MPA-Net: multi-path attention stereo matching network

Haiwei Sang¹,², Zuliu Yang³, Xiaowei Yang⁴, Yong Zhao³

¹School of Mathematics and Big Data, Guizhou Education University, Guiyang 550018, People’s Republic of China
²College of Computer Science and Technology, Guizhou University, Guiyang 550025, People’s Republic of China
³School of Electronic and Computer Engineering, Shenzhen Graduate School of Peking University, Shenzhen 518055, People’s Republic of China
⁴School of Mechanical Engineering, Guizhou University, Guiyang 550025, People’s Republic of China

Abstract: A novel learning-based end-to-end network for stereo matching, named Multi-path Attention Stereo Matching (MPA-Net), is introduced in this study. Different from existing methods, the multi-path attention aggregation module is designed firstly, named MPA, which is a unified structure using three parallel layers with a respective attention mechanism to extract the multi-scale informational features. Secondly, the method of cost volume construction, which differs from the traditional stereo matching methods, is extended. And then, the absolute difference between two input features is calculated. Furthermore, a u-shaped structure with 3D attention gate is selected as the encoder-decoder module. Specifically, the module is used to fuse the absolute difference (SAD), sum of squared intensity differences (SSD) [10] and normalised cross-correlation (NCC) [11]. Cost volume construction, which differs from the traditional stereo matching methods, is extended. And then, the absolute difference between two input features is calculated. Furthermore, a u-shaped structure with 3D attention gate is selected as the encoder-decoder module. Specifically, the module is used to fuse the absolute difference (SAD), sum of squared intensity differences (SSD) [10] and normalised cross-correlation (NCC) [11].

1 Introduction

Depth estimation has been widely used in many computer vision tasks [1–5]. A typical method is to employ the rectified stereo images for depth estimation by first acquiring the disparity map. Generally, the stereo matching on the stereo images is regarded as a 1D matching problem. For the reference pixel (x, y), if the corresponding pixel at (x − d, y) can be found in the target image, the depth of the pixel can be calculated by using f/B/d, where f is the focal length of camera, B is the distance between two camera centres and d is the disparity. The goal of stereo matching is to get an accurate disparity map from two stereo images.

Traditionally, stereo matching algorithms [6–8] follow four steps [9]: matching cost calculations, matching cost aggregation, disparity calculation and disparity refinement. It can be regarded as a problem with several stages for optimisation and each step in the pipeline is important. The goal of matching cost calculations is to find a reference pixel and its possible corresponding pixel in the candidate image, which is both initial and crucial step of stereo matching. Some algorithms have been proposed, including sum of absolute difference (SAD), sum of squared intensity differences (SSD) [10] and normalised cross-correlation (NCC) [11]. Cost aggregation refers to the spatial aggregation of initial matching costs over a support region around the pixel of interest. In the disparity calculation step, we need to define and select the optimal disparity with the lowest cost. Local approaches usually choose the winner-take all (WTA) optimisation [9]. Global optimisation approaches optimise all disparities for the entire images simultaneously and collectively by defining a global energy function. Common approaches are Markov Random fields (MRFs) [12]. Dynamic programming is generally adopted in semi-global matching [13] to optimise a global energy function. More recently, some contributions which integrate aggregated algorithms into the global energy function model have sprung up [14–17]. For more accurate modelling, instead of assigning a discrete disparity to each pixel, 3D cost-aggregation methods assign three continuous values to each pixel, representing the disparity and the surface normal direction simultaneously [18–20]. Left–right consistency check is usually used to further refine the disparity map. However, the performance of traditional methods is limited.

With convolutional neural network (CNN) and supervised learning method, stereo matching performs a state-of-the-art performance compared with traditional methods both in accuracy and speed. Although the learning-based method can achieve better disparity estimation, the results are often ambiguous in ill-regions, such as occlusions, reflections, textless regions etc. Some studies attended to relief the influences caused by the above problems, so they introduced extra information and effective structure to further improve the disparity estimation. EdgeStereo [21] and SegStereo [22] designed a multi-task learning network by using edge and semantic information, respectively, to improve the disparity prediction. GC-Net [23] formulated the network with stereo geometry, which first formed a 4D cost volume and then used a 3D encoder–decoder structure to regularise the cost volume. PSM-Net [24] employed the spatial pyramid pooling (SPP) and the stacked 3D CNN to further improve the utilisation of global context information.

Despite the above success, we argue that the results of the disparity map still have potential to be enhanced with some new designs and other modules. In this paper, the footsteps of existing methods are followed to further improve the performance of stereo network. Different from [22–24], we first design a multi-path attention aggregation module in the feature extraction section. More specifically, three paths of gradual down-sampling convolution layer are used in the module and each path aims to extract the features at a certain scale. In addition, attention mechanism [25] is introduced at the end of each path to increase representation power, and guide the network to focus on more informational features and suppress the unnecessary ones. Moreover, inspired from [9], the absolute difference of left-right features is calculated, so that it can be concatenated with the general used 4D cost volume [23, 24], to provide a better similarity understanding for the next cost regularisation section. Finally, the u-shaped structure is used as our encoder–decoder blocks for...
disparity prediction. And the 3D attention gate is applied to connect the decoding features to their corresponding decoding features by using skip-connection. By stacking the three encoder–decoder structures together with long-range skip connection, we acquire the final structure for cost regularisation. The main contributions of this paper can be summarised as follows:

- A carefully designed features extraction module, named MPA, is proposed. A multi-path attention aggregation module is designed to capture deep contextual features at multi-scale and achieves better feature representation compared with the general SPP module.
- The 4D cost volume with the extracted features by using disparity-level concatenation and disparity-level subtraction is used. Furthermore, a u-shaped encoder–decoder with attention gate is introduced to regularise the cost volume, leading to better disparity map prediction.
- The performance of the proposed network is verified on three different data sets and achieves favourable disparity maps. The results show our network can achieve favourable disparity maps against existing state-of-the-art networks.

2 Related work

2.1 End-to-end stereo network

Through careful design and supervision of the neural network, a fine disparity map can be end-to-end learned with stereo images. The first end-to-end network for disparity map estimation is DispNet [26], which proposed a correlation module first to encode the similarity of initial features, and then decode it directly to infer the disparity map regression by using deconvolution. Based on DispNet, Pang et al. [27] proposed a cascade CNN architecture to predict disparity map in coarse-to-fine manner. At the first stage, DispNet was adopted to produce the coarse disparity map. Then, at the second stage network, the initial disparity was refined with the residual maps produced at multiple scales. Although deep learning models have obtained a better success compared with traditional methods, there is still potential to improve the prediction in consideration of the specialty of stereo matching. The first network using the understanding of stereo geometry was GC-Net [23], in which a 4D cost volume was formed and a 3D convolution was used to regularise the cost. Inspired by GC-Net, Chang and Chen [24] proposed the pyramid stereo matching network with two main contributions: SPP and 3D stacked multiple hourglass networks, resulting in better results than GC-Net. Recently, new and better networks were proposed based on PSMNet with impressive results. Inspired by the SPP in PSMNet and the dilated convolution structure in DeepLab v3+ [28], Zhu et al. [29] introduced a new parallel cross-form pyramid network for stereo matching, resulting in better results than PSMNet. In addition, the general 4D cost volume was reconstructed in GwcNet [30] by introducing group-wise correlation, which preserved better performance when reducing parameters compared with PSMNet.

2.2 Attention mechanism

The attention structure is an important composition of the computer vision networks, including semantic segmentation [31], person re-identification [32, 33], depth estimation [34] and attention-based stereo matching [35]. Specifically, SE-Net [25] aims to improve the representation ability by introducing attention vector along the channel dimension, and then aggregating the features as the channel-wise multiplication between input features and attention vector. The non-local [36] structure is used to separate the input feature into query, key and value features. The query and key are applied to get attention matrix, and then the output features are calculated by a matrix multiplication by using value features and attention matrix. There are also other variants of attention structure, such as the new designed asymmetric pyramid non-local structure proposed by Zhu et al. [37] to reduce the computation and memory consumption. Also, Zhang et al. has [38] introduced the non-local block into their network for image synthesis and achieved better performance than prior work. Although the design and applications are widely used, the inherent role of attention structure has been just to advance the information-meaningful features and suppress the weak clues. Besides, with respect to huge computation cost of non-local operation, the attention structure adopted in this paper is mostly similar to [25]. This paper aims to effectively obtain the information-rich features without introducing too much computational cost, and the channel attention structure is used to boost the performance of our network. Moreover, the attention structure is further extended for 4D features.

2.3 Encoder–decoder structure

The encoder–decoder structure has been applied to many computer vision tasks, including semantic [28, 39, 40], object detection [41, 42], and classification [43, 44]. Typically, the beginning encoder module gradually reduces the feature maps and acquires the deeper features, the next decoder gradually recovers the required spatial size, by using deconvolution. Ronneberger et al. [45] utilised the skip-connection to their encoder-decoder structure, which could enrich the segmentation performance with more details. Specially, for stereo matching, the Disp-Net [26] applied the up-convolution to feature maps and then concatenated the corresponding encoder clues in their pipeline, the results conserved both low-level features and high-level features. In EdgeStereo [21], the general decoder structure was replaced by using the new designed residual pyramid network with different scales. The HSM [46] computed the encoder features in a coarse-to-fine with 3D residual convolution and volumetric pyramid pooling, and gradually increased the output resolution. It could predict accurate disparity map with low latency. Different from these works, our network not only explores the power of U-Net for multi-scale features learning, but also strengthens it with new embedding.

In our network, no post-processing step is conducted. Meanwhile, the features representation is further improved by introducing additional multi-path attentive aggregation module with attention mechanism. Then, the 3D u-shaped encoder–decoder structure with 3D channel attention is used to regularise the extended cost volume. Finally, the soft-argmin [23] is applied directly to regress the sub-pixel estimation of disparity map.

3 Network architecture

The network consists of four parts: feature extraction, cost volume construction, cost regularisation and disparity prediction. ResNet blocks and MPA module are first used to extract the feature of the original input binocular picture to obtain the feature maps for the left and right views, respectively, and the size of both is 1/4H*1/4W*32 (H: input image's height, W: input image's width, 32: feature map's channel numbers). Then the left and right features are used to build 4D cost volume in cost volume construction section. After obtaining the cost volume, we use the stacked codec structure to regularise the 4D cost volume to filter out noise. At the last step, we first change the cost distribution into a probability distribution, then the weighted summation to get the final predicted disparity value. The structure is shown in Fig. 1. Each element of the networks illustrated in details as follows.

3.1 Feature extraction

Convolutional operation is used to get a deep representation of the next procedure once we have two stereo images. Firstly, three 3*3 convolutional filters with stride one are adopted to keep the initial image size. Following the layer, two residual blocks (3*3 kernel size with stride two) and next two residual blocks (3*3 kernel size with stride one) are utilised. As a result, the deep features with 1/4 size of the input images are obtained. Compared with SPP applied in existing methods [24, 30], the newly designed multi-path attention aggregation module brings better experimental results, as shown in Fig. 2. To sum up, the module offers mainly two advantages: multi-scale attentive features and bigger receptive field. More specifically, three parallel convolutional layers are used in the module. For the first parallel layer, three 3*3 convolutional filters with stride one are applied to the input features, then two 3*3 convolutional layers with stride one and one 3*3 convolutional layer with stride two are used to down-sample the features of the
are applied to select effective features and suppress useless features used for multi-scale feature extraction. Compared with applying so as to help the network to make better predictions without.

The attention module and supervised training receptive field. Thus, both deepening the depth of the network and increasing the width of the network can alleviate the problems in performing downsampled. Compared with the baseline network, we proposed, SPP can get more context information, but pooling size. Furthermore, the parallel features are concatenated together channel reduction ratio. Inspired by the inception network, increasing.

This is an open access article published by the IET under the Creative Commons Attribution -NonCommercial License (http://creativecommons.org/licenses/by-nc/3.0/)

Fig. 1 Whole architecture of MPA-Net. MPA-Net consists of four parts: feature extraction module, 4D cost volume construction, 4D cost regularisation module and disparity map regression

Fig. 2 Architecture of multi-path attention module (MPA). we first use four ResNet-blocks to extract the initial deep feature maps, and adopt the fourth Res-Block output to the MPA module. Then, we concatenate the output with the second Res-Block output

Fig. 3 Architecture of channel attention

half size. And then the input features of 1/4 in the third convolutional layer with one 3*3 convolutional layer with stride one and two 3*3 convolutional layers with stride two are performing downsampled. Compared with the baseline network, we proposed, SPP can get more context information, but pooling operation will lose more detailed information. In order to maintain detailed information, multiple downsampling of different sizes is used for multi-scale feature extraction. Compared with applying SPP module only, the EPE error rate on the scene flow data set is reduced by 12.5%. Inspired by the inception network, increasing width of network can improve the accuracy of prediction results, so we conduct multiple downsampling of different sizes in parallel to introduce multi-scale feature, which is called as ‘Multi-path’ in this paper. In addition, deepening the network depth can increase the receptive field. Thus, both deepening the depth of the network and increasing the width of the network can alleviate the problems in the binocular network. At the same time, inspired by the excellent function of attention mechanism in other computer vision tasks, a lightweight channel attention mechanism is introduced in the MPA module we proposed. The attention module and supervised training applied are used to select effective features and suppress useless features so as to help the network to make better predictions without increasing the parameters and amount of calculation. Then, the channel attention to each path and bilinear interpolation are used to up-sample the last two feature maps to the 1/4 of an input image size. Furthermore, the parallel features are concatenated together and 1*1 convolutional layer is used to reduce the channel numbers to 32. Finally, the second res-block features, fourth res-block features and multi-path attention are concatenated, followed one 1*1 convolutional layer to obtain feature maps of 32 channels. For attention module, the SE-channel attention [25] module is exploited to each up-sampled feature before concatenation. By using SE-channel attention module, the feature maps along channel dimension are recalibrated to boost the meaningful features and the weak clues can be suppressed. The structure of SE-Channel Attention is shown in Fig. 3. For any given feature I, the global information is firstly squeezed into a channel descriptor by using global average pooling. In general, the goal can be achieved by shrinking the spatial dimensions H*W of input, which is defined mathematically in the following equation:

\[ z_c = F_c(I_c) = \frac{1}{H \times W} \sum_{i} \sum_{j} I_c(i,j) \]  

Then, by making use of the information acquired from the squeeze step. The excitation operation can be defined as follows:

\[ c_c = F_c(z_c, W) = \sigma(g(z_c, W)) = \sigma(W_c \sigma(z_c)) \]  

where \( \delta \) is the ReLU function, \( \sigma \) is the sigmoid function, \( W_1 \in \mathbb{R}^{c \times c} \) and \( W_2 \in \mathbb{R}^{c \times c} \). Firstly, the fully connected layers \( W_1 \) with channel reduction ratio \( r \) (in this paper, \( r=2 \)) is applied to the initial squeezed features, a ReLU and then the channel-increasing layer using \( W_2 \). The excitation produces channel attenuation coefficients \( \alpha_c \in [0,1] \) at each channel with a sigmoid function. The final output of our attention block is acquired by using channel-wise multiplication between the attention coefficients and the initial input features \( I_c \), which is given as

\[ \tilde{I}_c = \alpha_c \cdot I_c \]  

3.2 Cost volume

Instead of using correlation operation in DispNet [26], EdgeStereo [21], the left and right extracted features are directly concatenated to form a 4D cost volume along every disparity based on [23, 24].
Given the left and right features $f_l$ and $f_r$, and then get the 4D cost volume along the maximum disparity $d$ is given as follows:

$$C_{cat}(d, x, y) = \text{Concat}\{f_l(x, y), f_r(x-d, y)\}$$  \hspace{1cm} (4)

where $(x, y)$ is the location of every pixel in the feature maps, $d \in [0, 192)$. The process how to form the 4D cost volume is illustrated in Fig. 4. The grey pixels will be abandoned and replaced with zero value in the source code. Inspired from the traditional methods [9], it is found that the absolute difference between two features also can be regarded as another way to measure the similarity. And another 4D cost volume can be formed by using disparity-level subtraction.

$$C_{abs}(d, x, y) = \text{Abs}\{f_l(x, y) - f_r(x-d, y)\}$$  \hspace{1cm} (5)

Moreover, the disparity-level subtraction is displayed in Fig. 4. Finally, we concatenate those two kinds of cost volume to form the final cost volume for the next section.

$$C_{cat}(d, x, y) = \text{Concat}\{C_{cat}(d, x, y), C_{abs}(d, x, y)\}$$  \hspace{1cm} (6)

### 3.3 Cost regularization

Once the initial 4D cost volume is getting, there is a lot of redundant information to be discarded, and the useful raw information to be extracted. For example, for the fixed disparity value, the pixels in the whole images are considered to have the identical depth, and obviously, those useless information and the useful information need to be filtered. Generally, the encoder-decoder structure is chosen to effectively extract the information in learning-based method. For the encoder and decoder, several convolutional layers will be used in the encoder module to reduce the spatial size and achieve deeper features, and the deconvolution is used in the decoder to recover the spatial information. However, this imperfect structure, especially for the decoder, may not successfully maintain the context details because of the separated design. Inspired from PSMNet [24], three u-shaped encoder-decoder structures with 3D channel attention are proposed to bridge the encoder features to the corresponding decoder features for cost regularisation. The 2D attention in the feature extraction is extended to the 4D cost features, named 3D attention, and two attention modules are same in their essence. Finally, the recalibrated features are obtained. As shown in Fig. 5, the features to the 1/4 of input size in our encoder module are down-sampled to get bigger receptive field and deeper features, and the 3D attention module proposed is applied to the low encoder features. Then the decoder features and their corresponding encoder features are concatenated before transiting to the next stage. The decoder feature maps are gradually up-sampled by 2 until the spatial size is recovered as input. Three different choices are provided for the encoder-decoder design of our experiment, (i) the encoder-decoder structure with attentive connection, as shown in Fig. 5, (ii) the attention gate is abandoned and the encoder is concatenated and fused to their corresponding decoder features, (iii) the connection is removed and just the convolutional layer and de-convolutional
layer are used as general encoder-decoder. Among the choices, the experimental results show that the first choice achieves the best performance.

3.4 Disparity estimation and loss function

After acquiring the regularised cost, the estimated disparity map is regressed from the meaningful cost. In this section, two convolutional layers are used to gradually reduce the channel number, and squeeze the 4D cost to 3D ($D^2H^4W^4$). Following [23], the same way, named soft-argmin, is used to regress more smooth sub-pixel disparity estimation. The estimated disparity $\hat{d}$ can be calculated as

$$\hat{d} = \sum_{d=0}^{D_{\text{max}}-1} d \cdot \sigma(-c_d)$$

where $c_d$ is the 3D cost feature map, $D_{\text{max}}$ is the maximum value of disparity (in this paper, $D_{\text{max}}$=192) and $\sigma(\cdot)$ is the sigmoid function. Because of our stacked structure, more than one output can be got from the encoder-decoder structure. In fact, there are three outputs from each last decoder layer for disparity estimation. Therefore, three disparity maps can be got in the training step. For the test step, the last disparity map is taken as the final result. We refer to [24], the loss function in the training step is listed as follows:

$$\text{Loss}(d, \hat{d}) = 0.5 \cdot L_1(d - \hat{d}) + 0.7 \cdot L_1(d - \hat{d}) + L_1(d - \hat{d})$$

where $\hat{d}_i$ is the $i$ stage disparity map output, $d$ is the ground-truth and $L_1$ denotes $L_1$ smooth loss, which is calculated as

$$L_1(x) = \frac{0.5 \cdot x^2}{|x| < 1}$$

4 Experiments

In this section, qualitative and quantitative experiments will be provided on three different data sets. In Section 4.1, the basic situation of the three different data sets, training and test metrics are described. In Section 4.2, ablation study on the Scene Flow data set [26] and KITTI 2015 data set [48] are performed to justify the module design choice and the influence made by the proposed mechanism. In Section 4.3, the results obtained from our network are compared with several state-of-the-art methods on the KITTI 2012 [49] and KITTI 2015 [48] benchmarks.

4.1 Data sets and metrics

We mainly use the following data with sets to train and test our network.

| Feature extraction | Cost volume | Cost regularisation | Evaluation metrics |
|--------------------|-------------|---------------------|--------------------|
| SPP                | MP          | MPA                 | Concat             | Substraction | Basic | ED   | ED-S | ED-A | EPE | >1px | >2px | >3px |
| ✓                  | ✓           | ✓                   | ✓                  | ✓            | ✓     | ✓    | ✓    | ✓    |     | ✓    | ✓    | ✓    |
| ✓                  | ✓           | ✓                   | ✓                  | ✓            | ✓     | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    |
| ✓                  | ✓           | ✓                   | ✓                  | ✓            | ✓     | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    |
| ✓                  | ✓           | ✓                   | ✓                  | ✓            | ✓     | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    |
| ✓                  | ✓           | ✓                   | ✓                  | ✓            | ✓     | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    |
| ✓                  | ✓           | ✓                   | ✓                  | ✓            | ✓     | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    |
| ✓                  | ✓           | ✓                   | ✓                  | ✓            | ✓     | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    |

The test is on SceneFlow data set with ten epochs. 'MP' means MPA without attention. 'Concat' means 'disparity-level concatenation', 'Substraction' means 'disparity-level subtraction', 'Cost regu.' means 'Cost regularisation', 'ED' means 'encoder-decoder with skip-connection' and 'ED-A' means 'encoder-decoder module with attention structure'.

Table 1: Ablation study of MPA-Net

- SceneFlow: Scene Flow is a synthesised data set with 35,454 training image pairs and 4,370 testing image pairs. It has dense ground-truth disparity maps for both training and testing. The picture size is 540*960, and we take the disparity value less than 192 to train our network.
- KITTI 2012: KITTI 2012 is real-world data set with 194 stereo pairs with sparse ground-truth disparities for training and 195 stereo pairs without ground-truth. We evaluate our test result by submitting it to the KITTI 2012 benchmark website. The image size is 376*1240 and its valid disparity value is $D \in [0, 192]$.
- KITTI 2015: KITTI 2015 is another real-world data set acquired from a driving car. It contains 200 training stereo images with sparse ground-truth disparities and 200 testing stereo images without ground-truth. The image size is 375*1242 and its valid disparity value is $D \in [0, 192]$.

Implementation details: Our network is implemented by using PyTorch and Adam ($\beta_1=0.9; \beta_2=0.999$) to optimise it. Firstly, we train our network on Scene Flow data set with 256*512 input image size for totally 10 epochs. For the first five epochs, our learning rate is 0.01, and the learning rate of the next three epochs is 0.0005, and the last two is 0.0001. Next, we further fine-tune our network on KITTI 2012 and KITTI 2015 data sets. For KITTI 2012, we train the first 200 epochs with learning rate 0.001 and reduce to 0.0005 for next 100 epochs, the last 100 epochs we use 0.0001 as its learning-rate. For KITTI 2015, we take the same strategy as KITTI 2012. The batch-size is 4 for training on two GTX 1080 Ti GPUs.

Evaluation metrics: For Scene Flow data set, we usually use the end-point-error (EPE) as our evaluation metrics, which calculates the average Euclidean distance between estimated and ground-truth. To provide more convincing results, we also extend the EPE by introducing the percentage of disparities with EPE larger than $t$ pixels (> $t$ pixels). For KITTI 2015, if the disparity EPE is < $3$ pixels, we consider the disparity is correct. For KITTI 2012, the four different thresholds will be used to measure the erroneous pixels.

4.2 Ablation study of our network

In this section, we will verify each component of our network on Scene Flow data set, including multi-path attention aggregation module, the extended cost volume and the three different encoder-decoder structures. We train the network on the Scene Flow data set for 10 epochs and take the EPE, >1 pixel, >2 pixels and >3 pixels as our measure metrics. We follow the baseline setting in [24]. The whole ablation study results are shown in Table 1. Baseline: after 10 epochs training, we test the baseline network on the Scene Flow test data set and the EPE is 1.153. MP: by using the multi-path module, it improves EPE by 10.23%, it indicates that the multi-scale context information is important for improving the network performance. MPA: we notice that adding the attention improves the EPE to 1.009, > 1 px, showing that the attention mechanism is effective in our network for motivating meaningful features. Especially, it reduces the > 1 pixel error by nearly 2.64% and indicates it can be helpful to achieve more accurate disparity prediction. Cost volume: after extending the cost volume, as we
can see in Table 1, the result is pretty good with a 7.14% decrease in EPE and 6.84% decrease in > 1 pixel error. **Cost regularisation:** we first replace the basic structure with the stacked encoder-decoder (SED) structure, the EPE is significantly reduced from 0.937 to 0.885. We then use the skip-connection to connect the decoder with their corresponding encoder features, which bring 2.15% improvement for EPE. To further improve our encoder-decoder with advanced attention module, we use the SED with attentive (SED-A) skip-connection and get the final EPE of 0.854. Compared with basic network, our final network has 25.93% improvement in EPE, 22.63% improvement in > 1 pixel error, 24.97% improvement in > 2 pixel error and 26.67% improvement in > 3 pixel error.

To test the generality of our network for real-world scene, we divide the KITTI 2015 training data set into a training data set (80%) and a validation data set (20%). We then use the evaluation metrics as SceneFlow data set to do our further ablation study on the KITTI data set. The whole ablation study results are on KITTI 2015 validation data set, as shown in Table 2. We acquire the best performance when all the proposed modules are used in stereo matching network. The gradually reducing of error rate brought by each module indicates that our proposed modules and the unified stereo matching network still work for the real-world scene, though its ground-truth is sparse and limited [48]. Moreover, we point out that how the parameters and running time to evaluate the proposed modules affect the whole performance. By using MP module, we reduce the parameters while achieving a lower error rate. The attention mechanism (MPA) further improves the network performance without introducing too many parameters. The extended cost volume also helps to improve the performance. To further reduce the error rate, we use the SED structure to replace the basic design in our baseline network. At last, our final network achieves ~19.2% lower EPE than the baseline network while the parameters are basically the same. And the running time follows the same discipline.

Two examples are given to further verify the effectiveness of attention mechanism in our stereo matching network, as shown in Fig. 6. For MPA module, we first output the feature map without using attention module. Compared with the MPA (with attenuation), we can easily find that our MPA module yields more reasonable visual results than that of the MPA without attention. As shown in the red box, the MPA effectively focuses more attention on the car and the road signs that we need to be well estimated in the stereo matching task. For the cost regularisation section, we do the same as MAP. We achieve better visual results compared with the cost regularisation section without attention, obviously, and the section with attention is effective in suppressing the image noise and the display effect is getting closer to the input image. Moreover, we can further see the outline of the car and the road signs in the red box.

### 4.3 Benchmark results

In this section, we will compare our results with several existing state-of-the-art algorithms on KITTI 2012 and KITTI 2015 benchmarks to verify the universality of our network, especially for the real-world scene. To adopt our model on KITTI data set, we further fine-tuning the best model acquired from SceneFlow data set on KITTI training data set. The comparisons are shown in Tables 3 and 4, respectively. Table 3 shows the comparison on the KITTI 2012 benchmark. From the results, our network achieves the best performance in both non-occluded regions and all pixels. Compared with state-of-the-art algorithm [24], we improve the > 3 px Out-all pixels error from 1.89 to 1.83%. Meanwhile, even the SegStereo [22] uses the multi-task learning method that using additional semantic information, our network still leads to better results. Note that, our

| Feature extraction | Cost volume | Cost regularisation | KITTI2015 | Params, M | Time, s |
|-------------------|-------------|---------------------|-----------|----------|--------|
| SPP   | MP          | MPA                | Concat    | Substraction | Basic | ED     | ED-S | ED-A | EPE >1px | >2px | >3px |
| ✓     | ✓           | ✓                  | ✓         | ✓             | 0.863 | 17.048 | 5.77 | 3.204 | 9.02    | 0.656 |
| ✓     | ✓           | ✓                  | ✓         | ✓             | 0.789 | 16.969 | 5.66 | 3.158 | 6.79    | 0.370 |
| ✓     | ✓           | ✓                  | ✓         | ✓             | 0.743 | 16.557 | 4.958| 2.356 | 6.8     | 0.400 |
| ✓     | ✓           | ✓                  | ✓         | ✓             | 0.736 | 16.409 | 5.135| 2.308 | 6.09    | 0.365 |
| ✓     | ✓           | ✓                  | ✓         | ✓             | 0.714 | 15.318 | 4.651| 2.104 | 8.95    | 0.629 |
| ✓     | ✓           | ✓                  | ✓         | ✓             | 0.708 | 14.988 | 4.684| 2.16  | 9.59    | 0.705 |
| ✓     | ✓           | ✓                  | ✓         | ✓             | 0.697 | 14.302 | 4.434| 1.998 | 9.62    | 0.712 |

The evaluation metrics are the same as the ablation study on SceneFlow data set. In order to avoid the initial errors which are large in general, we statistic the last 100 epochs error rate and take the average value as final results. We further provide the information of the parameters and running time.

**Bold value indicates the lowest value in the listed indicators.**

![Visual results of proposed module with attention mechanism. The two rows of each example show the input image, the feature map of two modules output and their corresponding pseudo-colour map](image)
network is also efficient in running time, though our network has 9.62M parameters. We keep the lowest error rate among those published methods and keep not bad running time at the same time. Some visualisations on the KITTI 2012 test disparity maps of our network are shown in Fig. 7, which shows that our model outperforms the other existing methods, including PSM, HSM and SegStereo. Specifically, we all achieve the lowest error rate in the listed examples. As we can see in the first image of a red box, our network achieves smoother and more continuous prediction of the cars roof. The below red box of error map also proves this conclusion well, which has least error estimated pixels. In consequence, we all achieve the lowest error rate in the listed examples. As we can see in the first image of a red box, our network achieves smoother and more continuous prediction of the car's roof. The below red box of error map also proves this conclusion well, which has least error estimated pixels.

Table 3 KITTI 2012 test results on the benchmark

| Networks | >2 px | >3 px | >4 px | >5 px | Time, s |
|----------|-------|-------|-------|-------|---------|
| Noc      | All   | Noc   | All   | Noc   | All     |
| GCNet    | 2.71  | 3.46  | 1.77  | 2.30  | 1.36    |
| HSM-1.5x | [46]  | 2.65  | 3.32  | 1.53  | 1.99    |
| PSM      | 2.44  | 3.01  | 1.49  | 1.89  | 1.12    |
| PDSNet   | 3.82  | 4.65  | 2.93  | 2.53  | 1.38    |
| SegStereo| [22]  | 2.66  | 3.19  | 1.68  | 2.03    |
| ours     | 2.28  | 2.86  | 1.41  | 1.83  | 1.05    |

Bold value indicates the lowest value in the listed indicators.

Table 4 KITTI 2015 test results on the benchmark

| Networks | All | Noc-occluded | Time, s | D1-bg | D1-fg | D1-all | D1-bg | D1-fg | D1-all |
|----------|-----|--------------|---------|-------|-------|--------|-------|-------|--------|
| DispNetC | [26] | 4.32 | 4.41 | 4.34 | 4.11 | 3.72 | 4.05 | 0.06 |
| CRL      | [27] | 2.48 | 3.59 | 2.67 | 2.32 | 3.12 | 2.45 | 0.47 |
| GCNet    | [23] | 2.21 | 6.16 | 2.87 | 2.02 | 5.58 | 2.61 | 1.16 |
| HSM-1.5x | [46] | 1.95 | 3.93 | 2.28 | 1.76 | 3.55 | 2.06 | 0.085 |
| PSM      | [24] | 1.86 | 4.62 | 2.32 | 1.71 | 4.31 | 2.14 | 0.41 |
| PDSNet   | [50] | 2.29 | 4.05 | 2.58 | 2.09 | 3.68 | 2.36 | 0.5 |
| SegStereo| [22] | 1.88 | 4.07 | 2.25 | 1.76 | 3.70 | 2.08 | 0.6 |
| ours     | 1.78 | 4.95 | 2.22 | 1.57 | 4.58 | 2.06 | 0.86 |

Bold value indicates the lowest value in the listed indicators.

For KITTI 2015 data set, we also adopt the acquired best model in the SceneFlow and finetune the network on the KITTI 2015 data set for 400 epochs. The comparison of KITTI 2015 is shown in Table 4. We also achieve the best performance in background and all regions among the listing algorithms with the same training set. Some visualisations on the KITTI 2015 test disparity maps of our network are shown in Fig. 8. In the first example, our result gets a good trade-off between the accuracy and robustness. The SegStereo achieves the lowest error rate, however, as we can see in the whole image, the predicted disparity map is not good at all. It only predicts the regions that have ground-truth data and lose generality, especially for the traffic light in the red box, we cannot even recognise it as a traffic light, and we doubt that this network may be overfitting. Compared with PSM, our network achieves a lower error rate and performs better on reconstruction effect. Our error rate is a bit lower than that of the HSM, we both perform well in the prediction of traffic light. However, our network performs better than the HSM in terms of the car as shown in the red box of error map. Our network does not perform well about the fence in the distance as shown in the next red box, which harms our final result. In the second example, our network achieves the lowest error rate and rebuilds the street sign well compared with the listed algorithms.

5 Conclusion

In this paper, we propose a novel end-to-end network for stereo matching, named MPA-Net, with MPA, extended cost volume and SEDA-A three different useful modules. The MPA improves the features extraction by introducing multi-scale meaningful features. We extend the manner of cost volume construction by using the disparity-level subtraction and disparity-level concatenation. The SegStereo, we can find the distant scenes are miscalcuated, even though it has a low error rate, that is because the ground-truth of KITTI data set is acquired from Laserscanner and the disparity value is sparse and the range is limited. We suppose the SegStereo network relies too much on the labelled knowledge and could not produce a fine prediction in those regions without valid disparity value, such as the sky and things in the distance. For our network, it is robust enough to estimate those regions even without ground-truth data in the training stage. In the second example image, the state-of-the-art algorithm PSM still has difficulties in handling the just mentioned problem as it has wrong prediction of the sky. Compared with the rest networks, our network performs well in keeping the vehicle edge as shown in the red box of error map.

Fig. 7 Qualitative results on the KITTI 2012 test sets. The two rows of each example show the input image, the predicted disparity map of different algorithms and corresponding their error maps. The error map scales linearly between 0 (black) and >= 5 (white) pixels error and more white pixels indicates a higher error rate. The number of erroneous pixels at 2 pixels thresholds are provided below the error maps with Out-Noc and Out-All, respectively.
SED-A further regularises the cost volume, and achieves better performance than general encoder–decoder structures. Comparison to recent stereo matching algorithms on the KITTI benchmark has shown that our network can lead to better disparity map prediction results. In the future work, we will aim to refine this network and achieve real-time display.

### 6 Acknowledgement

This work was supported by Foundation of GuiZhou Educational Committee (QianJiaoHe[2021]240) and Technologies Research and Development Program of QianNan (QianNankeHe[2020]15).

### 7 References

[1] Fu, H., Gong, M., Wang, C., et al.: ‘Deep ordinal regression network for monocular depth estimation’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 2002–2011

[2] Hu, J., Gray, M., Zhang, Y., et al.: ‘Revisiting single image depth estimation: toward higher resolution maps with accurate object boundaries’. 2019 IEEE Winter Conf. on Applications of Computer Vision (WACV), Honolulu, HI, USA, 2019, pp. 1043–1051

[3] Nezhazov, V., Dharmasiri, T., Spek, A., et al.: ‘Real-time joint semantic segmentation and depth estimation using asymmetric annotations’. 2019 Int. Conf. on Robotics and Automation (ICRA), Montreal, Canada, 2019, pp. 7101–7107

[4] Torralba, A., Oliva, A.: ‘Depth estimation from image structure’, IEEE Trans. Pattern Anal. Mach. Intell., 2002, 24, (9), pp. 1226–1238

[5] Uddin, M.Z., Hassan, M.M., Almogren, A., et al.: ‘Facial expression recognition utilizing local direction-based robust features and deep belief network’, IEEE Access, 2017, 5, pp. 4525–4536

[6] Bleyer, M., Rhenmann, C., Rother, C.: ‘Patchmatch stereo-stereo matching with slanted support windows’. British Machine Vision Conf. (BMVC), Dundee, Scotland, 2011, vol. 11, pp. 1–11

[7] Taniai, T., Matsushita, Y., Sato, Y., et al.: ‘Continuous 3D label stereo matching using local expansion moves’, IEEE Trans. Pattern Anal. Mach. Intell., 2017, 40, (11), pp. 2725–2739

[8] Yuan, Q.: ‘Stereo matching using tree filtering’, IEEE Trans. Pattern Anal. Mach. Intell., 2014, 37, (4), pp. 834–846

[9] Schuster, D., Szczesniak, R.: ‘A taxonomy and evaluation of dense two-frame stereo correspondence algorithms’, Int. J. Comput. Vis., 2002, 47, (1–3), pp. 7–42

[10] Taniai, T., Matsushita, Y., Naemura, T.: ‘Graph cut based continuous stereo matching using locally accurate object boundaries’. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 2014, pp. 1613–1620

[11] Yang, Q.: ‘A non-local cost aggregation method for stereo matching’. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 2012, pp. 1402–1409

[12] Li, L., Yu, X., Zhang, S., et al.: ‘3d cost aggregation with multiple minimum spanning trees for stereo matching’. Appl. Opt., 2017, 56, (12), pp. 3411–3420

[13] Yamaguchi, K., Cmeraus, D., Urtasun, R.: ‘Efficient joint segmentation, occlusion labeling, stereo and flow estimation’. European Conference on Computer Vision (ECCV), Zurich, Switzerland, 2014, pp. 756–771

[14] Zhang, C., Li, Z., Cheng, Y., et al.: ‘Meshtereo: a global stereo model with mesh alignment regularization for view interpolation’. Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 2057–2065

[15] Song, X., Zhao, X., Hu, H., et al.: ‘Edgestereo: a context integrated residual pyramid network for stereo matching’. Asian Conf. on Computer Vision, Perth, Australia, 2018, pp. 20–35

[16] Taniai, T., Matsushita, Y., Naemura, T.: ‘Graph cut based continuous stereo matching using locally accurate object boundaries’. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 2014, pp. 1613–1620

[17] Yang, Q.: ‘A non-local cost aggregation method for stereo matching’. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 2012, pp. 1402–1409

[18] Li, L., Yu, X., Zhang, S., et al.: ‘3d cost aggregation with multiple minimum spanning trees for stereo matching’. Appl. Opt., 2017, 56, (12), pp. 3411–3420

[19] Yamaguchi, K., Cmeraus, D., Urtasun, R.: ‘Efficient joint segmentation, occlusion labeling, stereo and flow estimation’. European Conference on Computer Vision (ECCV), Zurich, Switzerland, 2014, pp. 756–771

[20] Zhang, C., Li, Z., Cheng, Y., et al.: ‘Meshtereo: a global stereo model with mesh alignment regularization for view interpolation’. Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 2057–2065

[21] Song, X., Zhao, X., Hu, H., et al.: ‘Edgestereo: a context integrated residual pyramid network for stereo matching’. Asian Conf. on Computer Vision, Perth, Australia, 2018, pp. 20–35

[22] Yang, G., Zhao, H., Shi, J., et al.: ‘Segment: exploiting semantic information for disparity estimation’. Proc. of the European Conf. on Computer Vision (ECCV), Munich, Germany, 2018, pp. 556–565

[23] Kendall, A., Martirosyan, H., Dasgupta, S., et al.: ‘End-to-end learning of geometry and context for deep stereo regression’. Proc. of the IEEE Int. Conf. on Computer Vision, Venice, Italy, 2017, pp. 66–75

[24] Chang, J.-R., Chen, Y.-S.: ‘Pyramid stereo matching network’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 5410–5418

[25] Hu, J., Shen, L., Sun, G.: ‘Squeeze-and-excitation networks’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 5410–5418

[26] Mayer, N., Jig, E., Haussler, P., et al.: ‘A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 4040–4048

[27] Pang, J., Sun, W., Ren, J.S., et al.: ‘Cascade residual learning: a two-stage convolutional neural network for stereo matching’. Proc. of the IEEE Int. Conf. on Computer Vision, Venice, Italy, 2017, pp. 887–895

[28] Chen, L.-C., Zhu, Y., Papandreou, G., et al.: ‘Encoder-decoder with atrous separable convolution for semantic image segmentation’. European Conf. on Computer Vision, Munich, Germany, 2018, pp. 833–851

[29] Zhu, Z., He, M., Dai, X., et al.: ‘Multi-scale cross-conv pyramid network for stereo matching’, arXiv preprint arXiv:1904.11309, 2019

[30] Guo, X., Wang, K., Yang, W., et al.: ‘Group-wise correlation stereo network’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 3327–3332

[31] Fu, J., Liu, J., Tian, H., et al.: ‘Dual attention network for scene segmentation’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 3327–3332

[32] Chen, D., Li, H., Xiao, Y., et al.: ‘Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 1169–1178

[33] Li, S., Bak, S., Carr, P., et al.: ‘Diversity regularized spatiotemporal attention for video-based person re-identification’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 369–378

[34] Sang, H., Wang, Q., Zhao, Y.: ‘Multi-scale context attention network for stereo matching’. IEEE Access, 2019, 7, pp. 15152–15161

[35] Zhang, G., Zhu, D., Shi, W., et al.: ‘Multi-dimensional residual dense attention network for stereo matching’. IEEE Access, 2019, PP, (99), pp. 1–1
[36] Wang, X., Girshick, R., Gupta, A., et al.: ‘Non-local neural networks’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 7794–7803

[37] Zhu, Z., Xu, M., Bai, S., et al.: ‘Asymmetric non-local neural networks for semantic segmentation’. Proc. of the IEEE Int. Conf. on Computer Vision, Seoul, Republic of Korea, 2019, pp. 593–602

[38] Zhang, H., Goodfellow, I., Metaxas, D., et al.: ‘Self-attention generative adversarial networks’. Int. Conf. on Machine Learning, Taiwan, People's Republic of China, 2019, pp. 7354–7363

[39] Kong, H., Fan, L., Zhang, X.: ‘Semantic segmentation with inverted residuals and atrous convolution’. Technical Report, SAE Technical Paper, 2018

[40] Noh, H., Hong, S., Han, B.: ‘Learning deconvolution network for semantic segmentation’. Proc. of the IEEE Int. Conf. on Computer Vision, Santiago, Chile, 2015, pp. 1520–1528

[41] Lang, A.H., Vora, S., Caesar, H., et al.: ‘Pointpillars: fast encoders for object detection from point clouds’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 12697–12705

[42] Li, G., Xie, Y., Wei, T., et al.: ‘Flow guided recurrent neural encoder for video salient object detection’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 3243–3252

[43] Baccouche, M., Mamalet, F., Wolf, C., et al.: ‘Spatio-temporal convolutional sparse auto-encoder for sequence classification’. British Machine Vision Conf. (BMVC), Guildford, UK, 2012, pp. 1–12

[44] Kiraly, A.P., Nader, C.J.A., Grimm, R., et al.: ‘Deep convolutional encoder-decoder for prostate cancer detection and classification’. US Patent App. 15/831,819, 23 August 2018

[45] Ronneberger, O., Fischer, P., Brox, T.: ‘U-net: convolutional networks for biomedical image segmentation’. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 2015, pp. 234–241

[46] Yang, G., Manela, J., Huppold, M., et al.: ‘Hierarchical deep stereo matching on high-resolution images’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 5515–5524

[47] Naiz, V., Hinton, G.E.: ‘Rectified linear units improve restricted boltzmann machines’. Proc. of the 27th Int. Conf. on Machine Learning (ICML-10), Hafai, Israel, 2010, pp. 807–814

[48] Menze, M., Geiger, A.: ‘Object scene flow for autonomous vehicles’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Boston, MA, USA, 2015, pp. 3061–3070

[49] Geiger, A., Lenz, P., Urtasun, R.: ‘Are we ready for autonomous driving? the kitti vision benchmark suite’. 2012 IEEE Conf. on Computer Vision and Pattern Recognition, Providence, RI, US, 2012, pp. 3354–3361

[50] Tulyakov, S., Ivanov, A., Fleuret, F.: ‘Practical Deep Stereo (PDS): Toward applications-friendly deep stereo matching’. Proc. of the Int. Conf. on Neural Information Processing Systems (NIPS), Montreal, Canada, 2018, (to appear)