Displays 74 (2022) 102243

Contents lists available at ScienceDirect

Displays

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

AMDCNet: An attentional multi-directional convolutional network for stereo matching[☆]

Hewei Wang [a], Yijie Li [a], Shijia Xi [a], Shaofan Wang [b], Muhammad Salman Pathan [c,d], Soumyabrata Dev [c,d,*]

[a] Beijing-Dublin International College, Beijing University of Technology, Beijing 100124, China
[b] Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
[c] The ADAPT SFI Research Centre, Dublin D04V1W8, Ireland
[d] School of Computer Science, University College Dublin, Dublin D04V1W8, Ireland

ARTICLE INFO

Keywords:
Stereo matching
Convolution aggregation network
Visual sensitive
Cost aggregation

ABSTRACT

Stereo matching refers to finding the correspondence of a point in the real world between two different storage mediums (*e.g.*, intensity images, depth images, three-dimensional points). There are existing stereo matching methods in the literature, but they exhibit two shortcomings. Firstly, during the feature region extraction of stereo matching, these methods require measuring the distance of regions, but measuring the texture distribution of the region is difficult and might lead to the failure of matching. Secondly, the templates used in these methods are rectangles with a fixed size, while most of the natural images exhibit rich information and are more suitable for flexible templates. In this paper, we propose an attentional multi-directional convolutional network (AMDCNet) for circumventing these issues. Our AMDCNet approach consists of three stages: extract the visual sensitivity factor, construct the multi-directional aggregation template and utilize left–right consistency detection to optimize. We evaluate our approach using standard images in the Middlebury test dataset, Scene Flow and KITTI 2015. Experimental results show that AMDCNet can reduce the mismatch rate, and also show significant improvement in accuracy compared with some classical method. In some scenarios, it surpasses some advanced methods based on deep learning. The model code, dataset, and results of the experiments in this paper are available at: https://github.com/WangHewei16/Attentional-Multi-Directional-Convolution-Network.

1. Introduction

Stereo matching, referred to as finding the correspondence of a point in the real world between two different storage mediums (e.g., intensity images, depth images, three-dimensional points), serves as one of the depth perception techniques for scene understanding, has wide applications in the fields of computer vision [1,2], robot industry [3,4], earth observations [5,6] and automatic driving technology [7,8]. According to the statistics, the eyes are an important channel for us to obtain external information, and vision can bring us more abundant information. The visual information reflected by objects is transmitted to the brain through nerve tissue and will be effectively processed by the visual cortex of the brain in the way of content and space respectively. Human eyes have a strong ability to distinguish color information with a great subjective influence, whereas the perception of the gray level of the image is very poor. The emergence of machine vision gets rid of the influence of human factors where the resolution of gray can reach 256 levels and has low requirements for the environment that intensifies our deep understanding of the image. Stereo vision technology in the field of computer vision has always been one of the hot spots of research. In recent years, stereo matching technology in stereo vision was the research direction of many scholars. The disparity map obtained by stereo matching technology is widely used in many intelligent fields. Such stereo matching can be used in the volumetric estimation of tumors in the domain of e-health analytics [9–11], insertion of 3D products [12,13] in scenes for product placements [14–16], reconstruction of cloud base height for atmospheric analysis [17,18], and amongst others. The disparity information contained in each pixel of the disparity map reflects the three-dimensional information of the scene, which has important application value for robot map construction (three-dimensional reconstruction), obstacle avoidance navigation

[☆] This paper was recommended for publication by Guangtao Zhai.

[*] Corresponding author at: School of Computer Science, University College Dublin, Dublin D04V1W8, Ireland.
E-mail addresses: hewei.wang@ucdconnect.ie (H. Wang), yijie.li@ucdconnect.ie (Y. Li), shijia.xi@ucdconnect.ie (S. Xi), wangshaofan@bjut.edu.cn (S. Wang), salman.pathan@adaptcentre.ie (M.S. Pathan), soumyabrata.dev@ucd.ie (S. Dev).

https://doi.org/10.1016/j.displa.2022.102243
Received 25 January 2022; Received in revised form 2 May 2022; Accepted 25 May 2022
Available online 6 June 2022

and many other fields. The image source of this technology is usually captured by a binocular camera [19,20], which captures two images from left and right perspectives, so as to carry out the stereo matching operation.

Existing stereo matching methods can be roughly divided into two categories. The first category, named fixed template methods, are designed with the pre-defined convolution template parameters. The second category named adaptive template methods, only take the distribution of gray difference as one of the judgment conditions, with the limited cost value of calculation. Due to the simplicity of the efficiency, the second category is one of the mainstream methods of stereo matching. However, they exhibit the following shortcomings and our motivation is to solve these issues.

- During the feature region extraction of stereo matching, these algorithms require measuring the distance of regions (*viz.*, Hamming distance). However, it is difficult to accurately measure the texture distribution of the region. This may lead to difficulty during matching.
- The templates used in such stereo matching algorithms are rectangles with a fixed size. However, most of the natural images exhibit rich information. Therefore, applying fixed size templates tend to lose some visual sensitive information, such as edge points, and contour points, leading to a reduced performance of stereo matching accuracy.

To solve the aforementioned issues, we propose an attentional multi-directional convolutional network (AMDCNet) for stereo matching. AMDCNet first detects the edge points, contour points and other visual sensitive areas of an image, and analyzes the texture distribution difference between them and smooth areas. Then it builds the weight factor of visual sensitivity and guides the attentional direction of the degree of freedom extension in the subsequent aggregation stage. In the aggregation stage, AMDCNet constructs a multi-directional extension template by using the weight factor with vision induced attention and generates an image by processing covered pixels with the guidance of the template.

The main contribution of AMDCNet is twofold:

- AMDCNet introduces a visual sensitivity factor in the cost calculation stage and constructs a multi-directional aggregation template by using visual sensitivity factor as attention to improve the matching degree.
- AMDCNet produces promising results in mismatching rate, disparity estimation, matching precision experiments, compared with several state-of-the-art stereo matching methods. The code related to this paper is available here: https://github.com/WangHewei16/Attentional-Multi-Directional-Convolution-Network.

This paper is organized as follows. Section 2 analyzes and discusses the relevant algorithms of stereo matching with their advantages and disadvantages. Section 3 describes AMDCNet in details. Section 4 conducts experiments on three datasets: Middlebury, Scene Flow and KITTI 2015 datasets. Section 5 concludes this paper with future work.

2. Related works

In this section, we review the related works on stereo matching from three angles: the process of stereo matching, region-to-region matching algorithm, and point-to-point matching algorithm.

2.1. The process of stereo matching

The first stereo matching technology was proposed by Marr at the end of the 1960s [21]. This opened up the area for research of stereo matching technology. Stereo matching algorithms can be roughly divided into two categories: local matching and global matching. After decades of development, a variety of stereo matching research continues to emerge. The process of stereo matching can be divided into four stages:

1. Cost calculation;
2. Cost aggregation;
3. Calculation of initial disparity map;
4. Disparity optimization [22].

The process of cost calculation is usually to determine the initial generation value by calculating the gray value difference of three channels of a pixel in the left and right view images. The cost aggregation stage can be regarded as a filtering operation, in which the cost obtained in the previous stage is aggregated. In the first step, the parallax value of a single pixel is calculated at the cost, and these pixels are isolated points, often accompanied by a lot of noise, so a similar filtering process is needed to optimize. After aggregating the isolated parallax values in the previous step, we can get an image that is close to the real parallax image. In the next step, we need to select the most appropriate parallax value for the generated multiple parallax values, and often select the parallax corresponding to the minimum generation value. In global stereo matching, the energy function is used to generate a disparity map. The optimization of parallax can be considered a post-processing process. It further optimizes the parallax value obtained above. This is one of the commonly used methods that includes left–right consistency detection, regional voting, linear interpolation and other methods, as well as relatively simple median filtering and other operations. The optimization phase can usually improve the matching effect.

2.2. Region-to-region matching algorithm

With the development of stereo matching technology, many scholars have used this area for research purposes. Simoncelli et al. proposed Sum of Absolute Differences (SAD) stereo matching algorithm based on the gray difference [23]. Based on the former, Da et al. improved the algorithm and used the gray features of the image to find the optimal matching region [24]. Some scholars use the whole image of the characteristics of stereo matching to meet the requirements of the parallax map. Wang et al. in [25] proposed a novel stereo matching algorithm where they introduced the visual sensitivity factor in the cost calculation stage of the matching process for an matching. In [26], Sun et al. proposed a single direction stereo matching algorithm, which makes various limit constraints between the image and another image to be matched, and finds the optimal pixel in another image. Yin et al. used the threshold of color to segment the processing image, and then fitted the segmented region, and used the energy function minimization to optimize the result image [27]. However, the threshold used for color segmentation in his research is very strict, which is easy to cause over-segmentation and under segmentation. Similar to the SAD and Normalized Cross-Correlation (NCC) algorithm, Wang et al. extracted local operators and proposed a real-time matching method, which can quickly compute the result image, but its matching effect is poor, which affects the subsequent application [28].

Zabih et al. have first developed a nonlinear stereo matching algorithm that is based on the premise of less time consumption to obtain a better matching effect of the image [29]. The emergence of this method promotes the development of stereo matching technology, but its limitations are also relatively large, and the matching effect is poor in non-smooth areas. This transformation method is highly dependent on the only pixel in the region template. Noise has a great impact on the cost calculation process, resulting in the calculation of a large deviation of the parallax value that greatly reduces the image matching effect. After this method was proposed, many scholars have improved it, mainly from the overall image matching effect. For instance, Fan et al. divided the noise fluctuation range in the matching process to prevent excessive noise from affecting the image matching effect [30]. Similarly, Men et al. used the gradient value, changed direction and gray information as the cost of image matching [31]. Guo et al. constructed the energy function of local information based on census stereo matching algorithm, which improved the matching effect [32].

Some scholars put forward new ideas according to the limitations of the algorithm proposed by Zabih et al. [29] in the stage of cost aggregation and post-processing optimization. For example, Chai and Cao [33] fused the SAD algorithm with census algorithm according to the characteristics of better local matching effect, so that the gray difference can guide the cost calculation process. In the process of parallax refinement, a median filter is mainly used to eliminate isolated noise and "bad points", and to reduce the cost, quadratic linear interpolation is used to further improve the image matching effect. However, the initial cost of the above method is relatively simple and the robustness is low, so the matching effect of the initial disparity map is not high, and the improvement of the overall image quality is limited. In the process of cost aggregation, the rectangular template is used to aggregate the generation value. This fixed template usually loses a lot of important information and affects the parallax effect. Therefore, on this basis, Zhang et al. [34] proposed a relatively simple cross extension template, which sets the extension condition as gray difference and distance difference to guide the cost aggregation process. Mei et al. [35] enhanced the robustness of automatic growth based on the former and added a third constraint condition, which made the template construction closer to the distribution of texture features in the image region, and improved the matching degree of the image to a certain extent. However, the construction of constraint conditions in the aggregation process was relatively simple, and the isolated points in the disparity map were lacking. The matching accuracy still needs to be improved. Based on the improvement of disparity optimization and cost aggregation in stereo matching, more new matching methods are proposed, such as the sub-pixel difference method proposed by an et al. From the perspective of disparity optimization, which can improve the disparity map matching effect by 1%–2% [36], for instances, Zhang et al. [37] and Shan et al. [38] improve the matching effect from the aggregation process. One is to improve the matching accuracy and introduce the regularization term in the aggregation stage whose advantage is to combine a variety of aggregation algorithms; the other is to improve the time-saving performance of stereo matching, using micro adaptive template window for aggregation, and using parallel processing for the whole process to save time and cost. De-Maeztu et al. [39] introduced the concept of the gradient similarity-matching in the aggregation stage to further constrain the calculation method of the initial cost to improve the matching effect.

2.3. Point-to-point matching algorithm

In addition to the region based matching method, the method based on feature point extraction can also obtain the corresponding parallax image. At present, the widely used feature matching methods mainly include ORB [40], SIFT [41] and SURF [42], whose principle is basically to establish the scale by extracting the feature points on the image (often with obvious structural features such as edge contour, corner, inflection point, etc.). The degree space guarantees its scale in-variance, and then the descriptors are generated by the gradient direction of the feature points. Finally, the matching results are obtained by matching feature vectors, but the feature-based matching methods often cause disparity discontinuity due to the sparse nature of feature points. There are many mismatched points, and the generated 3D point cloud is sparse, which affects the reconstruction effect. In this paper, the processing method of feature points is not described in detail.

To be specific, we briefly introduce the classic census algorithm, which is a local region matching algorithm based on nonlinear transformation. It is mainly judged according to the gray difference of the neighboring pixels around the target point. Suppose that in a pixel covered by a 3 × 3 rectangular template, the gray difference between the central point and the surrounding pixels is made when the pixel is larger than the target point. When it is smaller than the target point, it is marked with 0. The principle formula is as follows:

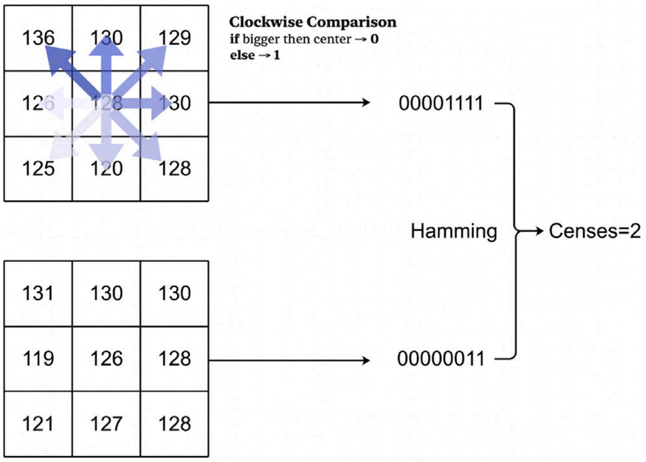


Fig. 1. Census Principle. Clockwise comparison begins in the position of upper left corner. The values in the 3 × 3 template represent the pixel distribution of a certain area in the image, and each area can be represented by a string of binary codes. Finally, the Hamming distance of the two binary strings is calculated to be 2.

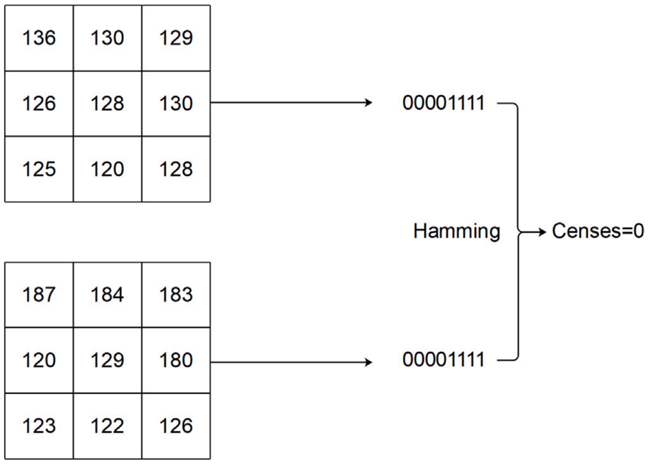
$$
\begin{aligned}
&l(p,p') = \begin{cases} 0, if\ \ (p > p') \\ 1, if\ \ (p < p') \end{cases} \\
&Cens(p) = \bigotimes_{p' \in W(u,v)} l(p,p')
\end{aligned}
\tag{1}
$$

where p represents the target point under the template, that is, the center pixel, p' represents the neighboring pixels, and $W(u,v)$ which is assumed to be a 3 × 3 rectangular template, represents the central pixel of the window. Through the above model, the pixels under the template can be reclassified, and then the template can be moved to traverse the whole image. Finally, many binary strings similar to bit code can be obtained. These strings reflect the texture distribution of the template region to a certain extent, and can roughly represent the feature information of the region. Then, from another image to be matched, we plan the same operation as above: find the region similar to the feature information of the region, complete the matching and calculate the parallax. The judging condition is selected by the Hamming distance. The specific schematic diagram is shown in Fig. 1. The census algorithm calculation is as follows: from clockwise from the top left, we compare each number to the center number. If the number is larger than the center number, the bit calculation result is 0. If the number is smaller or equal to the center number, the bit calculation result is 1.

However, this method also has many drawbacks. For instance, it only takes the distribution of gray difference as one of the judgment conditions, and the cost value of calculation is limited. Moreover, the template used in this method is a rectangle, and most of the images processed are rich in information, so it is easy to lose some visual sensitive information, such as edge points, contour points and so on. According to the algorithm principle, two groups of binary string codes are obtained as shown in Fig. 2, which are the same as those in Fig. 1. The calculated value of Hamming distance is also figured out, but the texture distribution of the region they represent is vastly different, so it is easy to cause the phenomenon of matching failure.

3. AMDCNet

AMDCNet consists of three stages (see Fig. 3 as a flowchart of AMDCNet). In the sensitive points extraction stage, AMDCNet detects



136	130	129
126	128	130
125	120	128

187	184	183
120	129	180
123	122	126

Fig. 2. Different regions of images. The values in the 3 × 3 template represent the pixel distribution of a certain area in the image, and each area can be represented by a string of binary codes. Finally, the Hamming distance of the two binary strings is calculated to be 0.

the edge points, contour points and other visual sensitive areas in the image, analyzes the texture distribution difference between them and smooth areas, and guides the judgment weight of degree of freedom extension. In the multi-directional template matching stage, AMDCNet constructs a multi-directional extension template, then, extended pixel length is guided according to the attention, which is induced by visual sensitivity, finally, AMDCNet completes the whole image processing. In the optimization stage, AMDCNet uses the left–right consistency detection method to further reduce the error matching rate.

3.1. Extract the visual sensitivity factor

As shown in Fig. 4, the template is divided into two directions: horizontal H_x and vertical H_y, and convolution operation is performed on the covered pixels through the two directions, and two candidate factors are extracted subsequently. One represents the brightness difference in each direction, and the other represents the directivity of the gradient. We use this factor to calculate the visual sensitive area. The brightness difference is calculated as follows:

$$\begin{aligned} G_x &= H_x * A \\ G_y &= H_y * A \end{aligned} \tag{2}$$

where $*$ denotes the convolution operator, A is the input image, G_x is the transverse brightness value, and G_y is the longitudinal brightness value. By summing the absolute values of the two, the gray value of a "point" representing the region can be obtained as follows:

$$G = \sqrt{G_x^2 + G_y^2} \tag{3}$$

The calculated G is used to match and guide the two images. We use the Hamming distance between the binary strings to represent the first generation value to highlight the grayscale changes in the two regions. Then, we further judge whether the difference of G between the two regions is less than a certain threshold. If two regions exhibit different gray levels but with the same sensitivity, then we set it as a candidate matching, and further obtain the parallax between the two images. The process of sequential judgment by two generations of values is expressed by:

$$\begin{aligned} f_s(p,d) &= 1 - \exp\left[-\frac{C_s(p,d)}{\lambda_s}\right] \\ f_g(p,d) &= 1 - \exp\left[-\frac{C_g(p,d)}{\lambda_g}\right] \end{aligned} \tag{4}$$

where p represents the current pixel, d represents the horizontal displacement difference of the points in the two images, C_s, C_g represent the cost of the points in the two positions, respectively, and f_s is the mapping value of C_s after a certain transformation and gray approximation. After gray screening, the second cost value, also visual sensitivity, f_g is calculated and screened. λ is an empirical parameter, which makes these two kinds of generational values project to the interval of 0 and 1 using an index.

The above method can be used to calculate the initial disparity map of the two images, but this is not the best result. This paper proposes a cost aggregation algorithm based on the non-directional growth template controlled by the visually sensitive directional factor.

3.2. Construct an attentional multi-directional aggregation template

We construct an attentional multi-dimensional aggregation template in this section. Aggregation of AMDCNet is not a simple superposition, but through visual sensitivity factor proposed in Section 3.1 as attention. The visual sensitivity direction of a certain area can be easily obtained. The calculation is as follows:

$$\theta = \arctan(\frac{G_y}{G_x}) \tag{5}$$

where the angle value θ obtained by the above formula is called the visual acuity direction factor, and can be regarded as the sensitivity direction especially when the gray level of images changes greatly. The pixels and areas in the image will produce such direction factors. Therefore, we can use this factor to guide the aggregation results in the next stage of aggregation. This paper proposes a multi-degree-of-freedom template based on orientation factors.

AMDCNet builds an aggregation template for pixels. This template is not a fixed rectangular window, but an adaptive template with a constantly changing shape. Such an adaptive template will meet the four-degree-of-freedom growth decision condition. The specific process is as follows:

$$\begin{cases} D_s(p_1,p) < L_1 \\ D_c(p_1,p) < f(\tau_1,\theta) \\ D_c(p_1,p_1+(1,0)) < \tau_2 \\ D_c(p_1,p) < f(\tau_3,\theta), if\ \ L_1 < D_3(p_1,p) < L_2 \end{cases} \tag{6}$$

with $\tau_3 < \tau_2 <= \tau_1, L_1 < L_2$

where p represents the central pixel, and p_1 represents a point extending a certain distance. The function D_s is used to generate the distance between two points and ensure that its value is less than a certain threshold L_1 or L_2. D_c represents the color similarity of the two points, which is also related to the sensitive direction. When the direction is similar to the extended distance, it means that the color difference between the point and the reference point is greater. We set a gray threshold γ_1 on this basis and judge the gray scale difference between the two points. If it is less than its threshold, it continues to extend in this direction, and L_1 represents a distance threshold. The third degree of freedom judgment is to compare the forward time during the extension process. When a newly extended point still out of the threshold with the reference point, it also needs to be compared with the point at the previous moment in grayscale. Similarly, it sets a gray scale threshold as the criterion for the two points. If only this condition is met, it is still can continue to extend. The judgment criterion for the fourth degree of freedom is to mutually restrict the distance and color. When the extension point reaches a certain distance but does not exceed the maximum distance limit, we reduce the previous color threshold γ_3 for more stringent screening. Similarly, when the extension continues, it is still adjusted by the direction factor, and the maximum distance parameter will be set at this time. When the extension distance exceeds the threshold, the extension will be terminated immediately. Fig. 5 shows a growth polymerization template of the degree of freedom

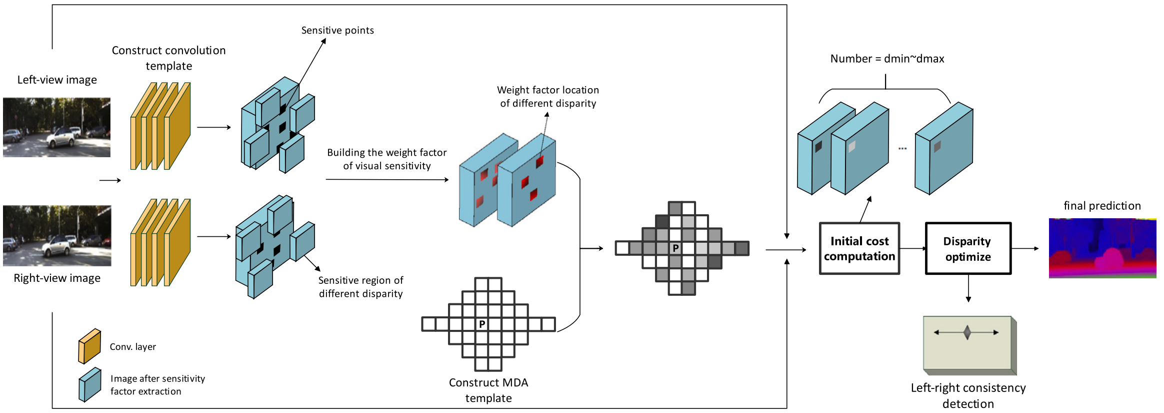

Sensitive points
Construct convolution template
Left-view image
Right-view image
Building the weight factor of visual sensitivity
Weight factor location of different disparity
Sensitive region of different disparity
Number = dmin~dmax
...
P
P
Initial cost computation
Disparity optimize
final prediction
Left-right consistency detection
Conv. layer
Image after sensitivity factor extraction
Construct MDA template

Fig. 3. Architecture overview of proposed AMDCNet. We perform three operations on the image: (a) Perform convolution template to get the visual sensitivity factor; (b) Calculate the initial cost of the image and construct the multi-directional aggregation template. (c) Detect left–right consistency to optimize.

-1	0	+1
-2	0	+2
-1	0	+1

+1	+2	+1
0	0	0
-1	-2	-1

Fig. 4. Convolution template graph. The template has vertical and horizontal directions. Left denotes the horizontal convolution template H_x and right denotes the vertical convolution template H_y. Convolute a region of the image in two directions to get a new factor representing the region.

based on the direction factor control. The leftmost figure is the cross template extended from each reference point, and the rightmost figure is the result of further expansion in the upper and lower directions according to the above rules based on each polymerization arm extended from the left figure. The template in the right subfigure of Fig. 5 is the cost aggregation template that should be used in the end.

Through the above aggregation template, the initial cost calculated at the beginning is aggregated to obtain a more refined parallax image.

3.3. Parallax optimization of AMDCNet: left–right consistency detection

We utilize an optimization algorithm named **left–right consistency detection** to further reduce the error matching rate. For the processed parallax image, AMDCNet carries out the final optimization algorithm, to make the test parallax value closer to the real value. The purpose of the optimization is to further propose some redundant points and improve the matching accuracy. This process is often considered the post-processing stage of the stereo matching process. The optimization algorithm used is the combination of left–right consistency detection and region voting: first, the left–right consistency detection is used to distinguish the occluded point and the mismatched point in the image. For the occluded point P, the first non-occluded point appears in the left and right directions, which is recorded as p_l and p_r. The minimum value of the two is assigned to the P point to complete the filling. For the mismatched points, the region voting method is adopted. The gray distribution of all pixels in the support window area is counted, and the parallax value with the most frequent occurrence is selected to replace the parallax value of the point. Finally, the image is processed with quadratic linear difference operation, and the parallax value of a point is interpolated with the following formula:

$$d_p = d - \frac{C(p, d+1) - C(p, d-1)}{2(C(p, d+1) + C(p, d-1) - 2C(p, d))} \tag{7}$$

where C represents the replacement value of p-pixel points calculated in this method, and d is the horizontal displacement interpolation for p points in the left and right view images.

4. Experiments and discussion

In this section, we discuss the benchmarking experiments that are conducted in this research.

4.1. Datasets

The experimental datasets used in this paper are from Middlebury official test datasets, Scene Flow dataset and KITTI 2015 dataset. These datasets have also been used in many pieces of literature [36,37,43–46]. To be specific, Middlebury datasets are mainly used for vision processing of binocular images and can be used for image restoration, stereo matching, image recognition and other operations as recommended by the majority of scholars. We have used the Middlebury dataset for benchmarking classical stereo matching algorithms. This is because most of these classical algorithms are compared in this dataset. In order to benchmark the more up-to-date methods, we have chosen the KITTI and Scene Flow datasets. We have provided a subjective evaluation of the recent stereo methods in the latest datasets. We have proceeded in this manner, because this helps in a fair comparison amongst all the benchmarking stereo matching algorithms. In this paper, we select four kinds of images in Middlebury 2.0 for experiments, which contains 24 kinds of left and right view images, including the color images and the real parallax images. AMDCNet has tested on Win7 64 bit system by using Visual Studio software and OpenCV2.4.9 framework.

4.2. Parameter setting

Firstly, we do quantitative experiments on distance threshold L_1 and color threshold γ_1 and choose a more appropriate parameter value. In this paper, four kinds of images in the dataset are selected for experiments, which are Tsukuba, Cones, Venus and Teddy. As shown in Fig. 6, we assign different sizes to the distance threshold and the color threshold and judge whether the selected parameters meet most of the images and are worth being selected as empirical parameters according to the error matching rate of the aggregation results.

Mismatch rate represents the percentage of the number of mismatches between the test result disparity map and the real disparity map in the whole image. From the information in Fig. 6, it can be found that when the color threshold γ_1 is between 20 and 21, the efficiency of image matching is better, and when the threshold is further increased, there is no more obvious effect. The distance threshold between 17 and

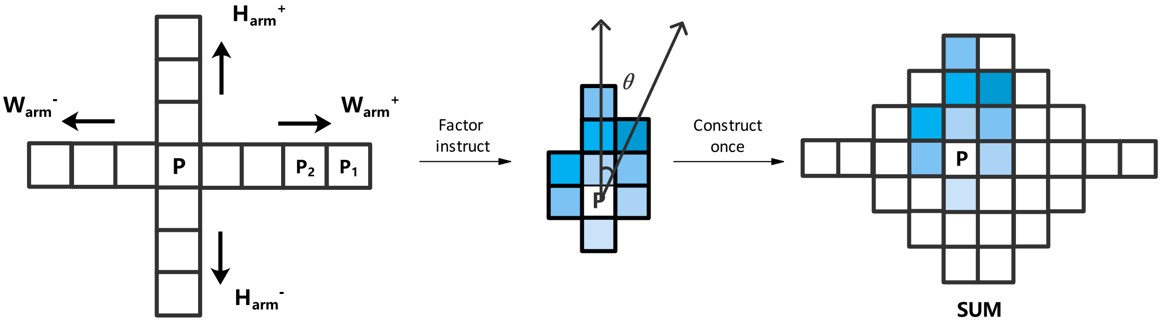
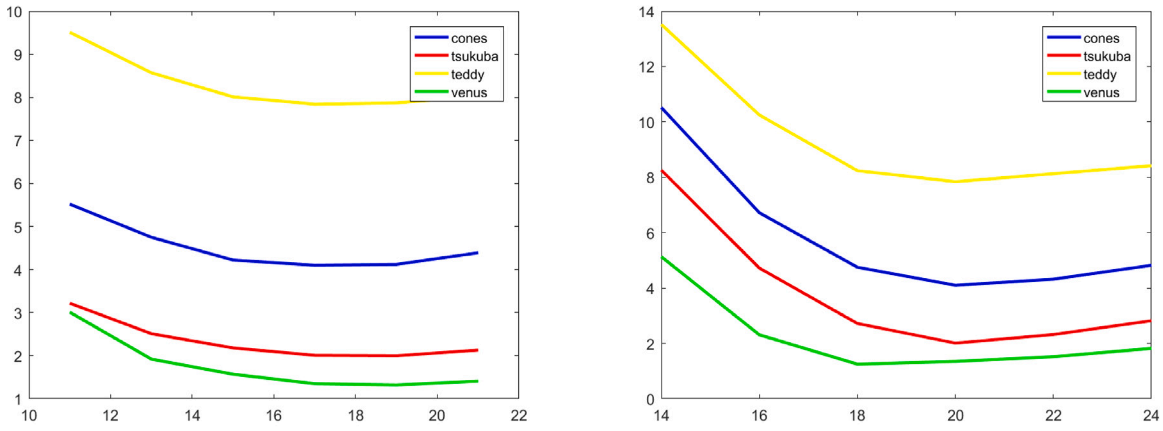


Fig. 5. The process of constructing the multi-directional aggregation template. For a cross template that each pixel may extend into at a single time (left), we adopt orientation factors to instruct the growth of template (middle), and then construct the final aggregation template once (right) for the target pixel.



(a) Mismatch Rate with Color Threshold

(b) Mismatch Rate with Distance Thresholds

Fig. 6. Mismatch rate with different threshold. The experiment is carried out from the (a) image: color threshold and the (b) image: distance threshold.

21 has a good matching performance of the image, starting from 17, and the increase of distance threshold L_2 has no great impact on the results. Therefore, this paper will control the color threshold at 20 to 21, the distance threshold L_2 choose 17 or 18.

4.3. Experiments of AMDCNet versus some non-deep learning methods

In order to prove the effectiveness of AMDCNet, we display the processed images, using Cones, Tsukuba, Teddy, and Venus in the dataset. Fig. 7 shows qualitative results of AMDCNet versus Census on Cones, Tsukuba, Teddy, and Venus (top to bottom). According to the results, AMDCNet can effectively calculate and generate the disparity of the original image, and the matching effect is better than that of classical Census method. Moreover, Fig. 8 shows qualitative results of visual sensitive area retention of AMDCNet versus classical Census method [29]. Through the display of subjective images, we can see that AMDCNet can effectively calculate and generate the disparity of images, and produce better matching effect than classical Census method [29]. Moreover, it has been improved and preserved in many visual sensitive areas, such as “peak”, “lamp” and “table”.

Next, this paper combines the objective data to prove the effectiveness of AMDCNet. Table 1 shows the error matching rate of the above four types of images. These values are obtained by comparing non-occluded areas. The non-occluded area refers to the image under two viewing angles. There must be a certain area that is invisible from the other viewing angle, which indicates that a certain displacement exists in the horizontal direction. Such invisible areas are called non-occluded areas. In general, the matching effect of comparison of non-occluded areas is better than the matching effect of all image areas. As shown in Table 1, in the non-occluded area, the matching performance of AMDCNet is better than the others.

Table 1
The error matching rate obtained by various methods.

Algorithms	cones	tsukuba	teddy	venus	avg%
Census [29]	21.89	27.41	23.43	27.78	25.13
AMDCNet	4.08	2.01	7.82	1.34	3.80
Ref. [30]	4.04	2.53	7.57	1.60	3.93
SG-C [47]	12.92	4.80	8.05	1.91	6.92
Mp-C [37]	6.78	4.50	11.32	3.55	6.53

Table 2
Mismatch rate under different Gaussian noise concentrations.

Noise	none	2%	5%	10%	15%
Ref. [33]	4.54	6.67	11.73	26.88	48.63
AMDCNet	3.81	5.01	8.59	17.32	38.41

AMDCNet is also more robust than other types of methods. In order to prove this, we add Gaussian noise of different concentrations to the four original images and show the comparative results in Table 2, where “None” means that there is no input noise. AMDCNet in the table also uses an improved adaptive aggregation template. In addition, the results in the table show that when the noise concentration increases, the change degree of the error matching rate of this method is lower, and the sensitivity to noise is lower.

For the above-processed images, AMDCNet optimizes these images by left–right consistency detection in order to make the test parallax value closer to the real value. The picture effect of the above four kinds of images after optimization is shown in Fig. 9. From the figure, we

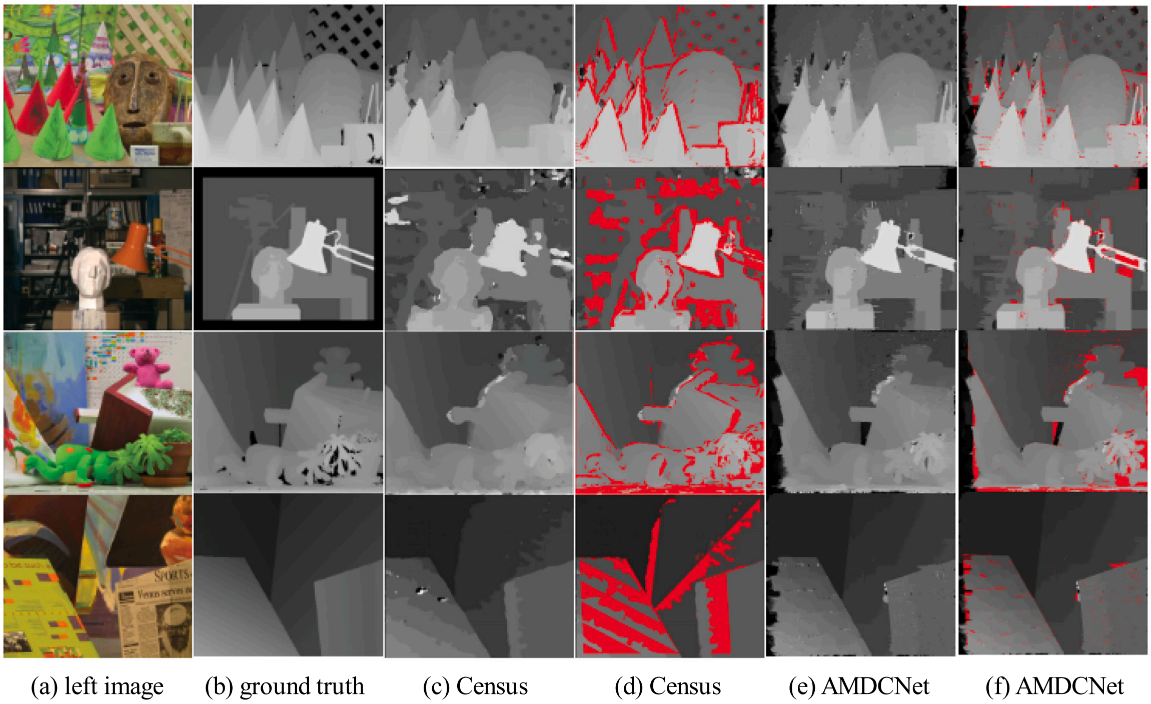


Fig. 7. Results of disparity estimation for Middlebury test images. (a) left image of stereo image pair, (b) ground truth disparity. (c) and (d) disparity map estimated using Census. (e) and (f) disparity map estimated using AMDCNet. The error area is shown in red. The more red, the worse the matching effect.

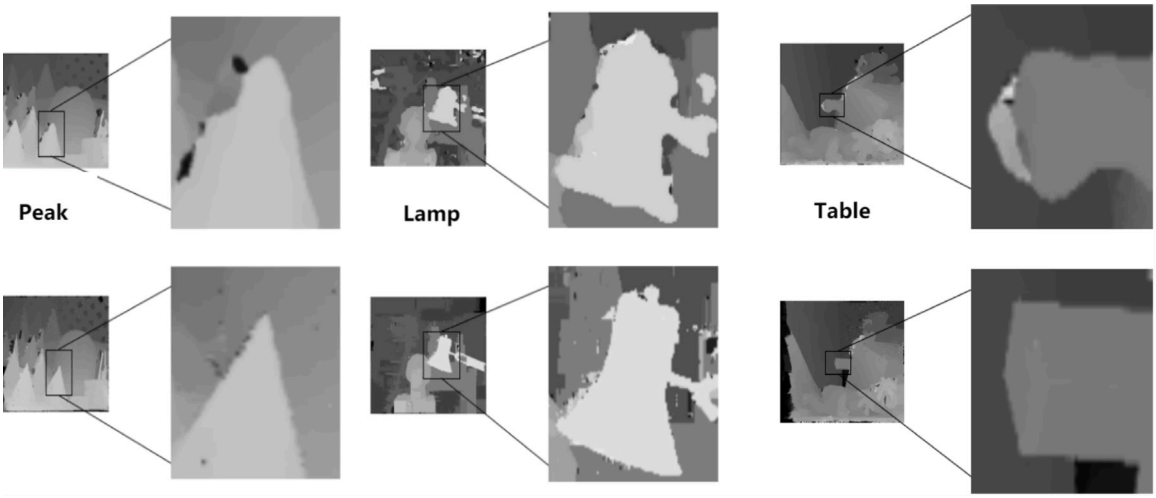


Fig. 8. Visual sensitive area retention. It shows the effect of the method for the visual sensitive area in the image, and compares with other methods. top: Census method, bottom: AMDCNet.

Table 3
Comparison results of different methods.

Algorithms	cones		tsukuba		teddy		venus		avg%
	all	nocc	all	nocc	all	nocc	all	nocc	
AMDCNet	8.85	3.46	1.26	0.92	10.72	6.09	0.86	0.34	4.06
Ref. [31]	9.26	3.79	4.15	3.62	11.6	5.68	1.87	1.08	5.13
Seg-CT [48]	13.77	6.15	5.40	4.57	11.44	6.27	1.93	1.19	6.34
SAD+CT [33]	7.65	4.09	1.98	1.29	12.60	6.02	0.74	0.53	4.36
GRD [37]	16.13	4.45	3.59	2.72	17.56	7.45	4.12	1.68	7.21
MCADSR [38]	11.1	3.51	4.15	3.62	14.7	7.57	0.87	0.48	5.75
GradAdaptWt [39]	7.67	2.61	2.63	2.26	13.10	8.00	6.99	1.39	5.58

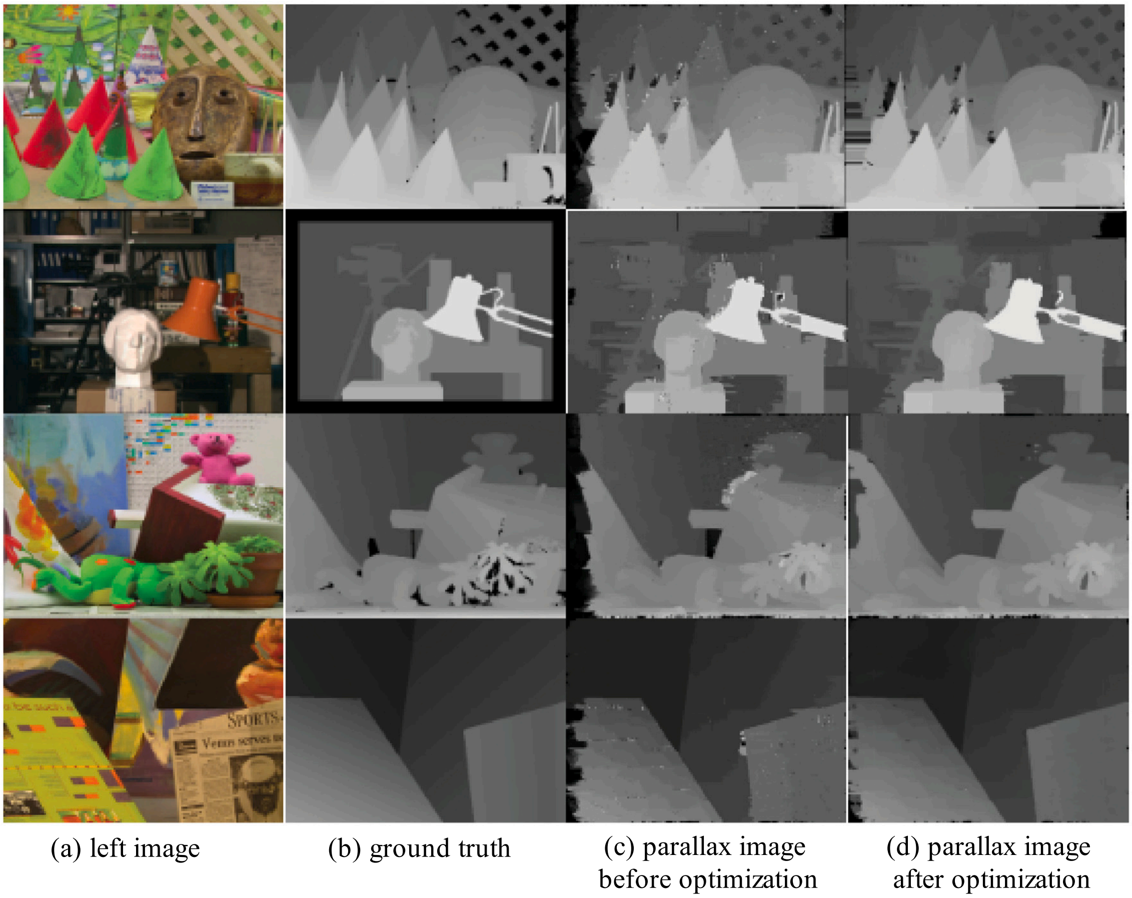

(a) left image
(b) ground truth
(c) parallax image before optimization
(d) parallax image after optimization

Fig. 9. Disparity images before and after optimization.

can see that the quality is significantly improved after optimization. After optimization, not only the large-scale "hole point" is removed for the non-occluded area, but also the isolated incorrect matching points are compensated by the quadratic linear interpolation, and the original boundary contour and other regional structures are maintained. This paper further compares the error matching rate of the optimized disparity map, and the results are shown in Table 3. We observe that GradAdaptWt's algorithm is suitable for processing and recognizing objects with shape features like cone. From Table 3, AMDCNet generally has better results, in addition to the optimization effect of algorithm GradAdaptWt on cones is better than that of AMDCNet. This is because the algorithm of GradAdaptWt is more sensitive to the shape of cone, but AMDCNet has better performance in tsukuba, teddy, venus and average values.

From Table 3, we can see the difference between NOCC and all, where the NOCC is the non-occluded area mentioned above, and "all" means the matching of all areas of the image, including the blind area of parallax. The data of mismatch rate in the table objectively proves the effectiveness of the AMDCNet, which has more efficiency than the same kind of traditional methods.

AMDCNet is not only applied to the mentioned four types of images but also experimented with other images in the dataset. The experimental images are shown in Fig. 10. From this figure, we can observe that AMDCNet produces a good matching effect for most of the binocular vision images.

4.4. Experiments of AMDCNet versus state-of-the-art deep learning methods

Next, we compare AMDCNet with some advanced deep learning methods. Wang et al. [44] proposed a multi-task attention stereo network (MASNet) to integrate the feature information from stereo image pairs for disparity estimation. Compared with the traditional CNN algorithm, AMDCNet deepens the number of network layers, embeds the pooling layer into the convolution layer, and provides a better measure for the calculation of parallax. However, this method is more inclined to the optimization process in stereo matching, and cannot provide a more detailed description for the calculation of early cost characteristics. Chang et al. [45] proposed a pyramid stereo matching network (PSMNet) consisting of two modules: spatial pyramid pooling and 3D CNN. The spatial pyramid pooling module takes the advantage of the capacity of global context information by the aggregating context in different scales and locations to form a cost volume. PSMNet is similar to the way of constructing an image pyramid, but for high-resolution images, its matching effect will be limited because the influence of the CNN convolution layer is ignored in the process of network construction. Liu et al. [46] proposed a richer convolutional feature network architecture (RCF), increasing the detection of edge features, having a good matching effect for high-resolution images, and using convolution mode to refine and extract features. However, this method does not have much workload for parallax optimization.

In [49], the authors proposed a local stereo matching algorithm that aims to improve the quality of stereo matching algorithm by solving the multi-label problems. Hong et al. in [50] proposed an almost near real-time local stereo matching algorithm possessing fast computational time. Similar fast computational time is achieved in [51], that performs depth enhancement using RGB-D guided filtering. In [52], the authors have described a generic framework for parallelizing dense matching procedures using CUDA (Compute Unified Device Architecture) programming. Furthermore, the idea of fast guided image filtering is not restricted to images, but extended to real-time 3-dimensional (3-D) video services with reduced computational complexity [53].

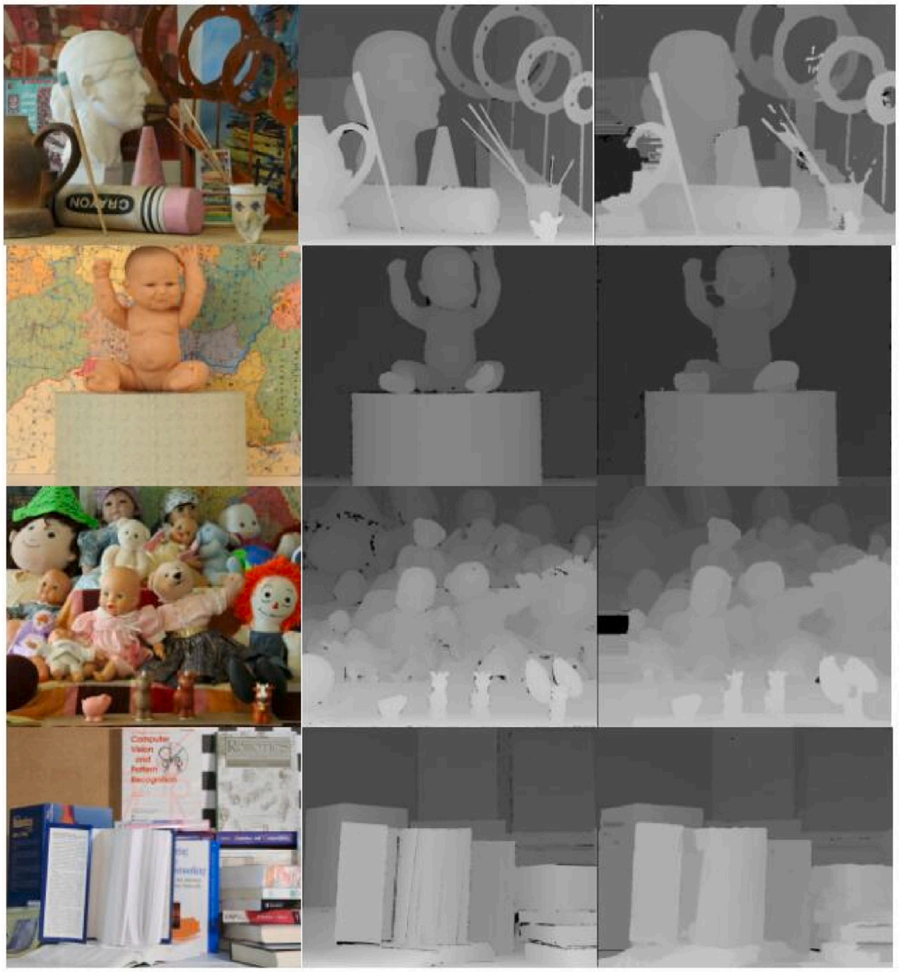


Fig. 10. Performance evaluation using Middlebury test data. (a) left image of stereo image pair, (b) ground truth disparity, and (c) disparity map estimated using AMDCNet. Top to bottom: four types of images are art, baby, dools and book.

Table 4
Matching precision of different methods before and after optimization in non-occluded areas.

Algorithms	cones		tsukuba		teddy		venus		avg%
	No-refine	refined	No-refine	refined	No-refine	refined	No-refine	refined	
AMDCNet	4.10	3.46	2.01	0.92	7.84	6.09	1.35	0.34	3.26
MASNet [44]	4.85	3.25	2.58	1.04	8.04	4.59	1.86	0.64	3.35
PSMNet [45]	4.21	3.98	1.86	0.99	9.51	8.45	0.89	0.29	3.77
RCF [46]	3.97	3.89	2.13	1.29	6.84	6.02	1.73	0.61	3.35

4.4.1. Middlebury

AMDCNet is compared with these advanced algorithms from before optimization and after optimization. The targeted area is non-occluded. The comparison of the error matching rate is shown in Table 4. It can be seen from the table that before optimization, the effect of stereo matching depends more on the description of detailed features. The false matching rate of MASNet is higher than AMDCNet. However, after adopting the optimization algorithm, the optimization effect is slightly better for some images. Although the optimization algorithm is better than that of AMDCNet, it needs to learn the features in advance, AMDCNet can be optimized directly through the left–right consistency of features, and the efficiency is relatively high. Through the image Teddy, we can see that the effect of PSMNet will be significantly reduced in high-resolution images, which shows that its general performance is poor. While the RCF has certain advantages in feature convolution learning before optimization, but its optimization stage is carried out by using the optimization algorithm of AMDCNet. Therefore, comprehensively speaking, AMDCNet is better than other methods. Moreover, AMDCNet does not need to rely on high hardware requirements, reduces the feature learning stage, and the equalization ability is more prominent. Fig. 11 shows the results of disparity estimation for Middlebury test images. It can be seen from the baby figure that the results produced by PSMNet show that the baby's arm and head are obviously connected, and its treatment of detail effect is not good. Although the results obtained by RCF can better retain the details, thanks to its relatively perfect convolution network, it has lost a certain optimization effect, resulting in many unnecessary "details" being exposed, thus reducing the overall matching effect. Other images have similar phenomena. Combined with these comparative experiments, it can be seen that although the method used in deep learning can greatly

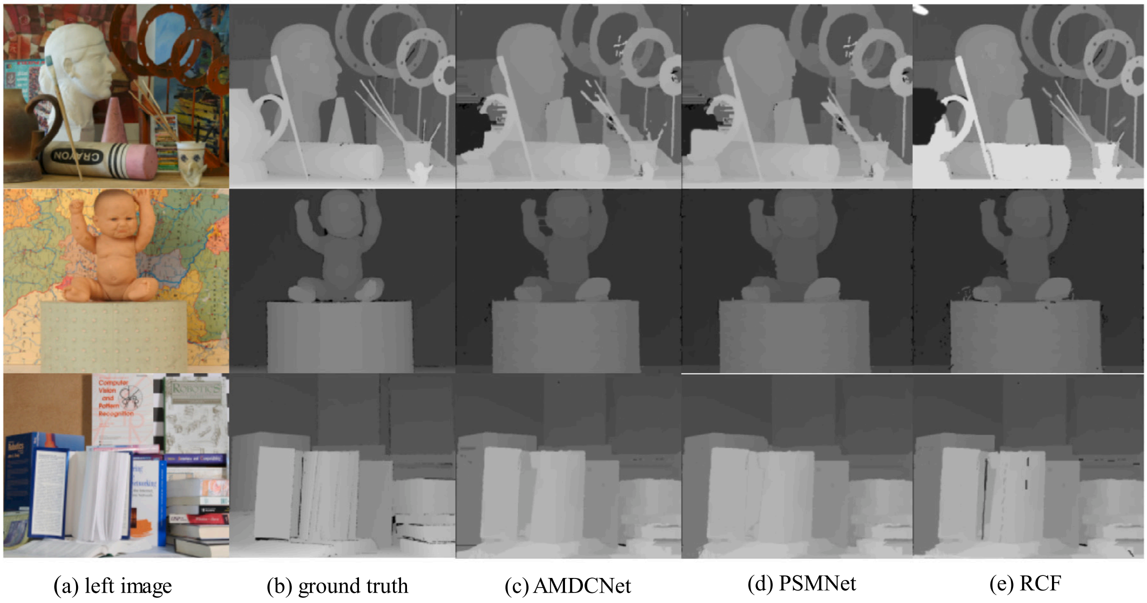


Fig. 11. Results of disparity estimation for Middlebury test images. (a) left image of stereo image pair. (b) ground truth disparity. For each input image, the disparity maps obtained by (c) AMDCNet, (d) PSMNet, and (e) RCF.

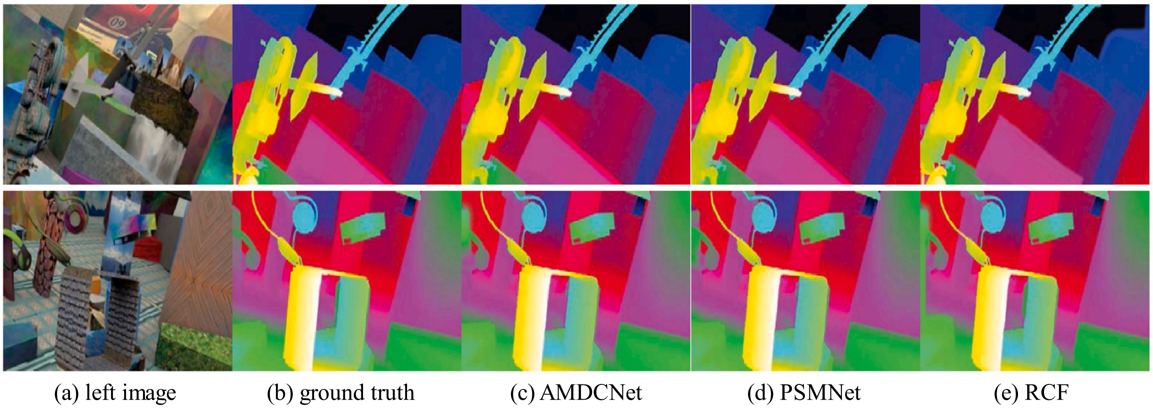


Fig. 12. Results of disparity estimation for Scene Flow test images. (a) left image of stereo image pair. (b) ground truth disparity. For each input image, the disparity maps obtained by (c) AMDCNet, (d) PSMNet, and (e) RCF.

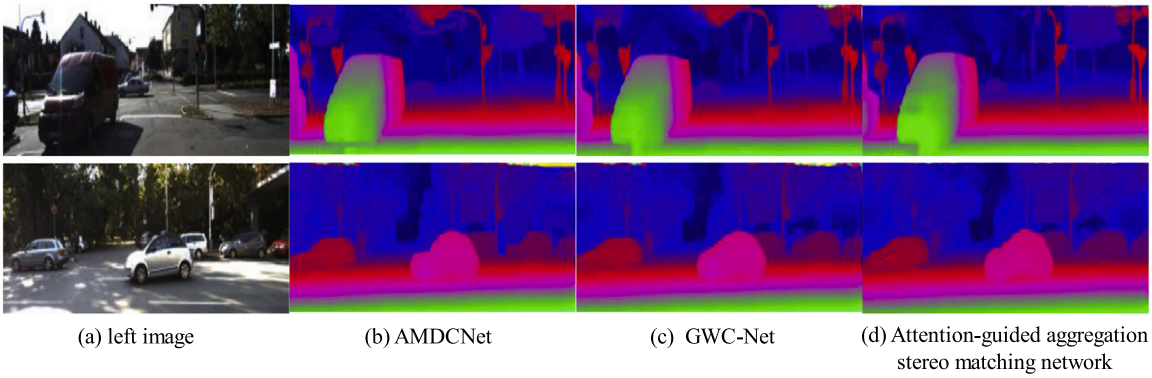


Fig. 13. Results of disparity estimation for KITTI 2015 test images. (a) left image of stereo image pair. For each input image, the disparity maps obtained by (b) AMDCNet, (c) GWC-Net, and (d) Attention-guided aggregation stereo matching network [43].

improve the matching effect, it will have some disadvantages more or less. For stereo matching technology, balance is often more important. Therefore, AMDCNet not only maintains good structural performance but also reduces the false matching rate. As it exhibits a satisfactory performance balance, there is no need to preprocess the image such as feature learning in advance, nor to build a complex network architecture trained on large-scale stereo-matching datasets. Most existing deep learning methods require training on different datasets for different instances to achieve acceptable performance, while AMDCNet gains a good performance on all kinds of instances without complex training and validation procedure.

4.4.2. Scene flow

We also use two images in the Scene Flow datasets for experiments, which are consistent with the previous figures. Fig. 12 shows the results of experiments. By analyzing the qualitative results of the figure, we can see that the optimized effect of AMDCNet is not as good as PSMNet, and the efficiency of edge matching is slightly lower than RCF, but AMDCNet has more advantages in a balance between optimization effect and efficiency of edge matching.

4.4.3. KITTI 2015

GWC-Net [54] proposes a high-dimensional feature grouping description method, which forms the high-dimensional feature description of two images by connecting multi-level unitary features, and groups them according to the dimension correlation to find the grouping feature correlation at the corresponding position of the two images. The method Attention-guided aggregation stereo matching network in Zhang et al. [43] further improves it and leads to the four-dimensional cost measurement, further guiding the distinction and connection of features. This kind of method can provide great support for the cost calculation stage. In the aggregation stage, the pyramid is constructed according to the scale pyramid, different scales are divided, and a cross-scale aggregation template is constructed for aggregation operation. However, it also has some disadvantages, for instance, its level is basically window level, and there is a lack of consideration for the accurate pixel level. Therefore, for image matching in complex scenes, there will be errors in the local detail effect.

In order to prove the universal applicability of AMDCNet, the current hot KITTI 2015 dataset is selected for experiments, and compared with the above two advanced algorithms to analyze the advantages and disadvantages of the algorithm from a qualitative point of view, as shown in Fig. 13. It can be seen from the experimental results in the first row that AMDCNet can display the contour of the car more comprehensively, and the parallax acquisition effect of street lamps is better. From the third and fourth columns, the detailed description effect of GWC-Net is significantly better than the Attention-guided aggregation stereo matching network. The second row of images is consistent with it and it can be seen that the parallax obtained by this method is more accurate and the detail display effect is better for closer scenes, but there are some deficiencies in the matching effect for areas close to the left and right boundaries that areas prone to occlusion and areas with far-field of view. However, the judgment of stereo matching efficiency is more for the more obvious regions. Therefore, AMDCNet is more stable.

5. Conclusions and future work

This paper studies the stereo matching technology in computer vision and identifies a few shortcomings in the existing methods. We propose AMDCNet for stereo matching that is based on sensitivity value and multi-directional template. In the multi-directional template matching stage, the sensitive factor is used as guidance for constructing a multi-directional extension template and calculating the second generation value whereas the direction is used to develop the construction process of the aggregation template. The effectiveness of AMDCNet is demonstrated via benchmarking experiments. It mainly uses the visual display in the image to analyze the regional structure retention and the matching performance. Finally, the experimental results show that AMDCNet can effectively deal with the matching image and it is very effective as compared with other relevant methods.

For future work, we intend to perform subjective evaluation on large-scale datasets such as KITTI and Scene Flow datasets. We will compare our proposed AMDCNet approach with other benchmarking guided filtering approaches. We also intend to further study the deep learning methods on stereo matching and its performance on the KITTI dataset. Especially, we will focus on the reuse and fusion of visual transformer on specific binocular stereo matching tasks, model design for binocular image inputs, the research on new encoder and decoder for depth output and the loss function together with training hyperparameters.

CRediT authorship contribution statement

Hewei Wang: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Yijie Li:** Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Shijia Xi:** Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Shaofan Wang:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Resources. **Muhammad Salman Pathan:** Writing – original draft, Writing – review & editing, Resources. **Soumyabrata Dev:** Conceptualization, Supervision, Writing – original draft, Writing – review & editing, Project administration, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at University College Dublin. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme. This research is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 801522, by Science Foundation Ireland and co-funded by the European Regional Development Fund through the ADAPT Centre for Digital Content Technology grant number 13/RC/2106_P2.

References

[1] D. Wan, J. Zhou, Stereo vision using two PTZ cameras, Comput. Vis. Image Underst. 112 (2) (2008) 184–194.

[2] W. Shi, Rajkumar Ragunathan, Point-GNN: Graph neural network for 3D object detection in a point cloud, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020.

[3] Z. Zhu, D.R. Karuppiah, E.M. Riseman, A.R. Hanson, Dynamic mutual calibration and view planning for cooperative mobile robots with panoramic virtual stereo vision, Comput. Vis. Image Underst. 95 (3) (2004) 261–286.

[4] Z. Al-Makhadmeh, A. Tolba, Dependable information processing method for reliable human–robot interactions in smart city applications, Image Vis. Comput. 104 (2) (2020) 104045.

[5] F. M. Savoy, S. Dev, Y. H. Lee, S. Winkler, Stereoscopic cloud base reconstruction using high-resolution whole sky imagers, in: Proc. IEEE International Conference on Image Processing (ICIP), 2017.

[6] F. M. Savoy, S. Dev, Y. H. Lee, S. Winkler, Geo-referencing and stereo calibration of ground-based whole sky imagers using the sun trajectory, in: 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2016.

[7] J. Fu, S. Pertuz, J. Matas, J.-K. Kämäräinen, Performance analysis of single-query 6-dof camera pose estimation in self-driving setups, Comput. Vis. Image Underst. 186 (2019) 58–73.

[8] C. Hane, L. Heng, G.H. Lee, F. Fraundorfer, P. Furgale, T. Sattler, M. Pollefeys, 3D visual perception for self-driving cars using a multi-camera system: Calibration, mapping, localization, and obstacle detection, Image Vis. Comput. 68 (2017) 14–27.

[9] M.S. Pathan, A. Nag, M.M. Pathan, S. Dev, Analyzing the impact of feature selection on the accuracy of heart disease prediction, Healthcare Analytics 2 (2022) 100060.

[10] S. Dev, H. Wang, C.S. Nwosu, N. Jain, B. Veeravalli, D. John, A predictive analytics approach for stroke prediction using machine learning and neural networks, Healthcare Analytics 2 (2022) 100032.

[11] C.S. Nwosu, S. Dev, P. Bhardwaj, B. Veeravalli, D. John, Predicting stroke from electronic health records, in: Proc. 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019.

[12] I. Bacher, H. Javidnia, S. Dev, R. Agrahari, M. Hossari, M. Nicholson, C. Conran, J. Tang, P. Song, D. Corrigan, F. Pitié, An advert creation system for 3d product placements, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2020.

[13] I. Bacher, H. Javidnia, S. Dev, R. Agrahari, M. Hossari, M. Nicholson, C. Conran, D. Corrigan, F. Pitié, Deepreal-a deep learning based 3d advert integration system, in: Proc. NEM Summit 2020 - B'Smart - European Media Science and Technology Meets Arts, 2020.

[14] A. Nautiyal, K. McCabe, M. Hossari, S. Dev, M. Nicholson, C. Conran, D. McKibben, J. Tang, W. Xu, F. Pitié, An advert creation system for next-gen publicity, in: Proc. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2018.

[15] M. Hossari, S. Dev, M. Nicholson, K. McCabe, A. Nautiyal, C. Conran, J. Tang, W. Xu, F. Pitié, Adnet: a deep network for detecting adverts, 2018, arXiv preprint arXiv:1811.04115.

[16] S. Dev, M. Hossari, M. Nicholson, K. McCabe, A. Nautiyal, C. Conran, J. Tang, W. Xu, F. Pitié, Localizing adverts in outdoor scenes, in: Proc. IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2019.

[17] S. Dev, F. M. Savoy, Y. H. Lee, S. Winkler, Short-term prediction of localized cloud motion using ground-based sky imagers, in: Proc. IEEE Region 10 Conference (TENCON), 2016.

[18] S. Dev, Y. H. Lee, S. Winkler, Multi-level semantic labeling of sky/cloud images, in: Proc. IEEE International Conference on Image Processing (ICIP), 2015.

[19] Y. Jin, M. Lee, Enhancing binocular depth estimation based on proactive perception and action cyclic learning for an autonomous developmental robot, IEEE Trans. Syst. Man Cybern. 49 (1) (2019) 169–180.

[20] H.H. Nagel, F. Heimes, K. Fleischer, Quantitative comparison between trajectory estimates obtained from a binocular camera setup within a moving road vehicle and from the outside by a stationary monocular camera, Image Vis. Comput. 18 (5) (2000) 435–444.

[21] D. Marr, T. Poggio, Cooperative computation of stereo disparity, Science 194 (4262) (1976) 283–297.

[22] O. Stankiewicz, G. Lafruit, M. Domański, Multiview video: Acquisition, processing, compression, and virtual view rendering, Academic Press Library in Signal Processing, 2018.

[23] E.P. Simoncelli, E.H. Adelson, D.J. Heeger, Probability distribution of optical flow, IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. (1991).

[24] F.P. Da, F. He, Z.W. Chen, Stereo matching based on dissimilar intensity support and belief propagation, J. Math. Imaging Vis. 47 (1–2) (2013) 27–34.

[25] H. Wang, M. S. Pathan, S. Dev, Stereo matching based on visual sensitive information, in: Proc. 6th International Conference on Image, Vision and Computing (ICIVC), 2021.

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

[27] C.L. Yin, D.M. Liu, J.Z. Song, An improved stereo matching algorithm based on image segmentation, J. Comput. Aided Des. Graph. 20 (6) (2008) 808–812.

[28] W. Fei, K. Jia, J. Feng, The real-time depth map obtainment based on stereo matching, in: Euro-China Conference on Intelligent Data Analysis & Applications, 2016.

[29] R. Zabih, J. Woodfill, Non-parametric local transforms for computing visual correspondence, in: Proceedings of European Conference on Computer Vision, 1994.

[30] H.R. Fan, F. Yang, X. Pan, An improved census transform and gradient fusion stereo matching algorithm, J. Opt. (2018) 267–277.

[31] Y. Men, G. Zhang, C. Men, L. Xiang, M. Ning, A stereo matching algorithm based on four-moded census and relative confidence plane fitting, Chin. J. Electr. 24 (4) (2015) 807–812.

[32] S. Guo, P. Xu, Y. Zheng, Semi-global matching based disparity estimate using fast census transform, in: International Congress on Image & Signal Processing, 2016.

[33] Y. Chai, X. Cao, Stereo matching algorithm based on joint matching cost and adaptive window, in: 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference, IAEAC, 2018.

[34] K.L. Zhang, G. Lafruit, Cross-based local stereo matching using orthogonal integral images, IEEE Trans. Circuits Syst. Video Tech. 19 (7) (2009) 1073–1079.

[35] M. Xing, S. Xun, M. Zhou, S. Jiao, H. Wang, On building an accurate stereo matching system on graphics hardware, in: IEEE International Conference on Computer Vision Workshops, 2012.

[36] J.C. Kwak, T.R. Park, S.K. Yong, K.Y. Lee, Implementation of improved census transform stereo matching on a multicore processor, in: Multimedia and Ubiquitous Engineering, Springer, 2013, pp. 989–995.

[37] Z. Kang, Y. Fang, D. Min, L. Sun, T. Qi, Cross-scale cost aggregation for stereo matching, IEEE Trans. Circuits Syst. Video Technol. 27 (5) (2017) 965–976.

[38] Y. Shan, Y.C. Hao, W.Q. Wang, Y. Wang, X. Chen, H.Z. Yang, W. Luk, Hardware acceleration for an accurate stereo vision system using mini-census adaptive support region, ACM Trans. Embed. Comput. Syst. (TECS) 13 (4) (2014) 1–24.

[39] L. De-Maeztu, A. Villanueva, R. Cabeza, Stereo matching using gradient similarity and locally adaptive support-weight, Pattern Recognit. Lett. 32 (13) (2011) 1643–1651.

[40] E. Rublee, V. Rabaud, K. Konolige, An efficient alternative to sift or surf, IEEE Int. Conf. Comput. Vis. (2012) 2564–2571.

[41] N. Tekin, K.A. Peker, Matching day and night location images using sift and logistic regression, in: 2015 23th Signal Processing and Communications Applications Conference, SIU, 2015.

[42] J. Su, Q. Xu, J. Zhu, A scene matching algorithm based on surf feature, in: International Conference on Image Analysis & Signal Processing, 2010.

[43] Y. Zhang, Y. Li, C. Wu, B. Liu, Attention-guided aggregation stereo matching network, Image Vis. Comput. 106 (2020) 104088.

[44] J. Wang, S. Zhang, Y. Wang, Z. Zhu, Learning efficient multi-task stereo matching network with richer feature information, Neurocomputing 421 (2021) 151–160.

[45] J.R. Chang, Y.S. Chen, Pyramid stereo matching network, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2018.

[46] Y. Liu, M.M. Cheng, X. Hu, K. Wang, X. Bai, Richer convolutional features for edge detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3000–3009.

[47] G. Z, M. Yubo, C. M, A stereo matching algorithm based on four-moded census and relative confidence plane fitting, Chin. J. Electr. 24 (4) (2015) 807–812.

[48] E.T. Baek, Y.S. Ho, Cost aggregation with guided image filter and superpixel for stereo matching, in: Signal & Information Processing Association Summit & Conference, 2017.

[49] G.-S. Hong, B.-G. Kim, A local stereo matching algorithm based on weighted guided image filtering for improving the generation of depth range images, Displays 49 (2017) 80–87.

[50] G.-S. Hong, J.-K. Park, B.-G. Kim, Near real-time local stereo matching algorithm based on fast guided image filtering, in: 2016 6th European Workshop on Visual Information Processing, EUVIP, IEEE, 2016, pp. 1–5.

[51] T.-W. Hui, K.N. Ngan, Depth enhancement using RGB-d guided filtering, in: 2014 IEEE International Conference on Image Processing, ICIP, IEEE, 2014, pp. 3832–3836.

[52] G.-S. Hong, W. Hoe, B.-G. Kim, Performance analysis of matching cost for stereo matching with CUDA, in: Computer Science and its Applications, Springer, 2015, pp. 623–629.

[53] G.-S. Hong, B.-G. Kim, Stereo matching algorithm based on fast guided image filtering for 3-dimensional video service, J. Digit. Contents Soc. 17 (6) (2016) 523–529.

[54] X. Guo, K. Yang, W. Yang, X. Wang, H. Li, Group-wise correlation stereo network, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020.