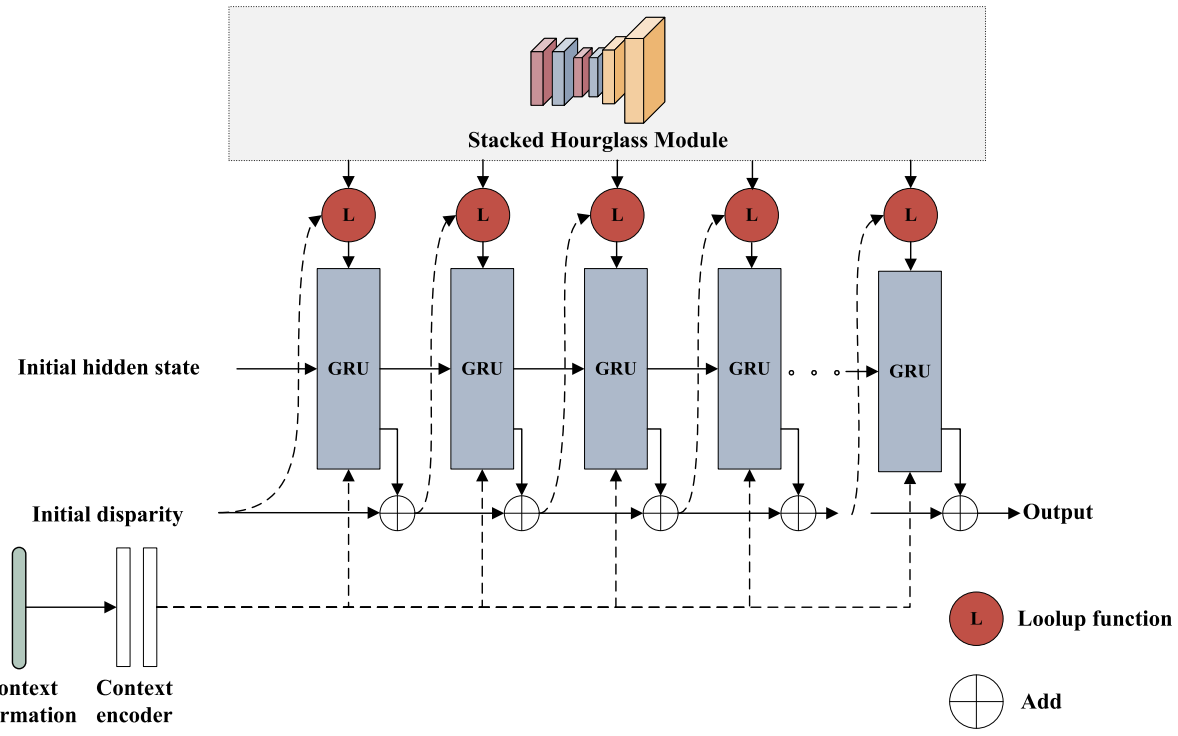

Fig. 3. The structure of GRU regression module is described in Sect.3.3. In every iteration, GRU use the current disparity estimation to sample from the cost volume, initial disparities are set as estimation from last output module (First initial set as 0).

task, while filling the disparity itself seems in principle inexplicable, since it does not yield valid information from the other image. Contrary to the above approaches, we do not explicitly estimate occlusion region neither handle occlusion in a post-processing way. Instead, we take an implicit approach to globally aggregate disparity, which provides extra information to predict disparity at occluded regions. This idea is similar to the belief propagation algorithm in classical stereo matching [9].

Attention mechanism had been achieved great achievement in recent years since transformer architecture [27] was introduced to computer vision. Among various modules in the transformer architecture, self-attention is the key design feature that model long-range dependencies. Recently, stereo related works [28,29] found global contextual information is critical for depth estimation, while using global contextual information to matching features of benign region is unnecessary. In this work, we do not use self-attention that query, key and value vectors coming from the same features, inspired by [30], query and key vectors come from the context features of left image while value vectors come from the correlations, which is an encoding of the cost volume.

3. Method

In this section, we describe our HCRNet which based on GwcNet. Firstly, we describe our model architecture in sect.3.1. More implementation detail is described in the following subsections. For the convenience of description, we assume that the input resolution of the stereo pairs is, and the max disparity range is.

3.1. Model architecture

The brief structure of HCRNet is illustrated in Fig. 1. The network consists of four main parts, cost volume construction, stacked hourglass networks, GRU regression module, and GDA module. The left and right image are conveyed to two weight-sharing feature extraction pipelines

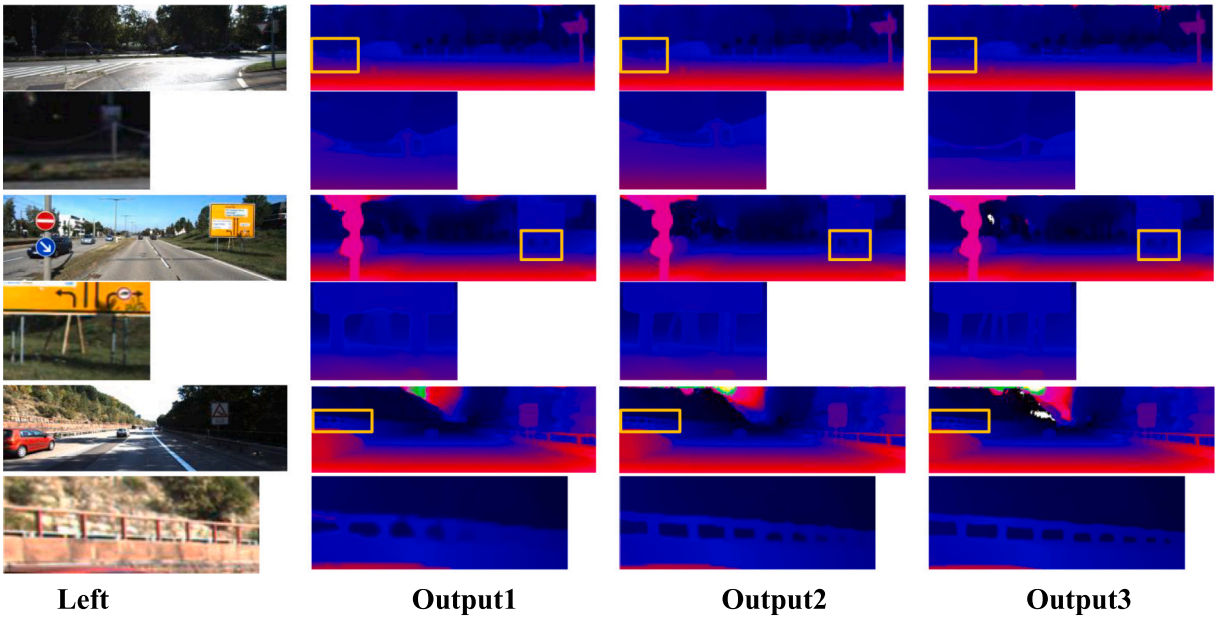

Fig. 4. From left to right: Left image and results of each output of HCRNet for KITTI 2015 test images.

like GwcNet. The left and right image features are then used to form a 4D cost volume (cost volume used in baseline). Hourglass modules export final 3D cost volume to the GRU regression module. Then, GRU regression modules associate context information of multi-level receptive fields to regress the disparity map, specifically, three GRU module initialize hidden state with different context information in turn. GDA act as a plug-in module that provides global information to GRU module. Following the ideal of CREstereo [22], low resolution and high-level feature maps are more beneficial to ill-posed regions such as textureless areas, while high resolution feature maps are more beneficial to preserve the details. GRUs handle multi-scale cost volume in three hourglass stages while share the same weights. Except for GRU of the first hourglass stage, which initialized to all zeroes, the subsequent stage is initialized by the up-sampled version of prediction from the previous stage. We freeze 2D and 3D batch normalization during the train and inference process.

3.2. Feature extraction and hourglass module

We defined feature encoder and the context encoder. The feature encoder is the same as GwcNet (adopt the ResNet-like network), except modified the output channel of last 3 residual blocks to 128 channels. The context encoder shares weights with feature encoder. The context features are used to initialize the hidden state of GRU regression units.

Stacked hourglass module has been dominated in stereo networks, it consists of multiple 3D convolution layers and 3D transposed convolution layers that form the encoder-decoder structure. Numerous related works had demonstrated its ability of aggregating cost volume. In our method, the stacked hourglass module consists of three hourglass networks, each of which aggregates cost volume and then generates different sizes of cost volume for GRU. More details are described in literature [3].

3.3. GRU regression units

Following RAFT-Stereo, after given final 3D cost volume C, and initial disparity map. Lookup function defined as:

$$S_t(x,y) = L(C, D_0) = C(x, y, D_0), \tag{1}$$

where S_t is part of the final cost volume C. To enlarge receptive field, we sample disparity features S^c from C,

$$S_t^c(x,y) = L(C, D_0 + \Delta r), \Delta r = 0, \pm 1, \pm 2, \pm 3, \pm 4, \tag{2}$$

where $S_t^c \in R^{H\times W\times 9}$, H and W are the height and width of the feature map. Then S_t^c will be took into GRU units as part of X_t.

GRU units defined as follows:

$$\begin{cases} z_t = \sigma((Conv_{3*3}([h_{t-1}, X_t])) \\ r_t = \sigma((Conv_{3*3}([h_{t-1}, X_t])) \\ \widetilde{h}_t = tanh(Conv_{3*3}([r \odot h_{t-1}, X_t])) \\ h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h}_t \end{cases} \tag{3}$$

where indicates tensor concatenation, σ and $\odot$ denotes sigmoid function and Hadamard product, initial hidden state t_0 is context information. Finally, updated hidden state h_t is used to compute disparity residual ΔD_t, specifically, by two layers convolutional neural network. After the update of h_t, $D_{t+1} = D_t + \Delta D_t$.

3.4. Context information of multi-level receptive fields

As described in Fig. 2, our three GRU regression modules handle different size cost volume in a coarse-to-fine manner. Inspired by Triden Networks [31], the small objects need smaller receptive field and the large objects benefit from the increasing receptive fields. We believe that the lower resolution cost volume is more appropriate to handle larger objects because object details are easily lost at lower resolutions. With this intuition in mind, with respect to the lower resolution cost volume, we give it context information of larger receptive field i.e. deeper layer image features, as for higher resolution cost volume, we give it context information of smaller receptive field. Experiments verify the effectiveness of this design as shown in Fig. 4, HCRNet could estimate disparity in a coarse-to-fine manner. One more reason for this module is that, considering the computation cost, when using a more complex regression method, regressing disparity on smaller cost volume is a good way to make a trade-off. In our ablation study, we found that there is no loss of accuracy, but also the inference speed is improved.

3.5. Global disparity aggregation

Let $x \in R^{(H*W)\times D_c}$ denote the flattened context information and $s \in S_t^c$

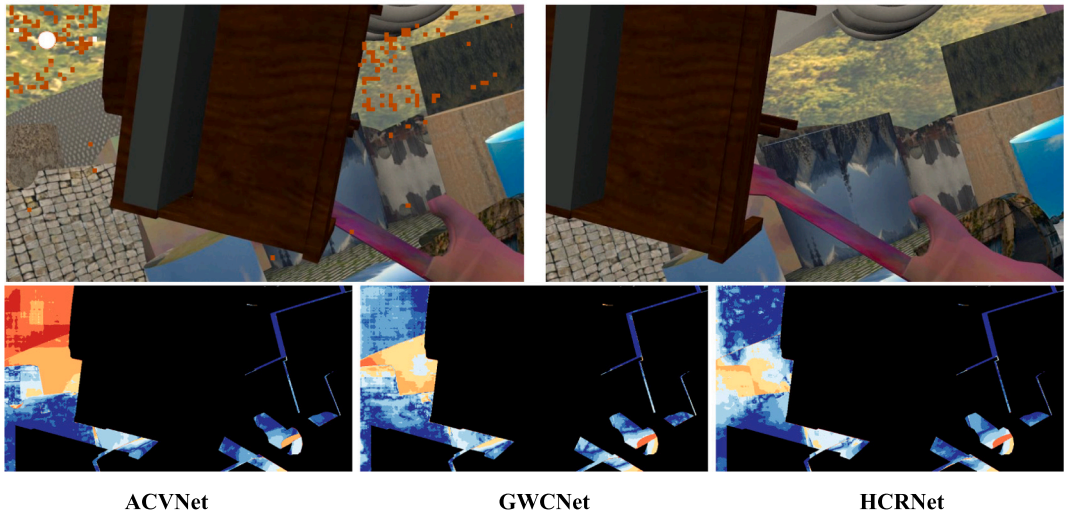

Fig. 5. Attention map visualization. We show the left image and right image on the top. For the query point (white circle) in left image, we show the attention map corresponding to this query (yellow point mean high attention weights). We also show a visual comparison with other methods on the occluded regions on the below. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

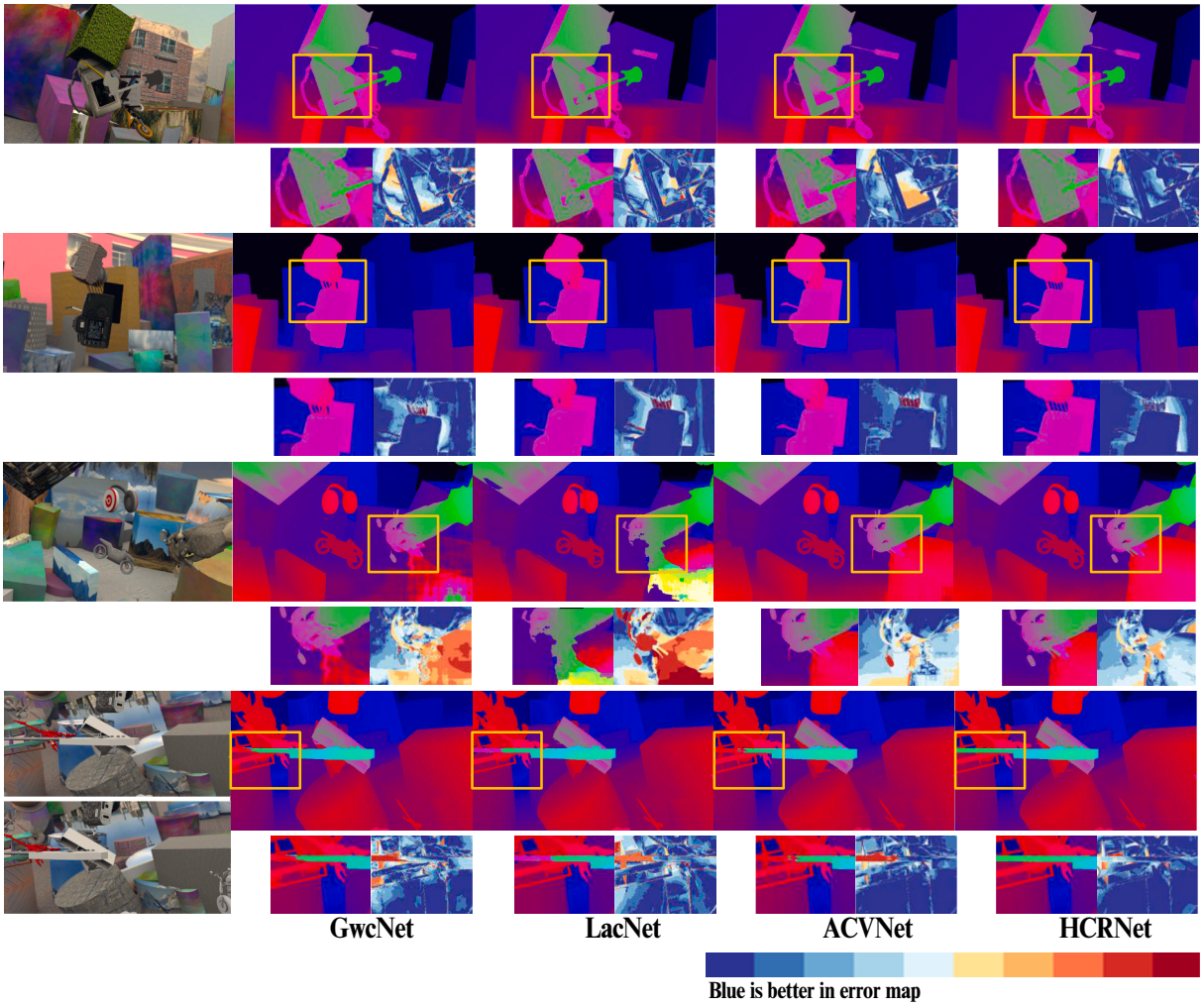

Fig. 6. Results of disparity estimation for Scene Flow test images. From left: left stereo input image, predictions of GwcNet, LacNet, ACVNet and our HCRNet prediction. We also observed in last row results, HCRNet was benefit from global context information when objects are occluded in the right frame.

Table 1
Ablation study of Different size of cost volume cooperates with CMR on Scene Flow.

Network setting	cost volume size		context information		EPE	D1(%)	>1px(%)	>2px(%)	>3px(%)	Time(s)
	same	different	same	CMR						
ablation1	✓		✓		0.50	1.86	5.10	3.07	2.32	0.26
ablation2	✓			✓	0.49	1.83	4.98	3.06	2.27	0.26
ablation3		✓	✓		0.51	1.87	5.21	3.11	2.41	0.19
ablation4		✓		✓	0.47	1.75	4.85	2.91	2.19	0.19

Table 2
Ablation study of different baseline.

Network setting	Scene Flow test EPE	KITTI 2015 val D1	Parm (M)	Time (s)
GwcNet	0.76	1.41	6.5	0.14
PSMNet	0.88	1.50	5.2	0.29
PSM + GRU(CMR)	0.48	1.35	7.7	0.34
ablation4: GwcNet+GRU (CMR)	0.47	1.35	8.9	0.19

Table 3
Ablation study of GDA on a subset dataset of Scene Flow that provides occlusion region mask.

Network	Occlusion region			Scene Flow test		
	EPE	D1	>3px (%)	EPE	D1	>3px (%)
GwcNet	2.33	10.86	12.27	0.76	2.71	–
RAFT-Stereo	2.32	12.40	13.71	0.56	–	2.85
ACVNet	1.97	8.21	9.43	0.48	1.59	2.05
ablation4	1.86	8.11	9.35	0.47	1.75	2.19
HCRNet: ablation4 + GDA	1.82	8.00	9.22	0.47	1.68	2.03

(S_t^c described in Sect.3.2) denote the disparity features, the i^{th}feature vector is denoted $x_i \in R^{D_c}$. GDA module computes the attention weights from left context information, the aggregated disparity features are given by:

$$\widehat{s}_i = s_i + \gamma \sum_{j=1}^{N} f\left(\eta(x_i), \phi(x_j)\right)\sigma(s_i) \tag{4}$$

where γ is a learnable parameter initialized to zero, η and ϕ, are the projection functions for the query, key, and value vectors, and f is a similarity attention function defined as:

$$f(x_i, x_j) = \frac{exp\left(\left(x_i^T xj - M_i\right)/\sqrt{D}\right)}{\sum_{j=1}^{N} exp\left(\left(x_i^T x_j - M_i\right)/\sqrt{D}\right)} \tag{5}$$

where $M_i = max\{x_i^T x_j, j \in [1, N]\}$. Subtracting a constant value M could increase numerical stability of soft max function. The projection functions for the query, key and value vectors are given by:

$$\begin{cases} \eta(x_i) = W_{qry} x_i \\ \phi(x_i) = W_{key} x_i \\ \sigma(s) = W_{val} s \end{cases} \tag{6}$$

where $W_{qry}, W_{key} \in R^{D_{in} \times D_C}$ and $W_{val} \in R^{D_m \times D_m}$. The final output of GDA is $[x|s|\widehat{s}]$ a concatenation of the three feature maps. This output is X_t in eq. (3). Concatenation allows the GRU module intelligently select from or combine the local and global context information. It is plausible that the network learns to decode the aggregated disparity features only when the model cannot be certain about the disparity based on local information. In Fig. 5, we show that when a large part of an object is occluded, considering global context information is beneficial for model. (See Fig. 6.)

3.6. Loss function

For each output module $s \in S, S = \{1, 2, 3\}$, we resize the sequence of the output $\{D_i^s, D_{i+1}^s, \ldots, D_n^s\}$ to the complete resolution size with the upsampling operator, where n denotes the number of iterations of GRUs, and following RAFT-stereo [14], we use the weighted L_1 distance as the loss function.

$$L = \sum_{s} \sum_{i=0}^{3n} \gamma^{3n-i} \left\| d_{gt} - \mu_s\left(D_i^s\right) \right\|_1 \tag{7}$$

Where d_{gt} denotes ground truth disparity, we set γ as 0.9, $n = 5$ during training and inference.

4. Experiment

In this section, we evaluate our proposed stereo model on the Scene Flow dataset [5], KITTI 2012 [32], KITTI 2015 [33], ETH3D [34]. To demonstrate our model performance in occlusion regions, we evaluate on the subset of FlyingThing3D [5] that contain some extremely hard samples such as complex textures and repetitive textures meanwhile occlusion is a natural occurrence. It consists of stereo pairs, ground truth of disparity, and occlusion mask of the left image which masks a visible point in the left image but is not visible in the right image. The experiment settings and training strategy are presented in Sect.4.2. We also performed ablation studies to verify the effectiveness of our network architecture in Sect.4.3. Finally, we compare HCRNet with other state-of-the-art stereo matching methods on public benchmarks. As provided in Sect.4.4, our method has a competitive performance for disparity estimation.

4.1. Datasets and evaluation metrics

4.1.1. Datasets

Scene Flow(finalpass): a large-scale synthetic dataset collection of dense ground-truth disparity maps. This dataset provides 35,454 stereo pairs for training and 4370 pairs for testing. It is large enough to directly train deep learning models without worrying about over-fitting when random cropping is used. Compared to the cleanpass version, finalpass is more like real-world images than the cleanpass. Our model is trained on this dataset, with the exception of normalization, we do not use any additional data augmentation strategy.

KITTI2012: a challenging and varied road scene dataset collected from a driving car which contains 194 stereo pairs for training and 195 pairs for testing. It only provides sparse disparity maps as ground truth for training images and the disparity range is 0–230. Most of the images are collected with real road scenes, illumination changes, complex textures, and specular reflection are the characteristics of this dataset.

KITTI2015: a real-world street views dataset contains 200 training scenes and 200 test scenes collected from a driving car. It only provides sparse disparity maps as ground truth for training and the disparity range is 0–230. The scenes were similar to KITTI2012, while it comprises dynamic scenes for which the ground truth has been established in a semi-automatic process that means more precision.

ETH3D: ETH3D contains indoor and outdoor scenes which provides 27 training and 20 testing image pairs with sparse disparity maps. The

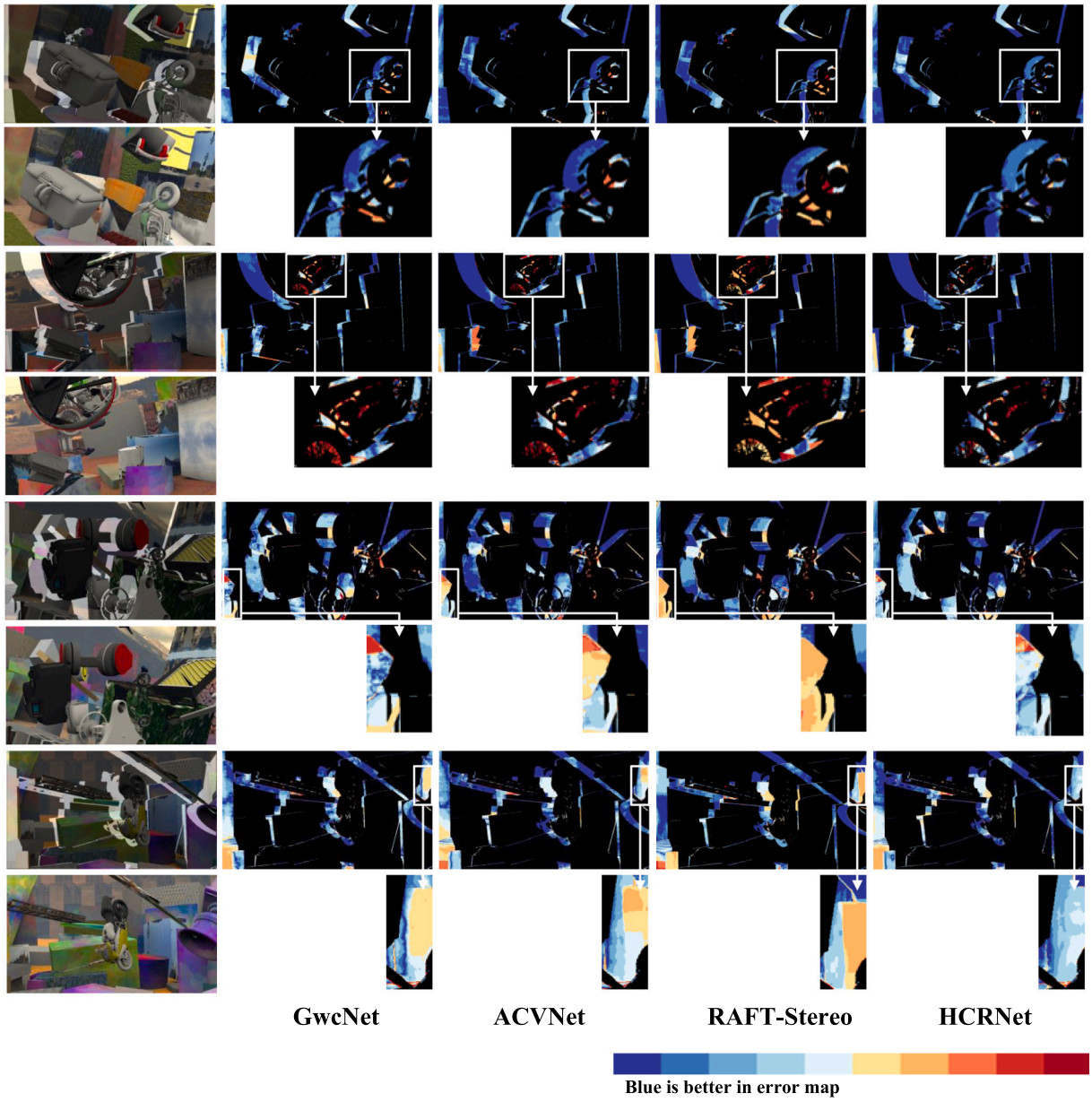

Fig. 7. Visual comparisons with other methods on the occluded regions. In the left image, we indicate with a transparent mask that the region is occluded in the right image.

Table 4
Comparison with other methods on the KITTI2015 stereo benchmark. We report D1 metric in background regions (bg), foreground areas (fg), and all. The foreground is not occluded denoted as D1-fg Noc, all foreground is denoted as D1-fg All. **Bold**: Best, Underscore: Second best.

Method	KITTI2015					
	D1-bg All	D1-bg Noc	D1-fg All	D1-fg Noc	D1-all All	D1-all Noc
MDCNet [37]	1.76	1.61	3.68	3.26	2.08	1.88
AcfNet [12]	**1.51**	**1.36**	3.80	3.49	1.89	1.72
AANet+ [38]	1.65	1.49	3.96	3.66	2.03	1.85
GwcNet-g [15]	1.74	1.61	3.93	3.49	2.11	1.92
EdgeStereo-v2 [10]	1.84	1.69	3.30	2.94	2.08	1.84
PSMNet [3]	1.86	1.71	4.62	4.31	2.32	2.14
HITNet [7]	1.74	1.54	**3.20**	**2.72**	1.98	1.74
HCRNet(ours)	**1.51**	1.38	3.51	3.32	**1.85**	**1.70**

disparity range of ETH3D is 0–64. It covers a variety of indoor and outdoor scenes, its images are not only low-resolution but also gray. The scenes contain strong light and low brightness cases.

4.1.2. Evaluation metrics

Flowing previous related work, for the Scene Flow dataset, the evaluation metric is the average end-point error defined as $EPE = \frac{1}{n}\sum_{i=1}^{n}|g_i - e_i|$, where g_i is the ground truth of one pixel, e_i is the estimation of the pixel, n is the number of valid pixels, for the KITTI 2015 dataset, a pixel to be incorrectly estimated if the disparity end-point error is $\geq 3px$ and $> 5\%$, i.e. $D1 = \frac{1}{n}\sum_{i=1}^{n}|g_i - e_i|$ (where valid pixels is $|g_i - e_i| > 3$, and $|g_i - e_i| > 0.05g_i$). for the KITTI2012 dataset, the percentage of pixels with errors larger that disparities in both non-occluded (x-noc) and all regions (x-all) are used, for ETH3D dataset, bad pixel percentages are computed, $Bad1.0 = \frac{1}{n}\sum_{i=1}^{n}|g_i - e_i|.|g_i - e_i| > 1$.

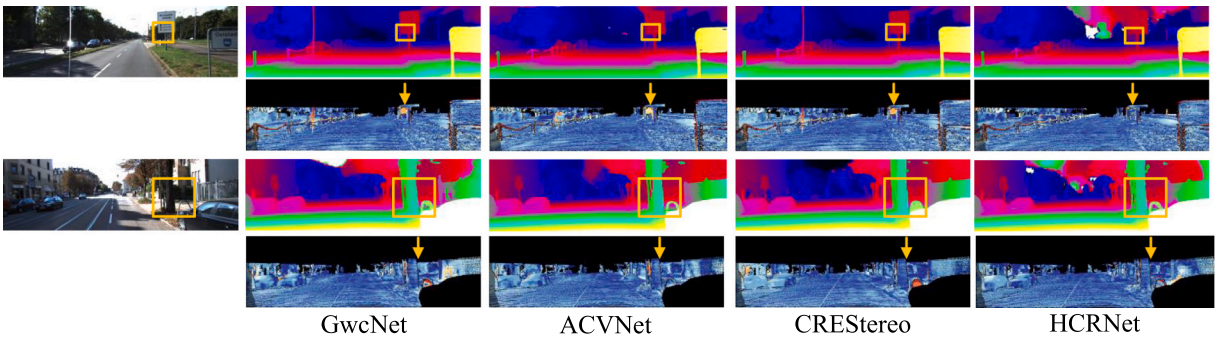

Fig. 8. Visual comparisons with other methods on KITTI2015 dataset. The first column shows the left input image. For each subsequent row, the first row shows the predicted colorized disparity map and the second row shows the error map.

Table 5
Comparison with other methods on the KITTI2012 stereo benchmark. **Bold**: Best, Underscore: Second best.

Method	KITTI2012							
	2-noc (%)	2-all (%)	3-noc (%)	3-all (%)	5-noc (%)	5-all (%)	Avg-Noc	Avg-All
CREStereo [22]	1.72	**2.18**	1.14	1.46	0.76	0.95	0.5	0.5
ACVNet [18]	1.83	2.34	1.13	1.47	0.71	0.91	**0.4**	0.5
OptStereo [11]	1.91	2.51	1.20	1.61	0.77	1.02	**0.4**	0.5
LEAStereo [39]	1.90	2.39	1.13	1.45	**0.67**	0.88	0.5	0.5
CFNet [40]	1.90	2.43	1.23	1.58	0.74	0.94	0.5	0.5
PSMNet [3]	2.44	3.01	1.49	1.89	0.90	1.15	0.5	0.60
GwcNet-gc [15]	2.15	2.71	1.32	1.70	0.80	1.03	0.5	0.5
HCRNet (ours)	**1.69**	**2.18**	**1.09**	**1.42**	0.69	**0.87**	**0.4**	**0.4**

4.2. Implementation details

We implement our architectures using Pytorch and the final model is trained using NVIDIA A40 GPUs. All model are optimized with AdamW [35] optimizer. The range of disparity value set as 192. In ablation study, models are trained on Scene Flow for 20 epochs with a batch size of 4. Our final results were trained for 45 epochs with a batch size of 8. All model trained and inference with 5 updates in each hourglass stage. We use a one-cycle learning rate schedule [36] with a minimum learning rate of $1e^{-5}$. All experiments were trained on random 320×640 crops. In the fine-tuning process on the KITTI 2012, KITTI2015, the minimum learning rate was set $1e^{-6}$, All experiments were set 300 epochs with a batch size of 4 and trained on random 256×960 crops. For ETH3D, the minimum learning rate was set $1e^{-6}$, we set 600 epochs with a batch size of 4 with random 256×512 crops.

4.3. Ablation studies

This section presents the ablation studies to verify effectiveness of model structure and explore the best setting for our network. Following previous related work, we compare the HCRNet results with different

Table 6
Comparison with other methods on the ETH3D stereo benchmark. **Bold**: Best, Underscore: Second best.

Method	ETH3D		
	Bad 1.0	Bad 2.0	AvgErr
AdaStereo [41]	3.09	0.65	0.24
NLCA-Net [42]	3.84	1.04	0.27
DANet [43]	6.02	1.74	0.35
GwcNet [15]	6.42	1.67	0.35
ACVNet [18]	2.58	**0.57**	**0.23**
HSMNet [44]	4.2	1.4	0.27
CBMV [45]	5.35	1.56	0.33
HCRNet(ours)	**2.56**	0.73	0.24

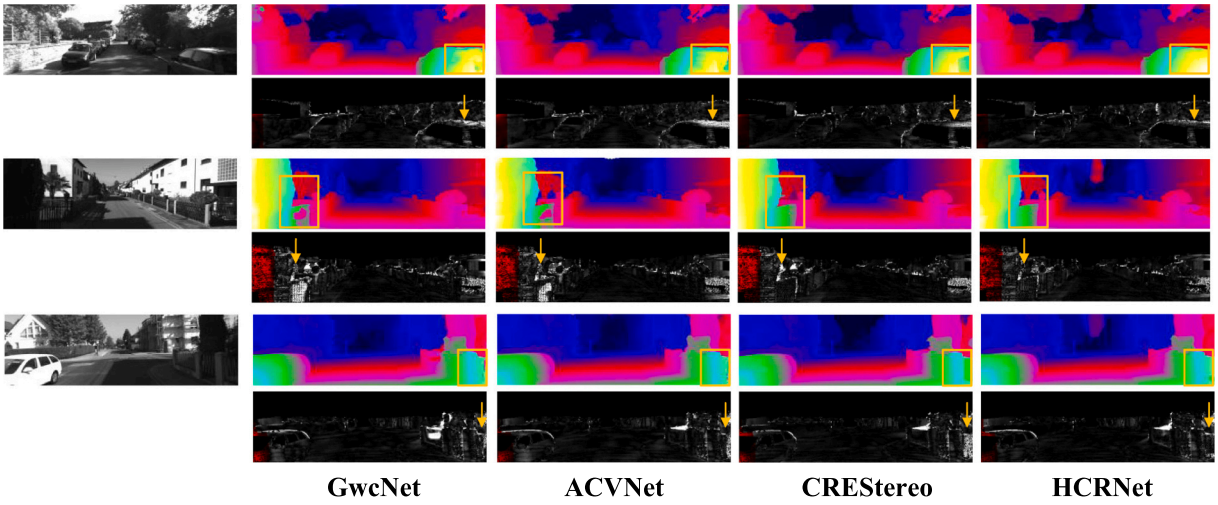

Fig. 9. Visual comparisons with other methods on the KITTI2012 dataset. The first column shows the left input image. For each subsequent row, the first row shows the predicted colorized disparity map and the second row shows the error map.

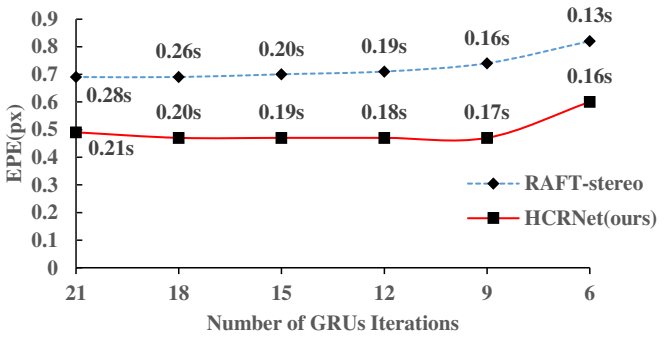

Fig. 10. Evaluate the performance of our method against RAFT-Stereo on the Scene Flow test set under varying iterations.

settings by computing EPE on Scene Flow, D1 on KITTI 2015 dataset, and inference time on KITTI2015. In addition, we evaluate the performance of the model in occluded region on a subset of Scene Flow which is selected by the literature [5] and provides occlusion mask.

4.3.1. Ablation study of context information of multi-level receptive fields

In this ablation study, we considered four different designs, first one is that the same size cost volume(1/4) cooperates context information of the same receptive fields (larger), second one is that the same size cost volume(1/4) cooperates with context information of multi-level receptive fields(CMR), third one is that different size of cost volume cooperates with context information of the same receptive fields(larger), last one is our HCRNet that different size of cost volume cooperates with CMR. In Table 1, we show that different size of cost volume cooperates with CMR could improve the performance of model while reduce the inference time compared to other settings. Specifically, integration of different context information can improve accuracy by about 7%, using different size of cost volume can improve accuracy by about 4% and reduce time consumption by 26%.

4.3.2. Ablation study of different baseline

In this ablation study, we demonstrated our method that GRU cooperates with CMR could improve the accuracy of popular baseline in Table 2, such as PSMNet and GwcNet(as the baseline of HCRNet). As a result, we improved GwcNet and PSMNet by 38% and 45% respectively.

4.3.3. Ablation study of GDA module

In this ablation study, we demonstrated that the GDA module could improve the accuracy of the estimate in the occluded region while retaining the accuracy of the model in the benign region, as shown in Table 3. For occlusion region, a subset dataset of Scene Flow that provided occlusion mask was used to evaluate (compute EPE only on the occluded region). Qualitative results on Scene Flow are shown in Fig. 7.

In Fig. 7, we focus on the occluded region of the scenes (the region that is not obscured by the black mask). From left to right, both GwcNet and ACVNet adopt 3D convolution to aggregate cost, the cost of neighborhood pixels is taken in to account but the local information is still insufficient to handle occluded region, RAFT-stereo can consider scope that limited by the context information which depends on the backbone network. For HCRNet, our GDA module could help the model consider global information so that we achieve better prediction results in occluded regions. In Table 3, we also find the same result. The performance in the occluded region exhibits a deterioration as the extent of regional information reduction decreases. RAFT-stereo is worse than both GWC and ACV on metric D1, while our method HCRnet achieves better results because global information is taken into account.

4.4. Comparison with other stereo matching methods

For the KITTI2015 benchmark, D1 is used to evaluate method performance, we report D1 metric in background regions (D1-bg), foreground areas (D1-fg), and all (D1-all). We compare our final HCRNet model with different state-of-the-art methods.

In Table 4, we show that our results outperform most soft argmin based regression methods, indicating that the GRU based regression method still has the advantage of accuracy in real-world scenarios. Our method surpasses baseline GwcNet by 13% on D1-all All. In Fig. 8, from top to bottom we show two different scenes in the road scene, in the first scene we find that only our model can predict the pure black indicator without texture well, in the second scene the model demonstrates the ability to handle the details of the scene.

For the KITTI2012 benchmark, the evaluation results are shown in Table 5. We report the percentage of pixels with errors larger that disparities in both non-occluded (−noc) and all regions(−all). Our method surpasses GwcNet by 16% on 3-all while achieve state-of-the-art accuracy on KITTI2012 dataset.

Table 5 shows that in comparison to soft argmin based methods, our method demonstrates superior performance, with a 2% increase in metric 3-all compared to CREStereo, another GRU based method. Our method surpasses baseline GwcNet by 16% on 3-all. In Fig. 9, from top to bottom we focus on three different difficult region that involves reflection, solid color, shadow. Overall, their issues in image representation are lack of texture, in these specific regions, our model exhibits advantage compared to others.

For the ETH3D benchmark, the evaluation results are shown in Table 6. Bad 1.0,2.0 denotes fraction of pixels with errors larger than the given number of disparities. Average error (AvgErr) denotes the per-pixel average disparity error. In Table 6, we show that our method out performs baseline GwcNet by 60% on ETH3D benchmark.

To demonstrate the advantages of our regression approach across different scales of cost volumes, we contrast it with RAFT-stereo, a method that performs regression at a constant scale. We consider the accuracy achieved by both models under varying numbers of iterations, as well as the computational time expended during these processes, with the detailed results presented in Fig. 10. The experimental platform utilized for this experiment is an Intel(R) Core(TM) i9-10850K processor and an NVIDIA GeForce RTX 3090 graphics card.As show in Fig. 10, we can find that our model has the advantage of accuracy under various iterations. In terms of time consumption of the model, although we added stacked hourglass module compared with RAFT-Stereo, as the number of iterations increases, our time efficiency advantage becomes more pronounced. Because in our approach, we take three iterations as a group, and within this group we perform regression on cost volumes at three distinct scales (1/4, 1/8, and 1/16), whereas RAFT-Stereo conducts its iteration at the same scale (1/4) three times.

5. Conclusion

In this paper, we proposed HCRNet to estimate disparity maps, which combines hourglass module and GRU. We showed that replace soft argmin with GRU could improve performance of learning-based model, specifically, hourglass based model. We show that global context information could improve the performance of model in occluded region. Experiment also show that our method has an exciting performance on Scene Flow, KITTI2012, ETH3D, and KITTI2015. Even there are many benchmarks current, due to the sparse disparity map and relatively less real-world ground truth, the performance of most models on real world is poor. We aim to improve the effect of sparse disparity utilization for the stereo matching method in our future work.

CRediT authorship contribution statement

Tuming Yuan: Writing – review & editing, Writing – original draft, Methodology, Investigation, Data curation, Conceptualization. **Jiancheng Hu:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization. **ShuangJiang Ou:** Writing – original draft,

Supervision, Methodology, Conceptualization. **Weijia Yang:** Writing – original draft, Conceptualization. **Yafang Hei:** Writing – original draft, Conceptualization.

Declaration of competing interest

The authors declared that they have no conflicts of interest to this work.

Data availability

The data that support the findings of this study are available in KITTI 2015 with identifier https://doi.org/10.1109/CVPR.2015.7298925 [33], KITTI 2012 with identifier https://doi.org/10.1109/CVPR.2012.6248074 [32], ETH3D with identifier https://doi.org/10.1109/CVPR.2017.272 [34] and Scene Flow with identifier https://doi.org/10.1109/CVPR.2016.438 [5].

References

[1] C. Kerl, J. Sturm, D. Cremers, Robust odometry estimation for rgb-d cameras, in: 2013 IEEE international conference on robotics and automation, IEEE, 2013, pp. 3748–3754.

[2] Y. Yao, Z. Luo, S. Li, T. Fang, L. Quan, Mvsnet: Depth inference for unstructured multi-view stereo, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 767–783.

[3] J.-R. Chang, Y.-S. Chen, Pyramid stereo matching network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5410–5418.

[4] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, 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.

[5] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, T. Brox, A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4040–4048.

[6] D. Scharstein, R. Szeliski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, Int. J. Comput. Vis. 47 (2002) 7–42.

[7] V. Tankovich, C. Hane, Y. Zhang, A. Kowdle, S. Fanello, S. Bouaziz, Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14362–14372.

[8] H. Hirschmuller, Stereo processing by semiglobal matching and mutual information, IEEE Trans. Pattern Anal. Mach. Intell. 30 (2) (2007) 328–341.

[9] J. Sun, N.-N. Zheng, H.-Y. Shum, Stereo matching using belief propagation, IEEE Trans. Pattern Anal. Mach. Intell. 25 (7) (2003) 787–800.

[10] X. Song, X. Zhao, L. Fang, H. Hu, Y. Yu, Edgestereo: an effective multi-task learning network for stereo matching and edge detection, Int. J. Comput. Vis. 128 (4) (2020) 910–930.

[11] H. Wang, R. Fan, P. Cai, M. Liu, Pvstereo: pyramid voting module for end-to-end self-supervised stereo matching, IEEE Robot. Autom. Lett. 6 (3) (2021) 4353–4360.

[12] Y. Zhang, Y. Chen, X. Bai, S. Yu, K. Yu, Z. Li, K. Yang, Adaptive unimodal cost volume filtering for deep stereo matching, in: Proceedings of the AAAI Conference on Artificial Intelligence Vol. 34, 2020, pp. 12926–12934.

[13] B. Liu, H. Yu, Y. Long, Local similarity pattern and cost self-reassembling for deep stereo matching networks, in: Proceedings of the AAAI Conference on Artificial Intelligence Vol. 36, 2022, pp. 1647–1655.

[14] L. Lipson, Z. Teed, J. Deng, Raft-stereo: Multilevel recurrent field transforms for stereo matching, in: 2021 International Conference on 3D Vision (3DV), IEEE, 2021, pp. 218–227.

[15] X. Guo, K. Yang, W. Yang, X. Wang, H. Li, Group-wise correlation stereo network, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 3273–3282.

[16] Z. Shen, Y. Dai, Z. Rao, Cfnet: Cascade and fused cost volume for robust stereo matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13906–13915.

[17] Z. Shen, Y. Dai, X. Song, Z. Rao, D. Zhou, L. Zhang, Pcw-net: Pyramid combination and warping cost volume for stereo matching, in: European Conference on Computer Vision, Springer, 2022, pp. 280–297.

[18] G. Xu, J. Cheng, P. Guo, X. Yang, Attention concatenation volume for accurate and efficient stereo matching, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 12981–12990.

[19] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII 14, Springer, 2016, pp. 483–499.

[20] X. Yang, Z. Feng, Y. Zhao, G. Zhang, L. He, Edge supervision and multi-scale cost volume for stereo matching, Image Vis. Comput. 117 (2022) 104336.

[21] Y. Zhang, Y. Li, C. Wu, B. Liu, Attention-guided aggregation stereo matching network, Image Vis. Comput. 106 (2021) 104088.

[22] J. Li, P. Wang, P. Xiong, T. Cai, Z. Yan, L. Yang, J. Liu, H. Fan, S. Liu, Practical stereo matching via cascaded recurrent network with adaptive correlation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 16263–16272.

[23] Z. Teed, J. Deng, Raft: Recurrent all-pairs field transforms for optical flow, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, Springer, 2020, pp. 402–419.

[24] X. Mei, X. Sun, M. Zhou, S. Jiao, H. Wang, X. Zhang, On building an accurate stereo matching system on graphics hardware, in: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), IEEE, 2011, pp. 467–474.

[25] T. Suzuki, Implicit integration of superpixel segmentation into fully convolutional networks, arXiv preprint arXiv:2103.03435, 2021.

[26] F. Yang, Q. Sun, H. Jin, Z. Zhou, Superpixel segmentation with fully convolutional networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 13964–13973.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Proces. Syst. 30 (2017).

[28] X. Cheng, Y. Zhong, M. Harandi, T. Drummond, Z. Wang, Z. Ge, Deep laparoscopic stereo matching with transformers, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 464–474.

[29] W. Guo, Z. Li, Y. Yang, Z. Wang, R.H. Taylor, M. Unberath, A. Yuille, Y. Li, Context-enhanced stereo transformer, in: European Conference on Computer Vision, Springer, 2022, pp. 263–279.

[30] S. Jiang, D. Campbell, Y. Lu, H. Li, R. Hartley, Learning to estimate hidden motions with global motion aggregation, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 9772–9781.

[31] Y. Li, Y. Chen, N. Wang, Z. Zhang, Scale-aware trident networks for object detection, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6054–6063.

[32] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? the kitti vision benchmark suite, in: 2012 IEEE conference on computer vision and pattern recognition, IEEE, 2012, pp. 3354–3361.

[33] M. Menze, A. Geiger, Object scene flow for autonomous vehicles, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3061–3070.

[34] T. Schops, J.L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, A. Geiger, A multi-view stereo benchmark with high-resolution images and multi-camera videos, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3260–3269.

[35] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101, 2017.

[36] L.N. Smith, N. Topin, Super-convergence: Very fast training of neural networks using large learning rates, in: Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications vol. 11006, SPIE, 2019, pp. 369–386.

[37] W. Chen, X. Jia, M. Wu, Z. Liang, Multi-dimensional cooperative network for stereo matching, IEEE Robot. Autom. Lett. 7 (1) (2021) 581–587.

[38] H. Xu, J. Zhang, Aanet: Adaptive aggregation network for efficient stereo matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1959–1968.

[39] X. Cheng, Y. Zhong, M. Harandi, Y. Dai, X. Chang, H. Li, T. Drummond, Z. Ge, Hierarchical neural architecture search for deep stereo matching, Adv. Neural Inf. Proces. Syst. 33 (2020) 22158–22169.

[40] Z. Shen, Y. Dai, Z. Rao, Cfnet: Cascade and fused cost volume for robust stereo matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13906–13915.

[41] X. Song, G. Yang, X. Zhu, H. Zhou, Z. Wang, J. Shi, Adastereo: a simple and efficient approach for adaptive stereo matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10328–10337.

[42] Z. Rao, M. He, Y. Dai, Z. Zhu, B. Li, R. He, Nlca-net: a non-local context attention network for stereo matching, APSIPA Transactions on Signal and Information Processing 9 (2020) e18.

[43] Z. Ling, K. Yang, J. Li, Y. Zhang, X. Gao, L. Luo, L. Xie, Domain-adaptive modules for stereo matching network, Neurocomputing 461 (2021) 217–227.

[44] G. Yang, J. Manela, M. Happold, D. Ramanan, Hierarchical deep stereo matching on high-resolution images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5515–5524.

[45] K. Batsos, C. Cai, P. Mordohai, Cbmv: A coalesced bidirectional matching volume for disparity estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2060–2069.