SIP (2020), vol. 9, e18, page 1 of 13 © The Author(s), 2020. Published by Cambridge University Press in association with [APSIPA Transactions on Signal and Information Processing]. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.
doi:10.1017/ATSIP.2020.16

ORIGINAL PAPER

NLCA-Net: a non-local context attention network for stereo matching

ZHIBO RAO,[1] MINGYI HE,[1] YUCHAO DAI,[1] ZHIDONG ZHU,[1] BO LI[1] AND RENJIE HE[1,2]

Accurate disparity prediction is a hot spot in computer vision, and how to efficiently exploit contextual information is the key to improve the performance. In this paper, we propose a simple yet effective non-local context attention network to exploit the global context information by using attention mechanisms and semantic information for stereo matching. First, we develop a 2D geometry feature learning module to get a more discriminative representation by taking advantage of multi-scale features and form them into the variance-based cost volume. Then, we construct a non-local attention matching module by using the non-local block and hierarchical 3D convolutions, which can effectively regularize the cost volume and capture the global contextual information. Finally, we adopt a geometry refinement module to refine the disparity map to further improve the performance. Moreover, we add the warping loss function to help the model learn the matching rule of the non-occluded region. Our experiments show that (1) our approach achieves competitive results on KITTI and SceneFlow datasets in the end-point error and the fraction of erroneous pixels (D_1); (2) our proposed method particularly has superior performance in the reflective regions and occluded areas.

Keywords: Stereo matching, Non-local attention, Geometry context, Geometry refine

Received 30 April 2019; Revised 7 June 2020

I. INTRODUCTION

Stereo matching plays an essential role in computer vision tasks, including autonomous driving [1, 2], object detection and recognition [3, 4], and 3D reconstruction and understanding [5–7]. For a couple of rectified stereo images, disparity refers to the apparent pixel difference or motion between a pair of corresponding pixels on the left and right images [8, 9].

The dense disparity map estimation methods have been studied for many years. For the traditional stereo matching methods (e.g. semi-global matching (SGM) [9], non-local cost aggregation [10], second-order smoothness priors [11]), the classical pipeline involves the finding of corresponding points by matching cost, cost aggregation, optimization, disparity refinement, and post-processing based on the local or global features. In general, the traditional methods often focus on using the prior knowledge of images to construct the warping function through for improving matching accuracy.

For disparity prediction based on deep learning, recent efforts have yielded many high-quality outputs due to deep fully convolutional neural networks (FCN) [12, 13] and a large amount of training data [14–16]. Classically, mainstream methods contain a four-step pipeline, while each step is important to the overall matching performance: 2D feature extraction, cost volume construction, 3D feature matching, and disparity regression [13, 17, 18]. Zbontar and LeCun first calculated the matching costs by convolutional neural networks (CNNs) to improve the performance, and the result showed that CNNs could learn more robust features from images and produced reliable matching cost in this task [17]. Following this work, many researchers [13, 19–22] proposed several methods to improve matching accuracy [13, 22], reduce some parameters [18], or achieve self-supervision ability [21].

Albeit the above success, the stereo matching methods based on deep learning still exist some limitations. First, the prediction pixels have terrible performance in the occluded, repeated object, and reflective regions [23, 24] due to a lack of sufficient understanding of the scene. Second, it is difficult to improve further accurate correspondence estimation if solely applying the concatenation operation between different viewpoints in cost volume construction [13, 20]; it is caused by the concatenation operation that is lack of the physical meaning about similarity. Due to the above problems, the performance of used networks has encountered bottlenecks.

In this paper, we propose a novel non-local context attention network (NLCA-Net) to exploit the global context information for stereo matching. First, we utilize spatial

[1]Northwestern Polytechnical University, Xian 710129, China
[2]Nanyang Technological University, 639798 Singapore, Singapore

Corresponding author:
Mingyi He
Email: myhe@nwpu.edu.cn

pyramid pooling (SPP) and dilated convolutions to extract the semantic information. Next, we apply a variance-based method to build cost volume. Then, we use the hierarchical 3D convolution and non-local block [25] to set up the non-local attention matching (NLAM) module for regularizing the cost volume. Finally, we adopt a soft argmin operation to get the initial disparity map and then refine it via combining the semantic information. At last, we further improve the accuracy of the non-occlusion region by the warping loss function.

Our main contributions are listed below:

- We design a non-local context attention module to exploit the global context information for regularizing the cost volume, thus improving the performance of the matching task, particularly on the occlusion.
- We use a variance-based method instead of traditional concatenation operation to build cost volume, which provides the similarity information and reduces some memory.

II. RELATED WORK

To improve the accuracy of disparity map estimation in stereo matching, many researchers have tried to optimize cost volume or matching cost computation and got fantastic achievements. Interested readers are suggested to read the surveys to get an overview of the typical matching algorithms and different optimization methods [26–28]. In this section, we will focus on a brief discussion about the related methods, involving traditional methods, deep leaning matching methods, and semantic segmentation methods, respectively.

In general, traditional stereo matching methods care more about how to compute the matching cost accurately and how to apply local or global features to refine the disparity map [29, 30]. Guney and Geiger used inverse graphics techniques to integrate objects as a non-local regularizer, then applied the conditional random field (CRF) framework to refine the disparity map; its result showed the value of this method on the KITTI dataset [31]. Seki and Pollefeys developed deep neural networks based on SGM for predicting accurate dense disparity map and introduced a novel loss function that fully uses sparsely annotated disparity maps features; their method replaced manually-tuned penalties for regularization [23]. Moreover, Gidaris and Komodakis proposed a generic architecture that improved the labels by detecting incorrect labels, replacing incorrect labels with new ones, and refining the renewed labels (DRR); their method achieved a significant improvement surpassing prior approaches [32]. These methods used the ideas of traditional disparity map post-processing to reduce the mismatch in ambiguous regions and improve disparity estimation.

Recently, in stereo matching areas, the end-to-end networks have been developed to predict whole disparity maps without post-processing. Mayer *et al.* introduced two end-to-end networks for estimating disparity (DispNet) and optical flow (FlowNet), and created a large synthetic dataset called SceneFlow, which improved the performance [14]. Chang and Chen introduced PSMNet, an end-to-end network for feature fusion using SPP and dilated convolution architectures [20]. Zhong *et al.* used image warping error as the loss function to drive the learning process, achieving a self-improving ability [21]. Kendall *et al.* exploited the way of cost volume regularization and shown 3D convolutions' effect in the context learning of stereo matching [13]. Guo *et al.* divided left–right features into different groups to obtain multiple matching cost proposals for measuring feature similarities and reducing some parameters [18]. Yin *et al.* composed local matching distributions to form the global match density for lessening the disparity candidates [22]. The main idea of these methods was to construct the cost volume or use external information (e.g. optical flow or edge) to improve the accuracy of disparity estimation, ignoring the effect of global scene understanding.

In the field of semantic segmentation, how to fuse the context information is an important topic. Many researchers proposed different methods to exploit global context information and make substantial progress in recent years. Long *et al.* demonstrated the value of the FCN in the semantic segmentation, and the performance had been dramatically improved [33]. Chen *et al.* designed a DeepLab_v3 that could capture multi-scale context with further boost performance by adopting multiple atrous rates, and the system of DeepLab_v3 without DenseCRF post-processing [34]. Ranjan and Black proposed the SPyNet, which introduced image pyramids to predict optical flow by a coarse-to-fine approach [35]. The above approaches showed that the idea of multi-scale architecture was essential for exploiting global context information in the field of semantic segmentation.

In this work, we exploit the potential of the non-local attention mechanism to enhance the scene understanding at the global-scope level. Moreover, we construct the cost volume by the variance-based method to add the similarity information compared with traditional concatenation operation. As described above, we propose the NLCA-Net for improving the matching accuracy, especially in the occlusions and reflective regions.

III. OUR METHOD

In this section, we propose the NLCA-Net. The network architecture is illustrated in Fig. 1, and the detailed parameters are presented in Appendix A. Our model consists of five parts: feature extraction and fusion, cost volume construction, feature matching, disparity map regression, and refinement. The implementation detail is described in the following sub-sections, respectively.

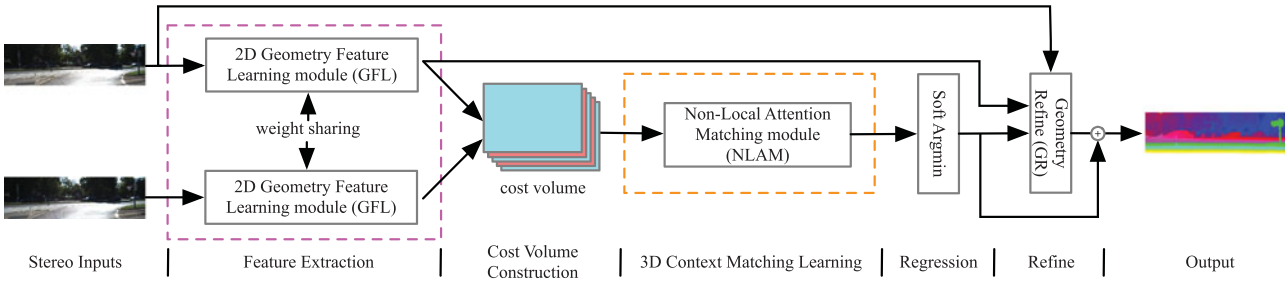

Fig. 1. Our end-to-end deep stereo regression architecture, NLCA-Net (Non-Local Context Attention network). Our model consists of three modules: 2D geometry feature learning (GFL), non-local attention matching (NLAM), and geometry refinement (GR) module.

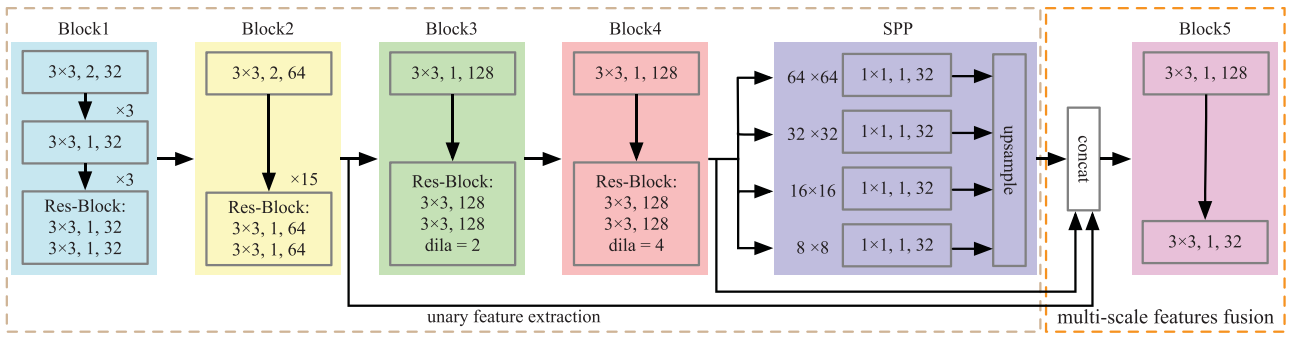

Fig. 2. The 2D geometry feature learning module (GFL). $x \times x, s, f$ denote the size of the convolution kernel, stride, and the number of convolution filters respectively. $\times n$ denotes the block repeats n times.

A) Feature extraction and fusion by 2D geometry feature learning module

Previous methods failed to predict an accurate disparity in ill-posed regions because the network did not understand the context well. On the other hand, semantic segmentation methods have a fantastic performance in understanding the context. Thus, we are inspired by many semantic segmentation methods and get a robust descriptor that determines the context relationship from the pixel (particularly for ill-posed regions) via the SPP struct. We design the 2D geometry feature learning (GFL) module, as shown in Fig. 2. For description convenience, we set the input resolution of the stereo pairs to be $H \times W$, and the basic number F of the convolution filter to be 32.

In this module, we apply a series of 2D convolutional operations to extract the semantic information, and each convolutional operation is followed by a batch normalization (BN) layer and a rectified linear unit (ReLU) layer. The GFL module consists of the unary feature extraction part and the multi-scale feature fusion part.

(1) The unary feature extraction part

The unary feature extraction part contains the block 1–4 and SPP, which are the basic or residual unit for learning the unary feature. In block 3 and block 4, dilated convolution is applied to enlarge the receptive field further. In SPP, we use the multi-scale average pooling to compress features and a 1×1 convolution to reduce feature dimension; then, the feature maps are upsampled to the original size. Next, we concatenate the unary feature of block 2, block 4, and SPP. After the unary feature extraction part, we could obtain the aggregated unary feature volume with the size $H/4 \times W/4 \times 10F$.

(2) The multi-scale feature fusion part

The multi-scale feature fusion part is block 5, which is used for fusing the aggregated unary feature volume. To avoid losing the critical information, we first adopt 128 convolutional filters with the size 3×3 to fuse them, then use the 32 convolutional filters with the size 3×3 to reduce feature dimension. After the multi-scale feature fusion part, we could obtain the semantic information with the size $H/4 \times W/4 \times F$.

B) Cost volume construction by variance-based cost metric

In previous works [13, 19–21], the cost volume is the critical step which links 2D and 3D convolution. To achieve better performance, we aggregate the semantics feature of left and right image V_l, V_r to one cost volume C via variance-based cost metric (VBCM) $\mathcal{M}$. Let W, H, D, F be the input image height and width, the disparity sample number, and the feature number. Thus, the size of the semantics feature is $V_l = V_r = (H/4) \times (W/4) \times F$, and the size of cost volume $C = (D/4) \times (H/4) \times (W/4) \times F$. We define the cost metric as the mapping $\mathcal{M} : \underbrace{\{V_l, V_{r,1}\}, \ldots, \{V_l, V_{r,(D/4)}\}}_{D/4} \rightarrow$

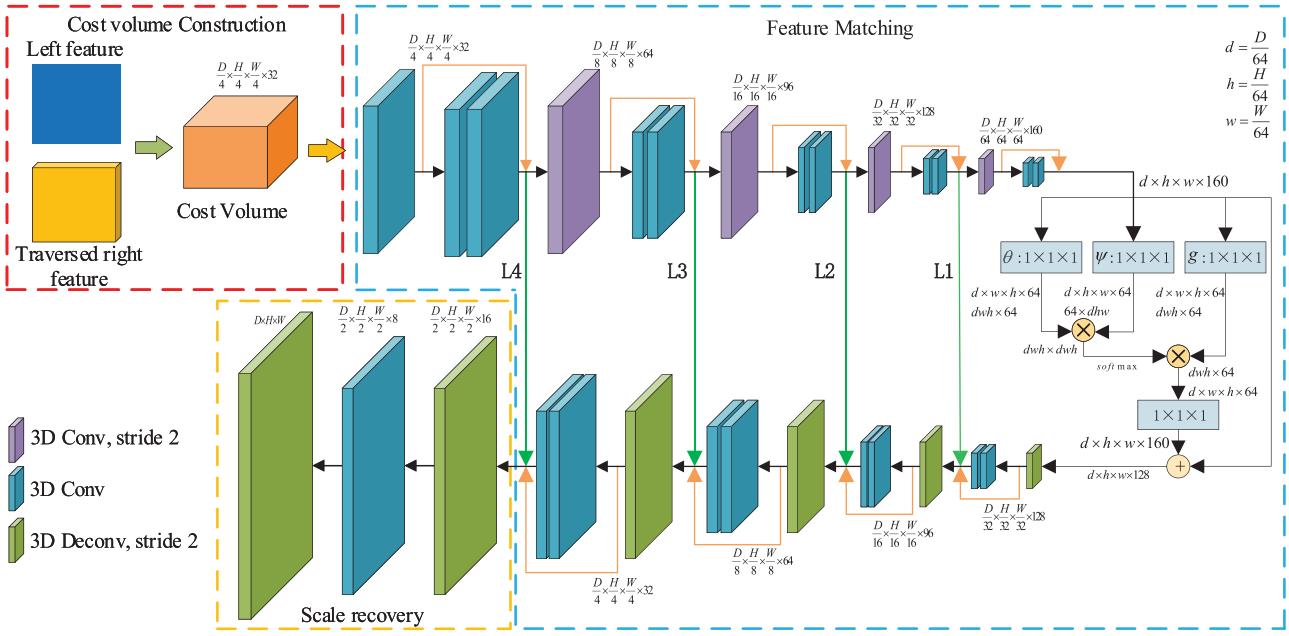

Fig. 3. The non-local attention matching module (NLAM). The NLAM module consists of feature matching part and scale recovery part. Note that the feature maps are shown as feature dimensions, e.g. $D \times H \times W \times F$means a feature map with disparity number$D$, height$H$, width W, and feature number$F$. Here, $L*$denotes different scale levels of the feature maps.

C that:

$$C = \mathcal{M}\left(\{V_l, V_{r,1}\}, \cdots, \left\{V_l, V_{r,\frac{D}{4}}\right\}\right) = stack\left(\frac{(V_l - \overline{V_i})^2 + (V_{r,i} - \overline{V_i})^2}{2}\right), \quad (1)$$

where $V_{r,i}$ means traversed right semantics feature with a preset disparity range i, $\overline{V_i}$ means the average of V_l and $V_{r,i}$, and all operations above are element-wise.

Most traditional stereo matching methods aggregate the cost volume between the left and right images in a heuristic way. However, recent works apply the concatenation operation instead of the mean or subtraction operation [13, 36]. This is the way that depends on network learning entirely. Here we choose the variance-based operation instead of the concatenation operation, due to which provides no direction about what the networks should do in the feature matching module. In contrast, our variance-based operation explicitly measures the left–right feature difference, which reflects the similarities between them and saves about half of memory. The true matched pair should have the lowest cost value, whereas it should have a higher cost. The output size of variance-based cost volume is $D/4 \times H/4 \times W/4 \times F$.

C) Feature matching by non-local attention matching module

To regularize the matching cost volume along the disparity dimension as well as spatial dimensions, we propose a 3D CNN architecture for learning the matching feature: the NLAM module. In [25], the non-local block was designed to compute the response at a position as a weighted sum of the features at all positions, and it showed a significant improvement for video classification and poses estimation. However, the cost volume is too big for the non-local block, leading it cannot be directly applied to the matching task. From another point, the essence of the non-local block is the attention mechanism. Therefore, we could combine the non-local block and the hierarchical 3D convolution for setting up the NLAM module, as shown in Fig. 3.

In this module, we apply a series of 3D convolutional operations to obtain the matching volume, and each convolutional operation is followed by a BN layer and a ReLU layer. The NLAM module consists of feature matching part and scale recovery part.

(1) The feature matching part

The feature matching part contains 26 convolutions with stride one or two for regularizing the variance-based cost volume. This part has four levels, and we pass the feature maps between the same level to form the residual architecture, avoiding losing the critical information. Each level consists of an up-sampling or a sub-sampling convolution, and a residual block. After the $L4$ level, we adopt a non-local block as an attention block for further improving global matching learning.

To further understand the non-local attention mechanism, our feature matching part could be viewed as the group of the hierarchical 3D convolution block and the non-local block. The hierarchical 3D convolution block is an encoder–decoder architecture; it encodes the feature map by sub-sampling and decodes the encoded feature by up-sampling, as shown in Fig. 4.

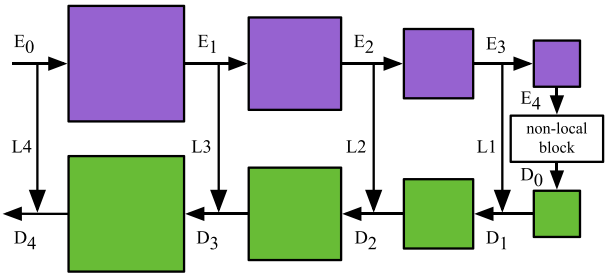

Fig. 4. The encoder–decoder architecture. The pink block means the encoding process. The green block means the decoding process.

As shown in Fig. 4, we define the encoder and decoder process, respectively as:

$$\begin{aligned} E_{i+1} &= \mathcal{F}_{E_{i+1}}(E_i) \\ D_i &= \mathcal{F}_{D_i}(D_{i-1}) + E_{n-i}, \end{aligned} \tag{2}$$

where n denotes the number of the encoder and decoder, i denotes the level of the encoder or decoder, and $\mathcal{F}_E$ or $\mathcal{F}_D$ denotes the process of each encoder or decoder.

In the encoder–decoder architecture, the worst drawback is that the convolution is a local operation. It causes the network to not further improve the receptive fields to evaluate the impact of global information on the current pixel. Thus, we use the non-local block as attention mode for promoting the understanding ability of global context, as shown in Fig. 3.

The non-local operation $\mathcal{N}(\cdot)$ could be represented as:

$$y_i = \frac{1}{\mathcal{C}(x)} \sum_{\forall j} f(x_i, x_j) g(x_j), \tag{3}$$

where x, y denotes the input and output, respectively, i denotes the index of an output position, j denotes the index of all possible positions, $f(\cdot)$ denotes the response function of global influence on current position x_i, $g(\cdot)$ denotes a representation of the input signal at the position x_j, and $\mathcal{C}(\cdot)$ denotes the total influence for normalizing the response.

In our non-local block, we apply the Gaussian function to compute similarity in an embedding space. We set the response function $f(\cdot)$ as:

$$f(x_i, x_j) = e^{\theta(x_i)^T \phi(x_j)}, \tag{4}$$

where $\theta(x_i) = W_\theta \cdot x_i$, $\phi(x_j) = W_\phi \cdot x_j$. Similarly, $g(\cdot)$ could be set as $g(x_j) = W_g \cdot x_j$. Thus, the $\mathcal{C}(\cdot)$ could be set as $\mathcal{C}(x) = \sum_{\forall j} f(x_i, x_j)$. In this response function, the process of $(1/\mathcal{C}(x)) \sum_{\forall j} f(x_i, x_j)$ could be viewed as a softmax operation. In addition, we add y and x for a residual learning. In our architecture, the non-local block could be defined as:

$$D_0 = \mathcal{N}(E_4) + E_4. \tag{5}$$

After the non-local attention step, we feed the fused global feature into the decoding process as presented in equation (2) or Fig. 4. The non-local block could improve the performance of matching effectively. After this part, we could obtain the matching volume but in a low resolution $1/4H \times 1/4W \times 1/4D$. Thus, we should recover the scale to get the final matching volume.

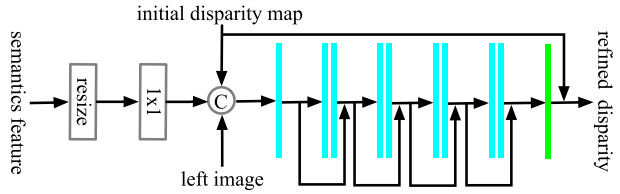

Fig. 5. Geometry refinement module (GR). The initial disparity map, the left image, and the semantics feature are fed to the GR module. After this module, we get refined disparity map. Here, blue block means the 32 convolutions with the size3 × 3, and green block means the 1 convolution with the size3 × 3.

(2) The scale recovery part

The scale recovery part contains one convolution and two de-convolutions for recovering the size of the input image. The output of our NLAM module is a final matching volume with size $D \times H \times W$ from the variance-based cost volume.

D) Disparity map regression by soft argmin

In this step, we will estimate the initial disparity map from the matching volume. Thus, we naturally embed our matching volume into a 3D to 2D process. The simplest way to recover the initial disparity map $\hat{d}$ from the matching volume M is the pixel-wise winner-take-all such as an argmax operation. However, this way is unable to predict sub-pixel estimation and less robust [5, 13]. Thus, we predict the disparity map by passing an argmin operation. First, we convert the matching volume M to the probability volume $\mathcal{P}$ via the softmax operation $\sigma(\cdot)$. Then, we take the sum of each disparity d weighted with its probability. The soft argmin process is defined as:

$$\hat{d} = \sum_{d=0}^{D_{\max}} d \times \mathcal{P}(d) = \sum_{d=0}^{D_{\max}} d \times \sigma(-M_d), \tag{6}$$

where $\mathcal{P}(d)$ denotes the probability estimation for all pixels of the image at disparity d. M_d denotes all value of the d-th layer in the matching volume M.

The above method could accurately approximate the disparity d in the range from 0 to $D_{\max}$. The output initial disparity map is the same size as the input image.

E) Disparity map refinement by geometry refinement module

The initial disparity map from the probability volume is a qualified output, but the boundaries may suffer from over-smoothing in the recovery size part of the NLAM module, or the completeness of the object suffers from the missing piece in the occlusion area. Notice that the input image contains complete boundary information, and the output of 2D GFL module contains the semantics feature, we thus use the input image and the semantics feature as guidance to refine the initial disparity map. Inspired by the recent multi-view stereo algorithm [5], we redesign a geometry refinement (GR) module at the end of NLCA-Net, as shown in Fig. 5.

Table 1. Evaluation of NLCA-Net with different settings.

Network setting							SceneFlow		KITTI 2015	
GFL	VBCM	NLAM	GR	$Loss_{L_1}$	$Loss_{L_1} + Loss_w$	Param	Mem.	D1-all (%)	Mem.	D1-all (%)
(Res-Net 50)				√		2.3 M	3.7 G	6.32	3.6 G	6.17
√				√		2.4 M	3.7 G	5.33	3.6 G	4.90
√	√	(3D-Conv)		√		2.8 M	5.6 G	3.80	5.5 G	3.63
√	√	√		√		5.3 M	7.6 G	3.09	7.6 G	2.14
√	√	√	√	√		5.4 M	7.6 G	3.06	7.6 G	2.08
√		√	√	√	√	5.4 M	7.9 G	2.98	7.9 G	2.00
√	√	√	√	√	√	5.4 M	7.6 G	2.87	7.6 G	1.96

Computed the percentage of three-pixel-error on the SceneFlow and KITTI 2015 test set.

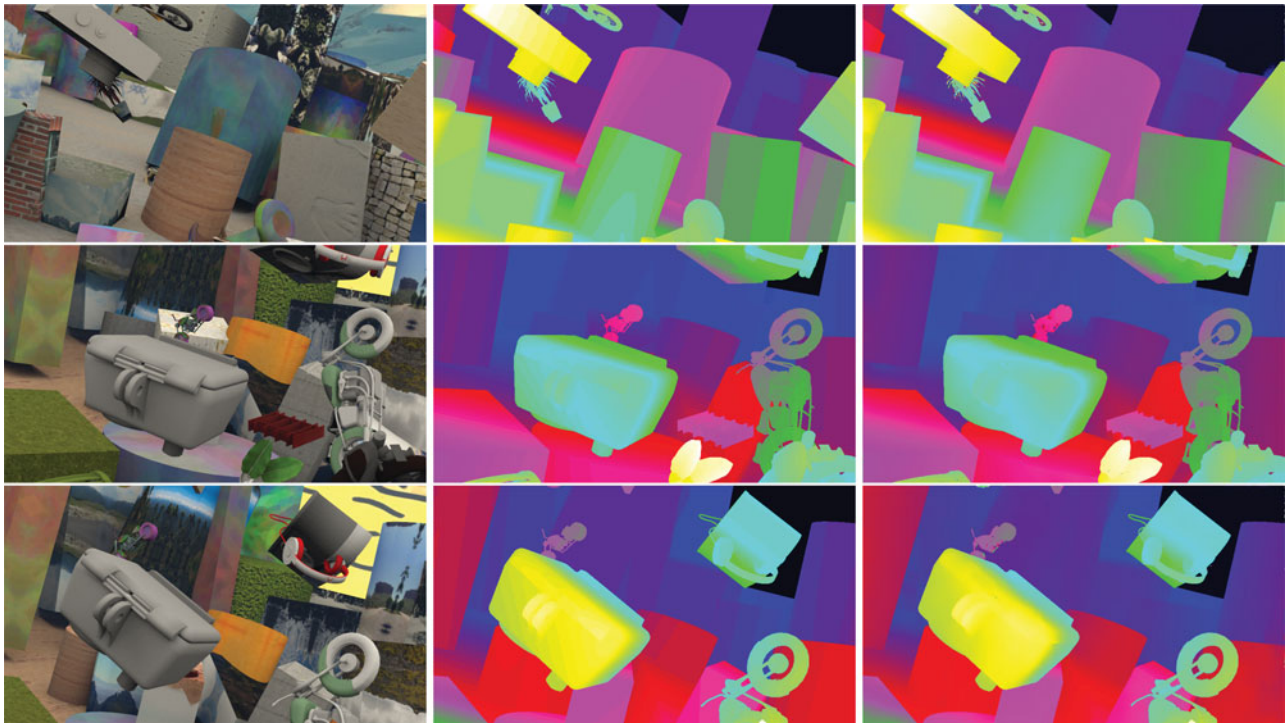

Fig. 6. SceneFlow test data qualitative results. From left: left stereo input image, ground-truth, disparity prediction.

In the GR module, the initial disparity map, the left image, and the semantics feature are concatenated as the input. Notice that the size of the semantics feature is only quarter in width and height dimension compared to input images. Therefore, we first upsample the semantics feature, then use the 1×1 convolutional filter to adjust it. Next, we concatenate them as a 38-feature input and send them to eight-layer residual network, which consists of the 32 convolutional filters with the size 3×3. After that, we apply one convolutional filter to obtain the difference, and the initial disparity map is added back to generate the refined disparity map. The last layer does not contain the BN layer and the ReLU layer. After the GR module, we could get the refined disparity map.

F) Loss function

The loss functions consider both the initial disparity map and the refined disparity map. We use the L_1 loss to evaluate the difference between the ground truth and the predicted disparity map. Due to the ground truth sparse disparity map, we only consider those pixels which are valid. The $Loss_{L_1}$ is defined as:

$$Loss_{L_1} = \frac{1}{N_1} \sum_{p \in P_{valid}} ||d(p) - \hat{d}_i(p)||_1 + ||d(p) - \hat{d}_r(p)||_1, \tag{7}$$

where P_{valid} denotes the set of valid ground truth pixels, N_1 denotes the total number of valid ground truth pixels, $d(p)$ denotes the ground truth of pixel p, $\hat{d}_i(p)$ denotes the initial disparity map value of pixel p, and $\hat{d}_r(p)$ denotes the refined disparity map value of pixel p.

To further improve the robustness and performance of the non-occluded area, we apply the warping loss to our loss function. The warping function is widely used in the traditional methods, and the previous learning method utilizes the warping function to structure the warping loss which

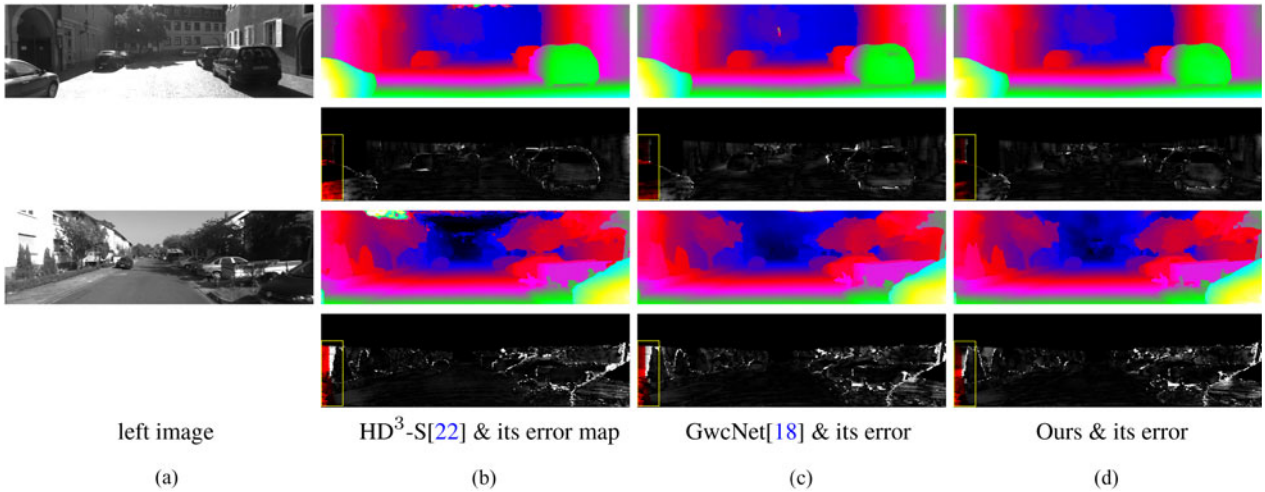

Fig. 7. KITTI 2012 test data qualitative results. We compare our approach with state-of-the-art methods (HD³-S and GwcNet), and we highlight our advantage in the error maps. Note that, in the error maps, the deeper red pixels mean higher error rate in the occluded regions and white pixels denote ≥5 pixels error in the non-occluded regions.

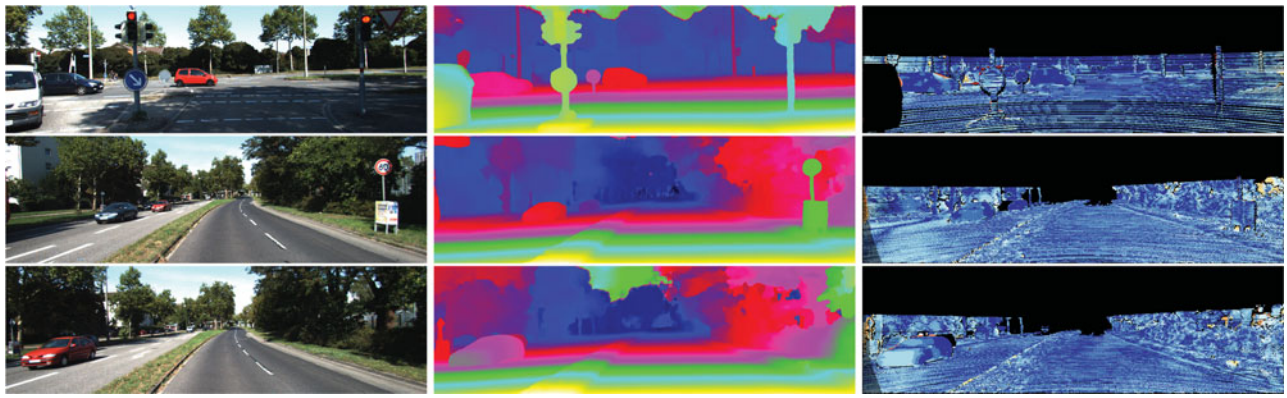

Fig. 8. KITTI 2015 test data qualitative results. From left: left stereo input image, disparity prediction, error map. Note that, in the error maps, depicting correct estimates (<3 px or <5% error) in blue and wrong estimates in red color tones.

Table 2. Influence of weight values for λ_1, λ_2, α, and β on three-pixel-error.

Network setting				D1-all	
λ_1	λ_2	α	β	SceneFlow (%)	KITTI 2015 (%)
0	0	1	0	3.43	2.20
0.5	0.5	0.5	0.5	3.57	2.24
0.5	0.5	0.6	0.4	3.60	2.28
0.5	0.5	0.7	0.3	3.54	2.16
0.5	0.5	0.8	0.2	3.37	2.13
0.5	0.5	0.9	0.1	3.29	2.21
0.6	0.4	0.8	0.2	3.13	2.19
0.7	0.3	0.8	0.2	3.05	2.08
0.8	0.2	0.8	0.2	2.99	1.99
0.85	**0.15**	**0.8**	**0.2**	2.87	1.96
0.9	0.1	0.8	0.2	2.95	2.02

We empirically found that 0.85/0.15/0.8/0.2 yielded the best performance on the SceneFlow test set.

Table 3. Influence of the different numbers of the non-local blocks on the model.

		SceneFlow		KITTI 2015	
Model	R	EPE	D1-all (%)	Out-Noc (%)	Out-ALL (%)
NLCA-Net	0	0.92	2.96	1.84	2.13
	1	0.87	2.87	1.79	1.96
	2	0.84	2.80	1.76	1.92
	3	0.82	2.76	1.74	1.90

Here R denotes the number of the non-local blocks.

gives the network self-supervised ability [21]. To solve the different illumination between left–right images, we introduce a structural similarity (SSIM) term $S(\cdot)$ [37] to improve the robustness. The $Loss_w$ is defined as:

$$Loss_w(I_L, I'_L) = \frac{1}{N_2} \sum_{p \in \mathbf{n}_{valid}} \lambda_1 \frac{1 - S(I_L, I'_L)}{2} + \lambda_2 |I_L - I'_L|, \tag{8}$$

where $\mathbf{n}_{valid}$ denotes the set of pixels in the non-occluded area, N_2 denotes the total number of valid pixels in the non-occluded area, I_L denotes the left image, I'_L denotes the warping image which is reconstructed from the right image

Table 4. Results on KITTI 2012 stereo benchmark.

Method	Error rates of 2 pixels		Error rates of 3 pixels		Error rates of 4 pixels		Error rates of 5 pixels		EPE	Runtime
	Out-Noc (%)	Out-All (%)	Out-Noc (%)	Out-All (%)	Out-Noc (%)	Out-All (%)	Out-Noc (%)	Out-All (%)	(px)	(s)
L-ResMatch [38]	3.64	5.06	2.27	3.40	1.76	2.67	1.50	2.26	0.7	48
MC-CNN-acrt [24]	3.90	5.45	2.37	3.63	1.90	2.85	1.64	2.39	0.7	67
GC-NET [13]	2.71	3.46	1.77	2.30	1.36	1.77	1.12	1.46	0.6	0.9
PSMNet [20]	2.44	3.01	1.49	1.89	1.12	1.42	0.90	1.15	0.5	0.41
SsSMnet [21]	3.34	4.24	2.30	3.00	1.82	2.39	1.53	2.01	0.7	0.8
SGM-Net [23]	3.60	5.15	2.29	3.50	1.82	2.39	1.60	2.36	0.7	67
SPS-St [30]	4.98	6.28	3.39	4.41	2.72	3.52	2.33	3.00	0.8	2
Displets v2 [31]	3.43	4.46	2.37	3.09	1.97	2.52	1.72	2.17	0.7	0.8
PBCP[39]	3.62	5,01	2.36	3.45	1.88	2.74	1.62	2.32	0.7	67
HD³-Stereo[22]	2.00	2.56	1.40	1.80	1.12	1.43	0.94	1.19	0.5	**0.14**
GwcNet[18]	2.16	2.71	1.32	1.70	0.99	1.27	0.80	1.03	0.5	0.32
NLCA-Net	2.01	2.55	1.25	1.62	0.94	1.22	0.76	0.98	0.4	0.43
NLCA-Net*	**1.97**	**2.51**	**1.22**	**1.59**	**0.92**	**1.20**	**0.75**	**0.97**	**0.4**	0.44

Average end-point-error (EPE) and the percentage of different pixel-error are used for evaluations on the KITTI 2012 test set. Compared with other algorithms, our approach achieves the best performance in most cases.
*The number of the non-local blocks is 3.

and the refined disparity map, and λ_1, λ_2 denote the balance between structural similarity and image appearance difference.

Our loss function for stereo matching is defined as:

$$Loss = \alpha Loss_{L_1} + \beta Loss_w, \tag{9}$$

where α denotes the weight of $Loss_{L_1}$, and β denotes the weight of $Loss_w$.

IV. EXPERIMENTS

In this section, we will evaluate the performance of our method on two widely used stereo datasets: SceneFlow and KITTI. First, we show our implementation details about the network setting and training method, as shown in Section A. Then, we compare the contribution of the different components in NLCA-Net, as shown in Section B. Finally, we quantize the performance of our method and compare it with the state-of-the-art methods on the KITTI 2012 and 2015, as shown in Section C.

A) Implementation details

In this section, we implement NLCA-Net by Tensorflow with 5.46 M trainable parameters, and the code will be released at the Github website.[1] To obtain the final model, we should choose the hyper-parameters, train, and evaluate the model.

(1) Hyper-parameters and datasets

For the hyper-parameters in this network, we set the max disparity $D = 192$ to ensure all possible disparity values in the image could be detected. In the loss function, we initially apply $\lambda_1 = 0$, $\lambda_2 = 0$, $\alpha = 1$, and $\beta = 0$ and then empirically test the best parameters based on our experiments.

[1]https://github.com/NPU-IAP/NLCA-Net

For the datasets, we will train and evaluate our approach on these stereo datasets as follows:

- **SceneFlow**: a large synthetic dataset consists of 35454 training and 4370 testing images with the size $H \times W = 540 \times 960$, which provides dense and clear disparity maps as ground truth. It could help us to adequately assess the performance of different model variants without worrying about over-fitting, and to make the pre-trained model have better generalization performance.
- **KITTI 2012**: a challenging and varied road scene dataset contains 194 training and 195 testing images with the size $H \times W = 376 \times 1236$, which only provides sparse disparity maps as ground truth for training images.
- **KITTI 2015**: a real-world street views dataset contains 200 training and 200 testing images with the size $H \times W = 374 \times 1236$, which only provides sparse disparity maps as ground truth for training images.

(2) Training method

For the training process, our network can be trained from random initialization in an end-to-end way with the supervision of stereo pairs and optimized using Adam Optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a batch size of 1 for each GPUs. Before training, we normalize stereo pairs with pixel intensities level ranging from 0 to1, and randomly crop them into 256×512. During the training process, we adopt the multi-step training method with four Nvidia 1080Ti GPUs. Thus, the training process consists of two parts: the pre-training process and the fine-tuning process.

- In the pre-training process on the SceneFlow dataset, the learning rate is initially set to 1×10^{-3} for 30 epochs and obtains the pre-train model.
- In the fine-tuning process on the KITTI dataset, the learning rate is set to 1×10^{-3} for 800 epochs and then reduced to 1×10^{-4} for the other 100 epochs. After the training process, we get the final model.

(3) Evaluating metric

For evaluating our model and comparing with the state-of-the-art methods published recently, we show our results' errors with the following metrics, which have been widely used in the website of KITTI dataset:

$$MAE = \frac{1}{|N|}\sum_{i\in T}|\hat{d}-d|,$$

$$D_{1-bg} = \frac{\sum_{i\in bg}[|\hat{d}-d| > m \wedge i = vaild]}{\sum_{i\in bg}[i = vaild]},$$

$$D_{1-fg} = \frac{\sum_{i\in fg}[|\hat{d}-d| > m \wedge i = vaild]}{\sum_{i\in fg}[d = vaild]},$$

$$D_{1-all} = \frac{\sum_{i\in T}[|\hat{d}-d| > m \wedge i = vaild]}{\sum_{i\in T}[d = vaild]},$$

where d denotes the ground truth of disparity, $\hat{d}$ denotes the estimated depth, m denotes the threshold of error pixel, $[\cdot]$ denotes the Iverson bracket, bg represents the set of all points in background regions of the images, fg represents the set of all points in foreground regions of the images, and T represents the set of all points in the images.

B) Ablation study for NLCA-Net

To verify our design's effectiveness, we conduct experiments with different settings to evaluate NLCA-Net on the SceneFlow dataset. We train different model variants like the pre-training process, as presented in Section A. We first compare the performance of different settings, including the 2D GFL module, VBCM, the NLAM module, and the GR module, as shown in Table 1. Then, we present the representative results of our model and ablation study of loss weight, as shown in Fig. 6 and Table 2. Moreover, we test the impact of the different numbers of the non-local on the model, as shown in Table 3.

As shown in Fig. 6 and Table 1, it qualitatively demonstrates the benefits of using the modules which we proposed. First, the 2D GFL and NLAM modules show a significant performance improvement for the matching accuracy. For the GFL module, it enhances the scene understanding ability of the model effectively; for the NLAM module, it has a strong regularization ability to learning the matching rules and can facilitate the learning process. Second, the GR module and $Loss_w$ function improve the matching accuracy a little. On a good baseline, the GR module could further improve the performance and $Loss_w$ function could achieve 1% improvement on whole pixels and 4% on non-occluded pixels. Finally, the VBCM provides better testing

Table 5. Results on KITTI 2015 stereo benchmark.

Method	All pixels			Non-occluded pixels		
	D1-bg (%)	D1-fg (%)	D1-all (%)	D1-bg (%)	D1-fg (%)	D1-all (%)
L-ResMatch [38]	2.72	6.95	3.42	2.35	5.76	2.91
MC-CNN-acrt [24]	2.89	8.88	3.89	2.48	7.64	3.33
GC-NET [13]	2.21	6.16	2.87	2.02	5.58	2.61
PSMNet [20]	1.86	4.62	2.32	1.71	4.31	2.14
SsSMnet [21]	2.70	6.92	3.40	2.46	6.13	3.06
SGM-Net [23]	2.66	8.64	3.66	2.23	7.44	3.09
SPS-St [30]	3.84	12.67	5.31	3.50	11.61	4.84
Displets v2 [31]	3.00	5.56	3.43	3.43	4.46	3.09
PBCP [39]	2.58	8.78	3.61	2.27	7.71	3.17
HD^3-Stereo [22]	1.70	**3.63**	2.02	1.56	**3.43**	1.87
GwcNet [18]	1.74	3.93	2.11	1.61	3.49	1.92
NLCA-Net	1.53	4.09	1.96	1.39	3.80	1.79
NLCA-Net*	**1.52**	3.79	**1.90**	**1.39**	3.55	**1.74**

Our approach achieves comparable performance to state-of-the-art methods.
*The number of the non-local blocks is 3.

Table 6. Comparisons of different state-of-the-art methods in the reflective regions.

Method	Error rates of reflective regions				EPE (px)
	2 pixels (%)	3 pixels (%)	4 pixels (%)	5 pixels (%)	
MC-CNN-acrt [24]	27.58	20.70	17.17	14.89	4.1
GC-NET [13]	19.07	12.80	9.77	7.99	2.0
PSMNet [20]	16.06	10.18	7.29	5.64	1.4
SsSMnet [21]	22.98	16.59	13.21	11.08	3.6
SGM-Net [23]	25.70	18.97	15.62	13.55	3.8
SPS-St [30]	24.35	18.00	14.88	13.07	3.6
Displets v2 [31]	16.25	10.41	8.02	6.61	2.2
GA-Net [40]	15.63	9.85	7.10	5.54	1.5
GwcNet [18]	14.57	9.28	6.70	5.22	1.4
NLCA-Net	14.49	9.00	6.43	4.88	**1.4**
NLCA-Net*	**14.11**	**8.78**	**6.19**	**4.68**	1.6

Average end-point-error (EPE) and the percentage of different pixel-error are used for evaluations on the KITTI 2012 test set. Compared with other algorithms, our approach achieves the best performance in most cases.
*The number of the non-local blocks is 3.

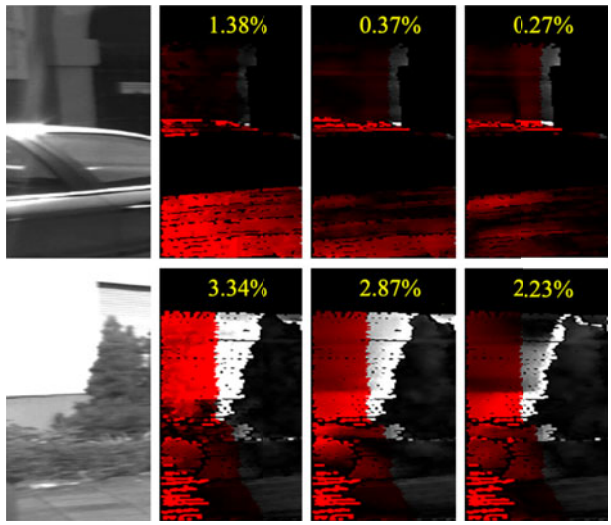

Fig. 9. Part zoom-up of the error maps on the occluded region. From left: original image, HD^3-Stereo, GwcNet, and ours. The result shows that our method can notably reduce the error rate on the occluded area and handle well with the large textureless regions.

accuracy compared with traditional concatenation way, and it saves half-memory of the cost volume, which reduces about 300 M running memory of the whole framework.

As shown in Table 2, we conduct experiments with various combinations of loss weights which have the relationship about $\lambda_1 + \lambda_2 = 1$ and $\alpha + \beta = 1$. For the baseline, we only use the L_1 loss function and then continuously adjust the loss weights to find the weights which yield the best performance. The result shows that the weight settings of 0.85 for λ_1, 0.15 for λ_2, 0.8 for α, and 0.2 for β obtain the best performance, which is a 2.87% three-pixel-error rate on the SceneFlow test set and a 1.96% three-pixel-error rate on the KITTI test set.

Finally, we also test our model with different non-local block (NL-block) numbers. In this part, we set a model without NL-block as the baseline. The impact of NL-block is tested by increasing the NL-block number in the model. As shown in Table 3, we list the results with different NL-blocks on the SceneFlow and KITTI 2015, indicating that performance can be improved by introducing NL-block into the model. It is also noticed that the improvement decays gradually as the number of NLB increases, showing the boundary effect.

C) KITTI 2012 and 2015 benchmark results

To evaluate the performance of our model, we compare the performance of NLCA-Net with other state-of-the-art methods on the KITTI dataset. We use the multi-step training method to train the model, as shown in Section A. We present the representative images of our model and other state-of-the-art methods, as shown in Figs 7 and 8. Then, we evaluate the performance of our model on the KITTI website,[2] as shown in Tables 4 and 5. Besides, we compare our model with other competing algorithms in the reflective regions, as presented in Table 6.

As shown in Figs 7 and 8, the proposed method could produce dense and clear disparity maps. For the non-occluded region, the proposed method shows the powerful performance; the disparity maps predicted by NLCA-Net are sharper and more complete than other learning methods. Even if in the occluded region, the proposed method still provides high-level performance and significantly outperforms other state-of-the-art methods, as shown in Fig. 9. The results show that our method can notably reduce the error rate on the occluded area and handle well with the large textureless regions. It also means the semantic information plays an important role, which offers the boundary information to perfect the edge pixels of objects. Overall, the proposed method shows the incredible expressiveness in the matching task.

As shown in Tables 4 and 5, it qualitatively demonstrates the performance of the proposed method. Compared with other methods, the proposed method yields more precise and robust disparity maps, particularly in the non-occluded regions. For the KITTI 2012 dataset, our approach is very close to HD^3-Stereo [22] on non-occluded (Out-Noc) in the error rate of 2 pixels; but for other quality indexes, our method all achieves the best performance as shown in the table. For the KITTI 2015 dataset, our approach shows comparable performance and markedly outperforms other competing algorithms, including the previous best result (HD^3-Stereo [22] and GwcNet [18]). In short, the proposed approach is superior to state-of-the-art methods in most cases.

As shown in Table 6, it qualitatively shows the advantage of our model in the reflective regions. In our designs, we use attention mechanisms and semantic information to enhance the ability of scene understanding. Thus, in the reflective region, the proposed method achieves the best performance of all quality indexes and significantly outperforms other state-of-the-art methods. It indicates the effectiveness of our matching network based on the contextual attention mechanism, which exhibits robustness to the reflective regions.

[2] http://www.cvlibs.net/datasets/kitti/eval_stereo.php

V. CONCLUSIONS

In this work, we present a highly efficient network architecture for stereo matching. The proposed model can exploit the global context information to achieve superior performance in the matching task. The NLAM significantly enhances the ability of scene understanding to improve the accuracy in the challenging regions, such as occlusions and large textureless/reflective areas. The variance-based cost volume can provide the similarity information and reduces some memory, thus further improving the performance. The experiment shows that the proposed method can improve the performance in challenging regions and outperform state-of-the-art methods in most cases. For future work, we are interested in exploring the generative adversarial network's potential to achieve higher accurate semi-supervised or unsupervised stereo matching.

ACKNOWLEDGEMENT

This work was supported in part by the Natural Science Foundation of China (61671387, 61420106007 and 61871325). The authors would like to thank the Editor-in-Chief, associate editor and the anonymous reviewers for their valuable comments and suggestions that helped improve the quality of this paper.

REFERENCES

[1] Chenyi, C.; Ari, S.; Alain, K.; Jianxiong, X.: Deepdriving: learning affordance for direct perception in autonomous driving, in *IEEE Int. Conf. on Computer Vision* (*ICCV*), 2015, 2722–2730

[2] Korbinian, S.; Teodor, T.; Felix, R.; Heiko, H.; Michael, S.: Stereo vision based indoor/outdoor navigation for flying robots, in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2013, 3955–3962

[3] Chen, X.; *et al.* 3d object proposals for accurate object class detection, in *The Int. Conf. on Neural Information Processing Systems* (*NIPS*), 2015, 424–432

[4] Zhang, C.; Li, Z.; Cheng, Y.; Cai, R.; Chao, H.; Rui, Y.: Meshstereo: a global stereo model with mesh alignment regularization for view interpolation, in *IEEE Int. Conf. on Computer Vision* (*ICCV*), 2015, 2057–2065

[5] Yao, Y.; Luo, Z.; Li, S.; Fang, T.; Quan, L.: Mvsnet: depth inference for unstructured multi-view stereo, in *The European Conf. on Computer Vision* (*ECCV*), 2018, 767–783

[6] Li, B.; Shen, C.; Dai, Y.; Hengel, A.V.D.; He, M.: Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs, in *IEEE Int. Conf. on Computer Vision and Pattern Recognition* (*CVPR*), 2015, 1119–1127

[7] Li, B.; Dai, Y.; He, M.: Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference. *Pattern Recognit.*, **83** (2018), 328–339.

[8] Hamzah, R.A.; Ibrahim, H.; Abu Hassan, A.H.: Stereo matching algorithm based on per pixel difference adjustment, iterative guided filter and graph segmentation. *J. Vis. Commun. Image. Represent.*, **42** (2017), 145–160.

[9] Hirschmüller, H.: Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.*, **30** (2) (2008), 328–341.

[10] Yang, Q.: A non-local cost aggregation method for stereo matching, in *IEEE Int. Conf. on Computer Vision and Pattern Recognition* (*CVPR*), 2012, 1402–1409

[11] Woodford, O.; Torr, P.; Reid, I.; Fitzgibbon, A.: Global stereo reconstruction under second-order smoothness priors. *IEEE Trans. Pattern Anal. Mach. Intell.*, **31** (12) (2009), 2115–2128.

[12] Laina, I.; Rupprecht, C.; Belagiannis, V.; Tombari, F.; Navab, N.: Deeper depth prediction with fully convolutional residual networks, in *2016 Fourth Int. Conf. on 3D Vision (3DV)*, 2017, 239–248

[13] Kendall, A.; Martirosyan, H.; Dasgupta, S.; Henry, P.: End-to-end learning of geometry and context for deep stereo regression, in *IEEE Int. Conf. on Computer Vision* (*ICCV*), 2017, 66–75

[14] Mayer, N.; *et al.* A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation, in *IEEE Int. Conf. on Computer Vision and Pattern Recognition* (*CVPR*), 2016, 4040–4048

[15] Andreas, G.; Philip, L.; Raquel, U.: Are we ready for autonomous driving? the kitti vision benchmark suite, in *IEEE Int. Conf. on Computer Vision and Pattern Recognition* (*CVPR*), 2012, 3354–3361

[16] Menze, M.; Geiger, A.: Object scene flow for autonomous vehicles, in *IEEE Int. Conf. on Computer Vision and Pattern Recognition* (*CVPR*), 2015, 3061–3070

[17] Zbontar, J.; LeCun, Y.: Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.*, **17** (1–32) (2016), 2.

[18] Guo, X.; Yang, K.; Yang, W.; Wang, X.; Li, H.: Group-wise correlation stereo network, in *IEEE Int. Conf. on Computer Vision and Pattern Recognition* (*CVPR*), 2019, 3273–3282

[19] Zagoruyko, S.; Komodakis, N.: Learning to compare image patches via convolutional neural networks, in *IEEE Int. Conf. on Computer Vision and Pattern Recognition* (*CVPR*), 2015, 4353–4361

[20] Chang, J.-R.; Chen, Y.-S.: Pyramid stereo matching network, in *IEEE Int. Conf. on Computer Vision and Pattern Recognition* (*CVPR*), 2018, 5410–5418

[21] Zhong, Y.; Dai, Y.; Li, H.: Self-supervised learning for stereo matching with self-improving ability, in *arXiv preprint*, 2017

[22] Yin, Z.; Darrell, T.; Yu, F.: Hierarchical discrete distribution decomposition for match density estimation, in *IEEE Int. Conf. on Computer Vision and Pattern Recognition* (*CVPR*), 2019, 6044–6053

[23] Seki, A.; Pollefeys, M.: SGM-Nets: semi-global matching with neural networks, in *IEEE Int. Conf. on Computer Vision and Pattern Recognition* (*CVPR*), 2017, 6640–6649

[24] Žbontar, J.; Le Cun, Y.: Computing the stereo matching cost with a convolutional neural network, in *IEEE Int. Conf. on Computer Vision and Pattern Recognition* (*CVPR*), 2015, 1592–1599

[25] Wang, X.; Girshick, R.; Gupta, A.; He, K.: Non-local neural networks, in *IEEE Int. Conf. on Computer Vision and Pattern Recognition* (*CVPR*), 2018, 7794–7803

[26] Scharstein, D.; Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vision*, **47** (1–3) (2002), 7–42.

[27] Tombari, F.; Mattoccia, S.; Stefano, L.D.; Addimanda, E.: Classification and evaluation of cost aggregation methods for stereo correspondence, in *IEEE Int. Conf. on Computer Vision and Pattern Recognition* (*CVPR*), 2008, 1–8

[28] Hamzah, R.A.; Ibrahim, H.: Literature survey on stereo vision disparity map algorithms. *J. Sens.*, **2016** (2016), 1–23.

[29] Bleyer, M.; Rhemann, C.; Rother, C.: PatchMatch stereo – stereo matching with slanted support windows, in *British Machine Vision Conf.*, 2011, 14.1–14.11

[30] Yamaguchi, K.; McAllester, D.; Urtasun, R.: Efficient joint segmentation, occlusion labeling, stereo and flow estimation, in *European Conf. on Computer Vision* (*ECCV*), 2014, 756–771

[31] Guney, F.; Geiger, A.: Displets: resolving stereo ambiguities using object knowledge, in *IEEE Int. Conf. on Computer Vision and Pattern Recognition* (*CVPR*), 2015, 4165–4175

[32] Gidaris, S.; Komodakis, N.: Detect, replace, refine: deep structured prediction for pixel wise labeling, in *IEEE Int. Conf. on Computer Vision and Pattern Recognition* (*CVPR*), 2017, 7187–7196

[33] Long, J.; Shelhamer, E.; Darrell, T.: Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **39** (4) (2015), 640–651.

[34] Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H.: Rethinking atrous convolution for semantic image segmentation, in *arXiv preprint*, 2017

[35] Ranjan, A.; Black, M.J.: Optical flow estimation using a spatial pyramid network, in *IEEE Int. Conf. on Computer Vision and Pattern Recognition* (*CVPR*), 2017, 2720–2729

[36] Zhang, K.; Fang, Y.; Min, D.; Sun, L.; Yan, S.Y.S.; Tian, Q.: Cross-scale cost aggregation for stereo matching. *IEEE Trans. Circuits Syst. Video Technol.*, **27** (5) (2014), 965–976.

[37] Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. *et al.*: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, **13** (4) (2004), 600–612.

[38] Shaked, A.; Wolf, L.: Improved stereo matching with constant highway networks and reflective confidence learning, in *IEEE Int. Conf. on Computer Vision and Pattern Recognition* (*CVPR*), 2017, 6901–6910

[39] Seki, A.; Pollefeys, M.: Patch based confidence prediction for dense disparity map, in *British Machine Vision Conf.*, 2016, 23.1–23.13

[40] Zhang, F.; Prisacariu, V.; Yang, R.; Torr, P.H.: Ga-net: guided aggregation net for end-to-end stereo matching, in *IEEE Int. Conf. on Computer Vision and Pattern Recognition* (*CVPR*), 2019, 185–194

APPENDIX A

Detailed network structure. The core architecture of our NLCA-Net contains three modules: (1) geometry feature learning module (GFL); (2) non-local attention matching module (NLAM); (3) geometry refinement module (GR). We illustrate the detailed structure of our method is presented in Table 7. Each 2D or 3D convolutional layer contains three steps: convolution, batch normalization (BN), and ReLU non-linearity (unless otherwise specified).

Zhibo Rao received his M.S. degree in electronic information engineering from Nanchang Hangkong University in 2017. He is currently a Ph.D. student in the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China. His research interests are pattern recognition, image processing, and deep learning. He has published some papers on the ICIP, APSIPA, 3DV, ICIEA, etc.

Table 7. The summary of our non-local context attention network, NLCA-Net.

	Layer description (k, s, f)	Output dimension
No.	Input images	$H \times W$
Geometry feature learning module (GFL)		
convo_1	$3 \times 3, 2, 32$	$1/2H \times 1/2W \times F$
convo_2	$[3 \times 3, 1, 32] \times 3$	$1/2H \times 1/2W \times F$
convo_3	$\begin{bmatrix} 3 \times 3, 1, 32 \\ 3 \times 3, 1, 32 \end{bmatrix} \times 3$	$1/2H \times 1/2W \times F$
convo_4	$3 \times 3, 2, 64$	$1/4H \times 1/4W \times 2F$
	$\begin{bmatrix} 3 \times 3, 1, 64 \\ 3 \times 3, 1, 64 \end{bmatrix} \times 15$	
convo_5	$3 \times 3, 1, 128$	$1/4H \times 1/4W \times 4F$
	$\begin{bmatrix} 3 \times 3, 128, dila = 2 \\ 3 \times 3, 128, dila = 2 \end{bmatrix}$	
convo_6	$3 \times 3, 1, 128$	$1/4H \times 1/4W \times 4F$
	$\begin{bmatrix} 3 \times 3, 128, dila = 4 \\ 3 \times 3, 128, dila = 4 \end{bmatrix}$	
SPP	$\begin{bmatrix} ave\_pooling : 64, 1 \times 1, 1, 32 \\ ave\_pooling : 32, 1 \times 1, 1, 32 \\ ave\_pooling : 16, 1 \times 1, 1, 32 \\ ave\_pooling : 8, 1 \times 1, 1, 32 \end{bmatrix}$	$1/4H \times 1/4W \times 4F$
	bilinear interpolation and concatenation	
	concat convo_4, convo_6, and SPP	$1/4H \times 1/4W \times 10F$
convo_7	$3 \times 3, 1, 128$	$1/4H \times 1/4W \times F$
	$3 \times 3, 1, 32$ (without BN and ReLU)	
	construct cost volume via variance-based cost metric	$1/4D \times 1/4H \times 1/4W \times F$
Non-local attention matching module (NLAM)		
conv3_0	$3 \times 3 \times 3, 1, 32$	$1/4D \times 1/4H \times 1/4W \times F$
	$\begin{bmatrix} 3 \times 3 \times 3, 1, 32 \\ 3 \times 3 \times 3, 1, 32 \end{bmatrix}$	
conv3_1 – conv3_4	$3 \times 3 \times 3, 2, (i+1) \times 32$	$\frac{1}{2^{(i+2)}}D \times \frac{1}{2^{(i+2)}}H \times \frac{1}{2^{(i+2)}}W \times (i+1)F$
	$\begin{bmatrix} 3 \times 3 \times 3, 1, (i+1) \times 32 \\ 3 \times 3 \times 3, 1, (i+1) \times 32 \end{bmatrix}, i = 1, 2, 3, 4$	
conv3_5	non-local block	$1/64D \times 1/64H \times 1/64W \times 4F$
deconv3_3 – deconv3_0	deconv: $3 \times 3 \times 3, 2, (i+1) \times 32$	$\frac{1}{2^{(i+2)}}D \times \frac{1}{2^{(i+2)}}H \times \frac{1}{2^{(i+2)}}W \times (i+1)F$
	$\begin{bmatrix} 3 \times 3 \times 3, 1, (i+1) \times 32 \\ 3 \times 3 \times 3, 1, (i+1) \times 32 \end{bmatrix}, \quad i = 3, 2, 1, 0$	
	add conv3_i	
deconv3_4	deconv: $3 \times 3 \times 3, 2, 16$	$1/2D \times 1/2H \times 1/2W \times 1/4F$
conv3_6	$3 \times 3 \times 3, 1, 8$	$1/2D \times 1/2H \times 1/2W \times 1/4F$
deconv3_4	deconv: $3 \times 3 \times 3, 2, 1$ (without ReLU and BN)	$D \times H \times W \times 1$
initial disparity map	Soft argmin	$H \times W$
Geometry refinement module (GR)		
conv4_0	resize convo_7 to $H \times W \times F$ via bilinear interpolation	$H \times W \times F$
	and apply convolutional layer with $1 \times 1, 1, 32$ to fine tuning feature	
	concat left image, disparity map, and conv4_0	$H \times W \times (F+4)$
conv4_1	$3 \times 3, 1, 32$	$H \times W \times F$
conv4_2	$\begin{bmatrix} 3 \times 3, 1, 32 \\ 3 \times 3, 1, 32 \end{bmatrix} \times 4$	$H \times W \times F$
conv4_5	$3 \times 3, 1, 1$ (without BN and ReLU)	$H \times W$
refined disparity map	add conv4_5 and initial disparity map	$H \times W$

Mingyi He received his B.Eng. and M.S. degrees in electronic engineering and signal processing from Northwestern Polytechnical University (NPU), Xian, China, in 1982 and 1985, respectively, and the Ph.D. degree in signal and information processing from Xidian University, Xian, China, in 1994. Since 1985, he has been with the School of Electronics and Information, NPU, where he has been a full professor since 1996. He is the Founder and Director of Shaanxi Key Laboratory (2003–2019) and International Research Center (2016–) for Information Acquisition and Processing, and the Director and Chief Scientist of the Center for Earth Observation Research (2011–), NPU. He had been a visiting scholar at Adelaide University, Adelaide, SA, Australia, and visiting professor at Sydney University, Sydney, NSW, Australia, and Adelaide University. His research interests are advanced machine vision and intelligent processing, including signal and image processing, computer vision, hyperspectral remote sensing, 3D information acquisition and processing, neural network, and deep learning. Dr. Mingyi He is a recipient of 11 national and provincial scientific prizes and two teaching achievement prizes in China. He won the Best Paper Award in IEEE CVPR 2012, the Best Deep/Machine Learning Paper Prize at APSIPA ASC 2017, the 2017 DICTA Best Student Paper Award (as supervisor and coauthor), etc. He was also a recipient of the government lifelong subsidy from the State Council of China in 1993 and the Baosteel Outstanding Teacher Award in 2017. He received awards or certificates of honor from IEEE Signal Processing Society in 2014, APSIPA in 2019, Chinese Institute of Electronics in 2019 and 2020. He has acted as a General Chair or TPC (Co)Chair, and Area Chair for dozens of national and international conferences. He has been a member of the Advisory Committee of National Council for Higher Education on Electronics and Information in China, a member of Chinese Lunar Exploration Expert Group, the Vice-President of Shaanxi Institute of Electronics (SIP committee chair), and the Vice-Director of the Spectral Imaging Earth Observation Committee of China Committee of International Society of Digital Earth. He is an Associate Editor for the IEEE Transactions on Geoscience and Remote Sensing and APSIPA Transactions on Signal and Information Processing, a Guest Editor for the IEEE Journal of Selected Topics in Applied Earth Observation and Remote Sensing. Dr. He is a senior member of IEEE and a member of IEEE GRSS Image and Data Fusion Technical Committee, Chair of SIPTM (Signal and Information Processing Theory and Methods) Committee and BoG member of the Asia-Pacific Signal and Information Processing Association (APSIPA).

Yuchao Dai is currently a Professor with the School of Electronics and Information at the Northwestern Polytechnical University (NPU). He received the B.E. degree, M.E degree, and Ph.D. degree all in signal and information processing from Northwestern Polytechnical University, Xian, China, in 2005, 2008, and 2012, respectively. He was an ARC DECRA Fellow with the Research School of Engineering at the Australian National University, Canberra, Australia from 2014 to 2017 and a Research Fellow with the Research School of Computer Science at the Australian National University, Canberra, Australia from 2012 to 2014. His research interests include structure from motion, multi-view geometry, low-level computer vision, deep learning, compressive sensing, and optimization. He won the Best Paper Award in IEEE CVPR 2012, the DSTO Best Fundamental Contribution to Image Processing Paper Prize at DICTA 2014, the Best Algorithm Prize in NRSFM Challenge at CVPR 2017, the Best Student Paper Prize at DICTA 2017, and the Best Deep/Machine Learning Paper Prize at APSIPA ASC 2017.

Zhidong Zhu is currently a master student in the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China.

Bo Li is currently a research associate in the School of Electronics and Information, Northwestern Polytechnical University (NPU), China. He received his B.E. degree and Ph.D. degree from NPU in 2011 and 2018, respectively. During 2013–2015, he was a visiting student at the University of Adelaide, Australia. His research interests mainly focus on the deep learning and computer vision. He has published more than 10 papers in the CVPR, PR, TMM, TGRS, etc.

Renjie He received the B.E. degree in automation from Xi'an Jiaotong University in 2008, and M.E. degree and Ph.D. degree in signal and information processing from Northwestern Polytechnical University in 2011 and 2017, respectively. During 2011–2013, He was a visiting student at the University of Sydney, Australia. During 2017–2020, He has been a Research Fellow at Nanyang Technological University, Singapore. His research interests include image enhancement, restoration, visual perception, and remote sensing image analysis. He won the Best Paper Award in IEEE ICIEA 2019.