

Adaptive Kernel Convolutional Stereo Matching Recurrent Network

Jiamian Wang ^{1,2}, Haijiang Sun ^{1,*} and Ping Jia ^{1,2}

¹ Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China; wangjiamian20@mailsucas.ac.cn (J.W.); jiap@ciomp.ac.cn (P.J.)

² University of Chinese Academy of Sciences, Beijing 100049, China

* Correspondence: sunhj@ciomp.ac.cn; Tel.: +86-135-7868-7727

Abstract: For binocular stereo matching techniques, the most advanced method currently is using an iterative structure based on GRUs. Methods in this class have shown high performance on both high-resolution images and standard benchmarks. However, simply replacing cost aggregation with a GRU iterative method leads to the original cost volume for disparity calculation lacking non-local geometric and contextual information. Based on this, this paper proposes a new GRU iteration-based adaptive kernel convolution deep recurrent network architecture for stereo matching. This paper proposes a kernel convolution-based adaptive multi-scale pyramid pooling (KAP) module that fully considers the spatial correlation between pixels and adds new matching attention (MAR) to refine the matching cost volume before inputting it into the iterative network for iterative updates, enhancing the pixel-level representation ability of the image and improving the overall generalization ability of the network. At present, the AKC-Stereo network proposed in this paper has a higher improvement than the basic network. On the Sceneflow dataset, the EPE of AKC-Stereo reaches 0.45, which is 0.02 higher than the basic network. On the KITTI 2015 dataset, the AKC-Stereo network outperforms the base network by 5.6% on the D1-all metric.

Keywords: stereo matching; GRU; adaptive; matching attention



Citation: Wang, J.; Sun, H.; Jia, P. Adaptive Kernel Convolutional Stereo Matching Recurrent Network. *Sensors* **2024**, *24*, 7386. <https://doi.org/10.3390/s24227386>

Academic Editor: Benoit Vozel

Received: 7 October 2024

Revised: 11 November 2024

Accepted: 18 November 2024

Published: 20 November 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Binocular stereo matching has always been a focal point in research on stereo vision, emulating human binocular vision to recognize 3D information within the field of view [1]. The binocular camera takes two viewpoint images of the same scene and estimates the disparity by finding the horizontal correspondence between the corrected left and right images [2], and then obtains the depth map. Depth map has a wide range of applications. Because it can record the distance between objects in the scene and the camera, it has a wide range of applications in robot navigation [3,4], visual basic models [5], smart healthcare [6], 3D measurement [7], and augmented reality [8,9]. In recent years, with the support of large synthetic datasets [10–12], stereo matching methods based on Convolutional Neural Networks (CNNs) have elevated the accuracy of disparity estimation to a new level [13–15]. However, due to various real-world challenges, achieving both high accuracy and efficiency is crucial for practical applications.

Traditional stereo matching algorithms [16–18] typically involve four steps: computing the matching cost, cost aggregation, disparity computation, and disparity refinement [1]. And the matching process can be divided into two types: global matching methods and local matching methods. Since the high-efficiency local algorithms [19] are based on the assumption that the disparity of local windows is the same, which is not true in many cases, it will lead to poor matching performance. The global algorithms with better performance [20,21], although achieving better matching results through constraints such as smoothness constraints between two-dimensional adjacent pixel disparities, have a larger memory footprint and slower speed. In this case, in order to combine the advantages of the above two methods while avoiding their disadvantages, the semi-global matching

method [22,23] was proposed. The disparity map obtained through the semi-global matching method was not significantly different in effect from the global algorithm, but there was a significant improvement in algorithm efficiency, making semi-global matching a widely used method at the time. However, in open-world environments, traditional methods like these still struggle with issues such as image noise, uneven lighting, and large textureless areas. It is therefore essential to develop targeted solutions to effectively handle real-world scenarios, particularly outdoor images.

In recent years, with the development of deep learning, each step in traditional stereo matching methods can be replaced by deep neural networks to improve performance [24]; thus, learning-based methods have been widely used in the field of binocular stereo matching. Compared to traditional methods, learning-based approaches produce more accurate and smoother disparity maps [14,15] and offer a slight advantage in computation speed [14,25]. The first end-to-end stereo matching network, DispNetC [11], computes a correlation volume from the features of the left and right images and uses a CNN to directly regress the disparity map. Since then, various learning-based methods have been proposed, and cost volume and correlation volume have been introduced into the network structure. GCNet [26] proposes using cascaded features to construct cost volumes and employing 3D convolution for cost aggregation. GA-Net [27] notes the drawbacks of 3D convolution and designs a semi-global aggregation layer and a local-guided aggregation layer, and tries to replace it with these two guided aggregation layers to further improve the accuracy. Meanwhile, GwcNet [28] proposes a group-wise correlation cost volume to enhance similarity measurements.

As mentioned above, most current algorithms use 2D CNNs to extract features, forming a cost volume that is subsequently fed into cost aggregation and regularization modules composed of 2D or 3D CNNs to compute the final disparity map. However, the aggregated information is often redundant [29], because it just concatenates all the information on a large number of feature dimensions. It does not require so much information to predict the disparity, and too much redundant information will have a bad impact on the network [30]. In fact, with the advancement of camera technology, the quality and resolution of initial images have greatly improved, leading to a significant amount of high-frequency information being disregarded during the aforementioned computation processes. This results in relatively blurry disparity maps lacking in detail, leading to insufficient matching accuracy and blurred depth estimation for the details. Consequently, the above algorithms are no longer compatible with the rapid advancements in current camera technology.

To adapt to higher-resolution images, the current state-of-the-art stereo methods [15,31–33] use an iterative structure based on ConvGRU [34] and have shown higher performance on both high-resolution images and standard benchmarks. Different from the above methods, the iterative approaches bypass the computationally expensive cost aggregation operations by only calculating the visual similarity between pixels at the same height. This substitutes the four-dimensional cost volume with a three-dimensional cost volume, while introducing multi-stage GRUs to progressively update the disparity map. This solution allows for the direct computation of cost volumes using high-resolution images, enhancing the ability of the update operators to propagate information across the image, and making it more suitable for high-resolution images. For instance, RAFT-Stereo [15] utilizes multi-layer convolutional gated recurrent units (ConvGRUs) [34] to cyclically update the disparity map using local cost volume retrieved from All-Pairs Correspondences (APCs); IGEV-Stereo [35] constructs a combined geometrically encoded volume to encode geometric and contextual information as well as local matching details, which is iteratively indexed to update disparity to make full use of all information. However, using iterative methods in place of cost aggregation results in the original cost volume lacking spatial-level information, and failing to better exploit the inherent geometric advantages present in stereo images. This leads to difficulties in handling pathological cases characterized by local blurriness, such as thin-structured objects (e.g., fences, wires) and textureless regions (e.g., highly reflective surfaces like glass). Another challenge is that due to their limited generalization

ability, they are often less effective when applied to real-world scenarios than on specific datasets [36].

Therefore, the purpose of this article is to explore a better stereo matching method to make up for the shortcomings brought by the iterative method based on the GRU structure and solve the problem of disparity estimation in the above ill-posed regions. This article proposes a new design based on the IGEV-Stereo network, which can effectively improve the accuracy of the stereo matching network, extract features more accurately, make full use of the effective information of the images, and process thin structures and textureless areas more perfectly. The main contributions can be summarized as follows:

- In the feature extraction phase, this paper proposes the Kernel Normalization Convolution Layer-based Adaptive Spatial Pyramid Pooling (KAP) module to extract multi-scale features from images. This approach takes into full account the spatial correlation between pixels, thereby enhancing the performance of the deep network. The KAP module extracts more plentiful contextual features, enabling the network to better handle issues such as overexposure, thin structures, and low-texture regions, while also allowing the network to learn error correction capabilities.
- For the cost matching computation phase, this paper believes that conventional methods of connecting two cost volumes and then performing regularized correlation volume calculations struggle to fully exploit the advantages of the two image samples. To obtain a more reliable correlation volume, this paper proposes a new attention module (MAR) that integrates the regularized cost volume to produce a more accurate similarity measure. It not only effectively utilizes the effective geometric information of the binocular images, but also reduces the number of subsequent GRU iterations.
- Furthermore, this paper also utilizes a kernel convolution-based gated recurrent unit [37] as the iterative update operator within the main network framework, which retrieves features from the previously obtained stereo matching correlation volume and regressively updates the disparity. It enables GRUs to consider wider contextual information around each pixel, promotes the fusion of local and global information, and provides a more comprehensive understanding of the feature space.

This article demonstrates the effectiveness of the proposed method on several publicly available datasets of binocular images. Due to the mature and widely used evaluation benchmarks of the KITTI dataset, the KITTI 2015 dataset is used as an example here. The AKC-Stereo network proposed in this article outperforms IGEV-Stereo [35] and CREStereo [33] by 5.6% and 11.24% on the D1-all metric in KITTI 2015 dataset [38], respectively. In the qualitative test, the AKC-Stereo in this article performs more excellently in light reflection regions and thin-structure regions, and also demonstrates better cross-dataset generalization ability. As shown in Figure 1, the accuracy of AKC-Stereo in this article is better than that of the basic network IGEV-Stereo. In Figure 1a, it can be seen that AKC-Stereo achieves a smaller EPE (that is, 0.52) using only 12 GRU iterations than IGEV-Stereo using 28 GRU iterations (that is, 0.56 EPE). In Figure 1b, it can be observed that AKC-Stereo also improves significantly for small batch training, which also helps the network to converge faster and reduce the footprint during training.

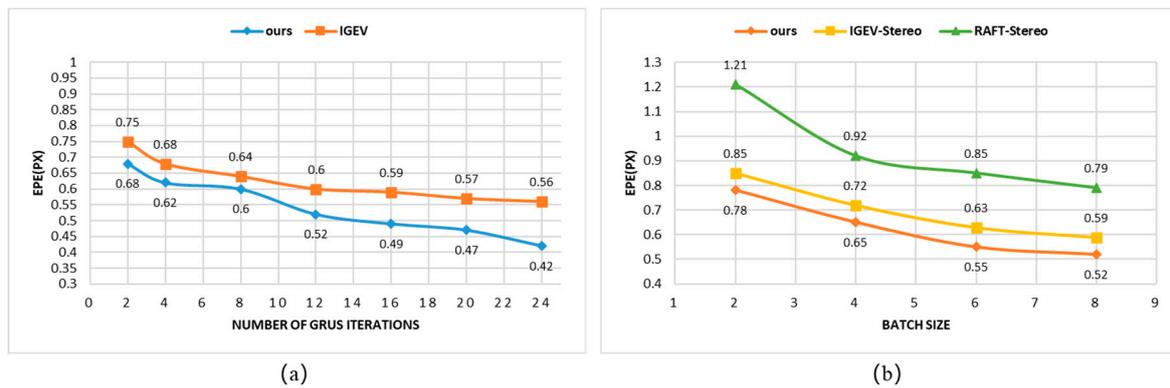


Figure 1. (a) shows the performance comparison between the AKC-Stereo network proposed in this article and the base network on the KITTI 2015 dataset with the change in the number of iterations; num_steps is set to 20,000, the abscissa is the number of iterations, and the ordinate is the EPE endpoint error data. In the figure, the orange part represents the performance of the base network, namely IGEV-Stereo, as the number of iterations changes, and the blue part represents the performance of the AKC-Stereo network as the number of iterations changes. (b) shows the effect of smaller batch size on the performance of the AKC-Stereo network proposed in this article (orange lines), RAFT-Stereo (green lines), and IGEV-Stereo (yellow lines) on the KITTI 2015 dataset with equal training rounds. The abscissa is the batch size number, and the ordinate is the EPE endpoint error data.

2. Materials and Methods

In this section, the structure of AKC-Stereo is described in detail (as shown in Figure 2), which consists of an adaptive multi-scale feature extractor KAP based on kernel-normalized convolutional layers, a matching attention MAR module that combines disparity and epipolar self-attention information, a GRU update algorithm based on kernel convolution, and other basic modules.

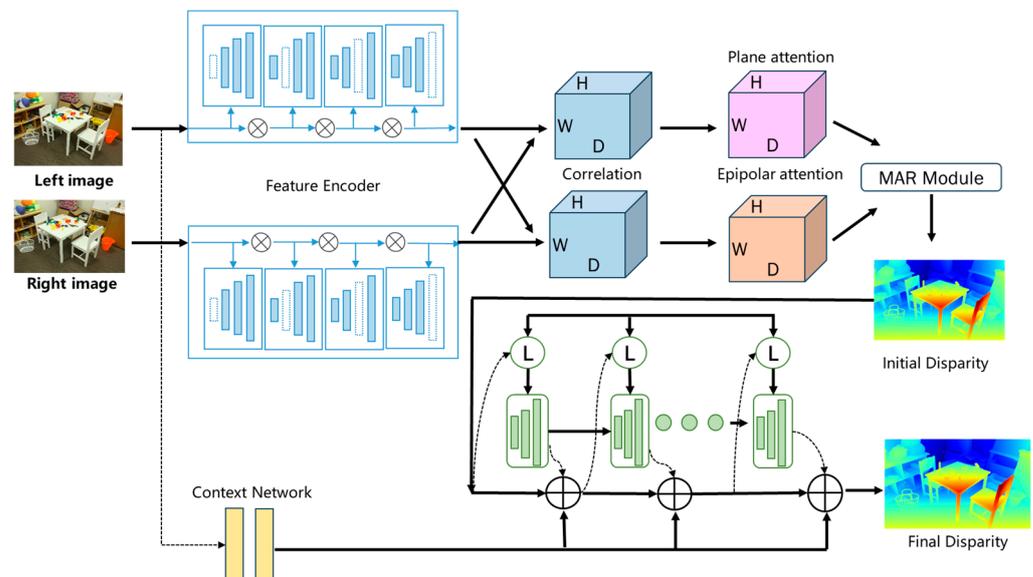


Figure 2. The overall structure diagram of the proposed AKC-Stereo. AKC-Stereo first constructs a multi-scale adaptive feature extractor (KAP), then computes the matching correlation volume by grouping correlation method, and then preliminarily refines it using the matching attention (MAR) module. Finally, the refined correlation volume is fused with the context-encoded features obtained through residual blocks, and the combined data are fed into the GRU iteration for further optimization through iterative updates.

2.1. Feature Extractor

The feature extractor consists of two parts: the context encoder, which extracts multi-scale contextual features, and the feature encoder applied to the left and right images.

The context encoder is based on IGEV-Stereo [35] and is composed of a series of residual blocks and downsampling layers. It generates multi-scale contextual features at 1/4, 1/8, and 1/16 resolutions of the input images, which have 128 channels. These contextual features are used to initialize the hidden state of the update operators and are injected into the GRU update operators during each iteration.

For the feature encoder, this paper was inspired by the success of the spatial pyramid pooling method in image segmentation technology [39]. By using dilated convolutions, multiple parallel convolutional layers with different dilation rates are employed to extract features, which are further processed and fused to produce the final result. This method allows the combination of semantic information with different receptive field sizes without sacrificing resolution (i.e., no downsampling), improving the precision of feature extraction. However, the superposition of dilated convolutions is not friendly to the disparity estimation task of dense pixels such as stereo matching, which will lose many pixel-level details of the object. To address this issue, this paper combines kernel convolution normalization and adaptive offset to create a variant of this approach, namely the KAP module, which will be described in detail in the following sections.

2.1.1. Adaptive Offset

Aiming at the problem that the dilated convolution will lose the dense features of the object, in order to obtain dense global contextual information and establish dependency between pixels, this paper extends the deformable convolutions [40] to multi-scale pyramid pooling layers. By utilizing learnable offset to create adaptive search windows for multi-scale feature extraction, we achieved adaptive sampling of the input feature maps without increasing memory and computational costs. This enables the network to dynamically adjust the receptive field according to the shape and orientation of objects, thereby capturing more relevant and discriminative features. Figure 3 shows how offset changes the formation of feature extraction boxes.

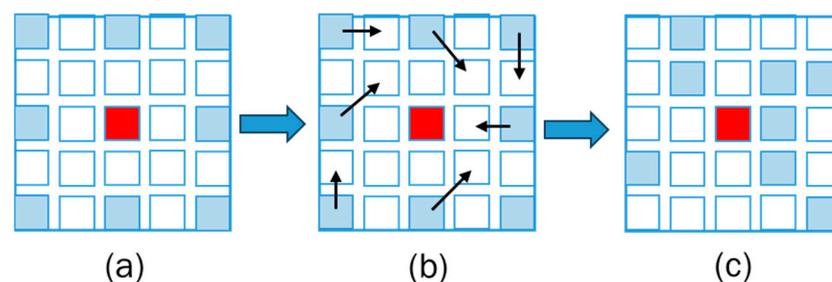


Figure 3. (a) shows the sampling search window with a normal dilation rate of 3 for feature extraction; (b) shows the positions where the sampling points should be shifted after adding offsets; (c) illustrates the more precise sampling search window obtained after the offsets are applied.

In feature extraction, $h(p_i)$ represents the weight associated with each pixel p_i ; the mapping relationship of any point p_0 in the feature map y relative to point p_i in the original image x can be expressed by Equation (1):

$$y(p_0) = \sum_{p_i \in y} h(p_i) \cdot x(p_0 - p_i) \quad (1)$$

After applying an offset Δp to the convolution, the mapping relationship is transformed into Equation (2):

$$y(p_0) = \sum_{p_i \in y} h(p_i) \cdot x(-p_i + \Delta p) \quad (2)$$

Firstly, this paper uses four parallel adaptive convolutional layers with different dilation rates (1, 3, 4, and 5) to generate more refined and comprehensive multi-scale features. Then, the extracted feature maps containing features of different scales are fused. In order to save computational cost and not lose information, only three of the feature maps obtained by convolution layers with different sampling rates were selected each time for fusion, that is, the feature maps obtained by adaptive convolution layers with expansion rates of 3, 4, and 5 are linked in block1. And they are normalized by kernel normalized convolution. In block 2, the feature maps obtained by linking adaptive convolutional layers with dilation rates of 1, 4, and 5 are also normalized using kernel normalization convolution. After that, the same process is applied to block3 and block4, using convolutional layers with dilation rates of (1, 3, 5) and (1, 3, 4), respectively. Finally, after normalization, we obtain a total of 4 blocks, which are eventually combined through residual cascading to produce the final feature map, as illustrated in Figure 4. Among them, the adaptive offset also undergoes end-to-end learning during the training process, ensuring that the network optimally adapts to the shape and scale of the target objects, thereby producing more accurate and detailed feature maps.

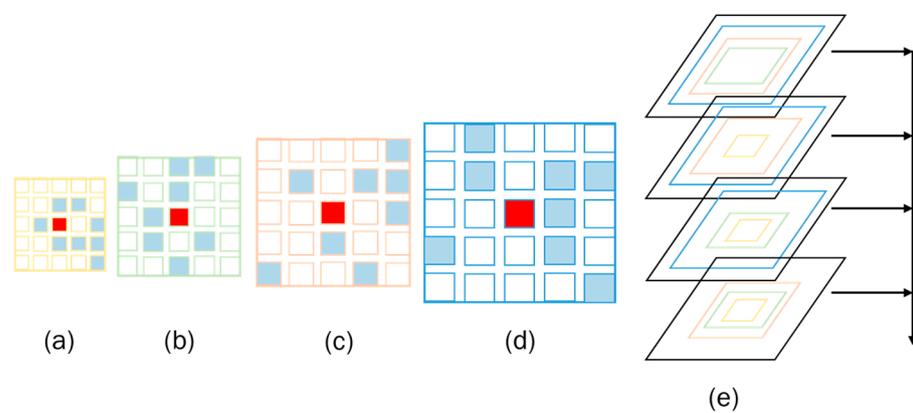


Figure 4. The figure illustrates some adaptive search windows with varying dilation rates for a 3×3 convolutional kernel. Specifically, (a) shows the adaptive search window with a dilation rate of 1; (b) displays the adaptive search window with a dilation rate of 3; (c) presents the adaptive search window with a dilation rate of 4; and (e) shows the fusion of features into four feature maps, respectively, by taking three layers of (a–c), three layers of (b–d), three layers of (a,b,d), and three layers of (a,c,d), and then the residual connection operation is performed to obtain the final feature map.

2.1.2. Kernel Normalized Convolutional Feature Fusion

Through a large number of experiments, it is found that in the final multi-scale feature fusion of the feature extraction stage, the effectiveness of BatchNorm [41] will be greatly reduced when only small batch data can be used for input training if the environment memory is limited. In this regard, we refer to kernel-normalized convolution (KernelNorm) [37] instead of BatchNorm. Kernel normalization first applies random dropout [42] to the normalization units to obtain dropout units, and then normalizes the original normalization units by calculating the average value and variance of the dropout units. Its hyperparameters include kernel size k , stride s , padding d , and dropout probability p .

$$U' = D_p(U) \quad (3)$$

$$\hat{U} = \frac{U - \mu_{u'}}{\sqrt{\sigma_{u'}^2 + \epsilon}} \quad (4)$$

In Equations (3) and (4), D_p represents the dropout operation; U stands for the old normalization unit; U' signifies the dropout units; $\mu_{U'}$ and $\sigma_{U'}^2$ correspond to the average value and variance of the dropout units; and \hat{U} is the new normalization unit.

Applying kernel normalization to the backend of each block ensures that the computation results do not depend on the statistical information of the entire batch. This approach emphasizes local information over global information within the feature space and reduces the dependency of multi-scale feature extraction on batch size, which is crucial for stereo matching tasks involving images under different scales and conditions. Moreover, by normalizing through kernel normalized convolution, the model is less prone to overfitting on specific scales or features, enhancing its generalization capabilities to unseen data. The specific network structure of the KAP module is illustrated in Figure 5.

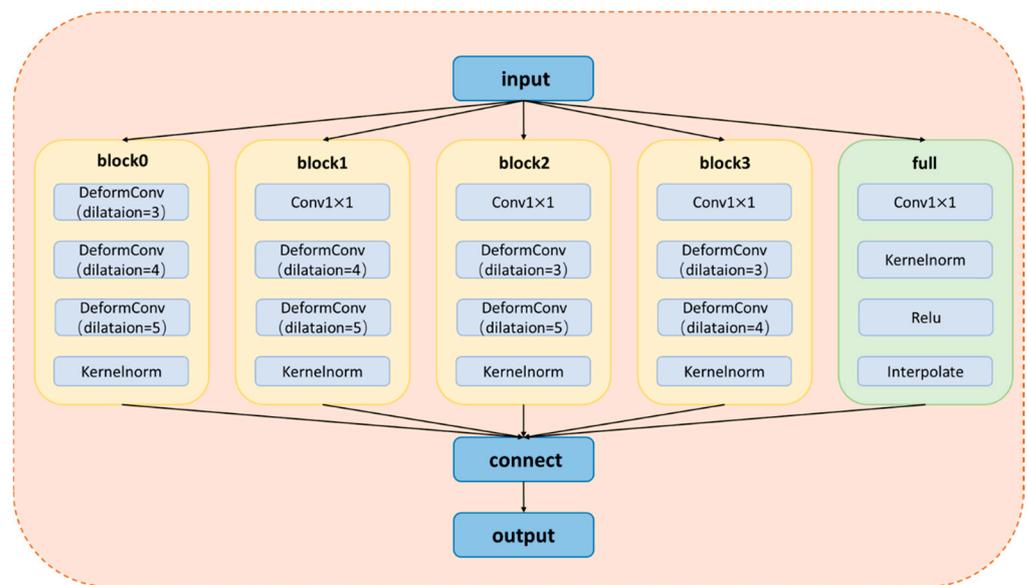


Figure 5. The concrete structure diagram of the KAP module.

2.2. Matching Attention

In the stage of calculating matching cost, there are three main steps: constructing the initial cost volume, generating matching attention weights, and feeding the final cost volume into the GRU-based update operator for further iterative refinement and optimization of the disparity map.

Usually, the construction of the initial cost volume is calculated using the group correlation method [28], that is, the channels of the left and right feature matrices are uniformly grouped along the channel dimension at a disparity level according to the vector inner product. Then, their inner product is calculated group by group, and the average value is calculated on the dimension of $\text{dim} = 2$ to obtain a three-dimensional matrix. When all disparity levels are traversed, the resulting matrices are split and concatenated to obtain the grouped correlation cost volume (group-corr volume).

If there are N_c channels in common, dividing them into N_g groups, the g -th feature groups f_l^g and f_r^g are composed of the $\left\{g \frac{N_c}{N_g}, g \frac{N_c}{N_g} + 1, \dots, g \frac{N_c}{N_g} + \left(\frac{N_c}{N_g} - 1\right)\right\}$ channels of the original features f_l and f_r , and the correlation of all feature groups g and all disparity levels d can be expressed by Equation (5).

$$C_{cor}(d, x, y, g) = \frac{1}{N_c/N_g} \left\langle f_l^g(x, y), f_r^g(x - d, y) \right\rangle \quad (5)$$

This paper believes that the traditional approach of connecting two correlation volumes and performing regularized correlation volume computation is insufficient to fully

utilize the advantages of two image samples. To obtain a more reliable correlation volume, the initial correlation volume is filtered to emphasize useful information and suppress irrelevant information. This paper proposes a novel hybrid matching attention method to preliminarily refine the initial correlation volume. This method applies criss-cross attention [43] (as shown in Figure 6) to traverse the initial 3D correlation volume from both the disparity and epipolar-line directions, and performs preliminary optimization on it. This method reduces the number of iterations and improves accuracy for the next GRU iteration, enhancing pixel-level representation while reducing computational complexity and memory usage.

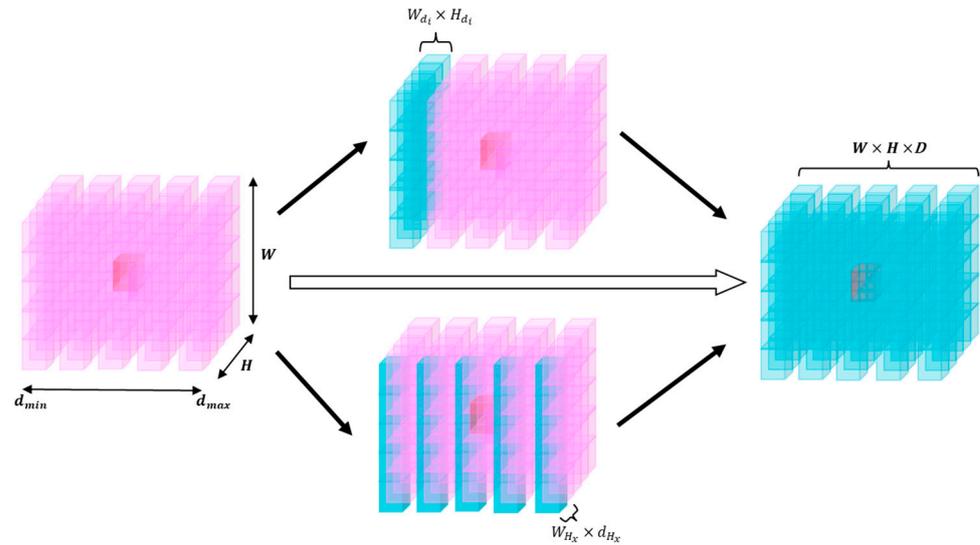


Figure 6. The pink cube on the left is the 3D correlation volume of size $W \times H \times D$ calculated by correlation. The cubes in the middle, positioned above and below, represent the disparity-level attention calculated for the $W_{d_i} \times H_{d_i}$ plane and the epipolar-line-level attention calculated for the $W_{H_x} \times d_{H_x}$ plane, respectively. The blue cube on the right illustrates the refined correlation volume after being processed by the MAR matching attention module. This refined correlation volume is then fed into the GRU for iterative updates to further optimize the disparity map.

As shown in Figure 6, firstly for each disparity level $d_i (i \in [d_{min}, d_{max}])$, we extract the $W_{d_i} \times H_{d_i}$ plane and apply the criss-cross attention mechanism [43] to traverse it. For any spatial position Q_{w_n, h_m} within it, we calculate the correlation weights A_{w_n, h_m} for positions $Q_{w_n, h}$ and Q_{w, h_m} , which align in a cross shape centered at Q_{w_n, h_m} . Thus, A_{w_n, h_m} contains the mutual mapping information of Q_{w_n, h_m} corresponding to any point in this cross-position relationship. Then, the above operation is repeated once, ensuring that for any other spatial position K_{w_a, h_b} , the correlation weights $A_{w, h}$ corresponding to Q_{w_n, h_m} can always be computed. At this point, after refining via attention, the plane obtains the disparity-level attention described by Equation (6).

$$\Phi'_{W_{d_i} \times H_{d_i}} = \sum_{w_n, h_m \in W_{d_i}, H_{d_i}} (A_{w, h} Q_{w_n, h_m} + \Phi_{W_{d_i} \times H_{d_i}}) \quad (6)$$

Then, the epipolar-level attention correlation weights $A_{w, d}$ are computed for any spatial position Q_{w_n, h_m} . Considering that binocular images typically consist of left and right images, based on the epipolar constraints of epipolar geometry, the $W_{H_x} \times d_{H_x}$ plane within $H_x (x \in [H_{min}, H_{max}])$ is extracted. By repeating the above operation, the epipolar-level attention Equation (7) can be obtained.

$$\Psi'_{W_{H_x} \times d_{H_x}} = \sum_{w_n, d_m \in W_{H_x}, d_{H_x}} (A_{w, d} Q_{w_n, d_m} + \Psi_{W_{H_x} \times d_{H_x}}) \quad (7)$$

Finally, by integrating the two aforementioned layers of attention mechanisms, the matching attention formula can be derived as Equation (8).

$$\Omega_{H \times W \times D} = \sigma \left(\Psi'_{W_{H_x} \times d_{H_x}} (\Phi'_{W_{d_i} \times H_{d_i}}) + \Phi'_{W_{d_i} \times H_{d_i}} (\Psi'_{W_{H_x} \times d_{H_x}}) \right) \quad (8)$$

Hence, the final correlation can be expressed as Equation (9).

$$C_{mar}(d, x, y, g) = \frac{1}{N_c / N_g} \left\langle f_l^g(x, y) \cdot \Omega_{H \times W \times D}, f_r^g(x - d, y) \cdot \Omega_{H \times W \times D} \right\rangle \quad (9)$$

Through the matching attention mechanism that combines the self-attention information of both parallax and epipolar directions, the correlation volume is preliminarily refined, and the result is input into the GRU-based update operator for further iterative disparity map optimization, which can make the network more effectively learn different features of the input data while taking into account the interdependence between the channel and the global information. It not only effectively utilizes the effective geometric information of the binocular images and improves the generalization ability of the network, but also greatly reduces the difficulty of the network to find the correct matching points in the search space.

2.3. Kernel Convolution-Based GRU Iterative Update Operator

According to the baseline network, the disparity update uses a sequence of gated recurrent units (GRUs) to combine all the previously acquired data. Starting from an initial starting point $d_0 = 0$, the update operator predicts a sequence of disparity fields $\{d_1, \dots, d_N\}$. Each iteration, it produces an update direction Δd , which is applied to the current estimate: $d_{k+1} = \Delta d + d_k$. The update operator takes disparity, correlation, and potential hidden states as input and outputs an update Δd and the updated hidden state. Mimicking the steps of the optimization algorithm, each iteration uses the current disparity estimate to index the correlation volume, producing a set of correlation features, and then the GRU updates the hidden state, and the new hidden state is then used to predict the disparity update.

The traditional GRU is well suited for handling sequential data but has certain limitations in directly capturing spatial dependencies. To address this, this paper adopts a GRU based on kernel convolution [37] as the iterative update operator within the main network framework. The GRU retrieves features from the stereo matching cost volume obtained earlier, and regresses iterative disparity updates.

First, the soft argmin function is used to regress the matching correlation volume to the initial disparity. The disparity is then updated through three GRU iterative update layers, each containing three kernel convolution layers. The activation function used is the sigmoid function, which connects and injects correlation, disparity, and contextual features into the GRU. The hidden state is then updated, and the new hidden state is used to predict the disparity update. During training, all stages of disparity iterative updates share the same weights, eliminating the need for fine-tuning.

The input-to-hidden and hidden-to-hidden transformations in the GRU make use of kernel convolutions, ensuring that spatial information is preserved and propagated efficiently through recurrent connections. By integrating kernel convolutions into the GRU, the network can process both spatial and temporal information, enabling the GRU to consider a wider range of contextual information around each pixel, facilitating the fusion of local and global information and providing a more comprehensive understanding of the feature space, especially in challenging regions such as thin structures and highly reflective regions. It enhances the network's ability to calculate the scene structure of larger areas.

2.4. Loss Function

Since smooth loss is considered robust to disparity discontinuities and outliers, we use smooth loss as our training loss function. The smooth L1 loss is computed between the predicted values and ground truth over the entire prediction sequence, using an

exponentially weighted L1 distance similar to IGEV [32] as the loss function. For all predicted disparities $d_i\{d_1, \dots, d_N\}$, given ground truth disparities d_{gt} , the loss function \mathcal{L} is defined as Equations (10) and (11).

$$\mathcal{L}_{\text{cor}} = \text{Smooth}_{L_1}(d_0 - d_{gt}) \quad (10)$$

$$\mathcal{L} = \mathcal{L}_{\text{cor}} + \sum_{i=1}^N \gamma^{N-i} |d_i - d_{gt}|_1, \gamma = 0.9 \quad (11)$$

3. Results

3.1. Implementation Details and Evaluation Metrics

In this experiment, the system used is Ubuntu 20.04 with an NVIDIA RTX 3090 GPU, 32 GB of memory, CUDA 11.2, and Python 3.9. Images are randomly cropped to 320×736 and trained using the same data augmentation methods as in [15]. During training, the network was trained for 20 k steps with a batch size of 2, using a single-cycle learning rate schedule with a learning rate of 0.0002, and 16 update iterations were employed.

Generally, the performance evaluation metrics used in stereo matching tasks are EPE, D1, and D1-BG [44]. In addition, we added the size of the model parameters as well as the computation time to the evaluation metrics, counting convolutional layers, fully connected layers, and other layers with parameters in the network structure.

EPE refers to the pixel-level average Euclidean distance between the predicted depth map and the true depth map, which reflects the disparity estimation error information of all pixels of the whole image. It is calculated as shown in Equation (12).

$$\text{EPE} = \sqrt{\left\{ \left(x_{\{pred\}} - x_{\{true\}} \right)^2 + \left(y_{\{pred\}} - y_{\{true\}} \right)^2 \right\}} \quad (12)$$

Both D1 and D1-BG are the percentage of erroneous pixels over all pixels. Generally, when the absolute disparity error of a pixel is greater than three pixels, the pixel is labeled as an error pixel. This index reflects the proportion of erroneous pixels in the whole image [45]. D1 is the proportion of incorrectly predicted pixels in the whole region and D1-BG is the proportion of incorrectly predicted pixels only in the background region; these are calculated by Equations (13) and (14), respectively.

$$D1 = \frac{1}{N} \sum_{\{i=1\}}^{\{N\}} \left\{ \left| D_{\{pred\}(i)} - D_{\{true\}(i)} \right| > 3 \right\} \quad (13)$$

$$D1 - BG = \frac{1}{N} \sum_{\{i=1\}}^{\{N\}} \left\{ \left| D_{\{pred-bg\}(i)} - D_{\{true-bg\}(i)} \right| > 3 \right\} \quad (14)$$

3.2. Datasets

This paper evaluates the proposed AKC-Stereo on several public datasets, including Sceneflow [2], ETH3D [46], Middlebury [47], KITTI-2012 [48], and KITTI-2015 [38]. Compared with the basic IGEV-Stereo method [32], AKC-Stereo improves the accuracy and also makes the network perform better in highly reflective areas and thin-structure areas.

Sceneflow [2] is a large-scale synthetic dataset that contains comprehensive data of more than 35,000 stereo image pairs, and contains the real values of disparity, optical flow, and scene flow. It is suitable for training convolutional networks and can also be used for the evaluation of other methods, and the image size is 540×960 .

KITTI 2012 [48] and KITTI 2015 [38] are two datasets widely used in the field of autonomous driving, composed of data from real driving scenes. They provide sparse ground truth disparity values obtained using LiDAR, for the evaluation of tasks such as stereo matching and optical flow. The KITTI 2012 dataset contains 194 training image pairs and 195 testing image pairs, with an image resolution of 1226×370 . Compared to the 2012

version, the KITTI 2015 dataset includes more dynamic scenes and consists of 200 training image pairs and 200 testing image pairs.

Middlebury 2014 [47] is an indoor dataset that offers 23 pairs of high-resolution images captured under different lighting conditions.

ETH3D [46] is a grayscale dataset consisting of 27 pairs of monochrome stereo images. It provides disparity samples obtained by laser scanning, covering both indoor and outdoor scenes.

3.3. Ablation Experiment

Ablation studies play a crucial role in interpreting the effectiveness of different components and strategies in the ever-changing deep stereo matching field [49]. In order to verify the effectiveness of integrating the matching attention mechanism and the adaptive multi-scale feature extraction module into the proposed network, this paper designs ablation experiments on the KITTI 2015 dataset. For all models in these experiments, it is set to 16 times in the number of iterative update operators. Using IGEV-Stereo as the backbone, this paper separately incorporated the adaptive multi-scale feature extraction module (KAP) and the matching attention mechanism (MAR) to validate their effects on improving the model's performance. As shown in Table 1, both the proposed adaptive spatial pyramid pooling layer (KAP) and the matching attention mechanism (MAR) significantly enhance prediction accuracy. By fully considering the spatial correlation between pixels and the geometric relationships in stereo images, the initial disparity fed into the iterative update network becomes more precise and substantially reduces prediction error.

Table 1. The table presents the results of the ablation study of the proposed AKC-Stereo on the KITTI2015 dataset. KAP denotes the adaptive multi-scale feature extraction module, and MAR refers to the matching attention mechanism. The baseline is IGEV-Stereo, and num_steps is set to 20,000.

Model	EPE (px)	D1-All (%)	Time (s)	Params (M)
Baseline	0.57	1.90	0.34	12.60
KAP	0.50	1.62	0.45	13.22
MAR	0.47	1.64	0.48	13.21
Full Model	0.42	1.52	0.52	13.60

3.4. Benchmark Evaluation

The proposed AKC-Stereo is compared with the state-of-the-art stereo matching methods published by KITTI 2012 and 2015. On the Sceneflow dataset, the EPE of the AKC-Stereo reaches 0.45, which is more accurate than most recent networks, and 0.02 higher than IGEV-Stereo. Detailed quantitative comparisons are provided in Table 2.

Table 2. Quantitative evaluation on Sceneflow dataset. Bold: best.

Model	PSMNet [50]	GwcNet [28]	AANet [25]	LEAStereo [13]	ACVNet [29]	IGEV-Stereo [35]	Ours
EPE(px)	1.09	0.76	0.87	0.78	0.48	0.47	0.45

Figure 7 shows the visual comparison between the computational results of the AKC-Stereo network proposed in this paper and IGEV-Stereo on the Middlebury dataset, where the proposed method is more robust to detailed regions and regions with a background.

In this paper, we conducted a quantitative evaluation of our network on the KITTI 2012 and KITTI 2015 datasets, as shown in Table 3. Compared to most state-of-the-art algorithms in recent years, the AKC-Stereo network proposed in this paper achieved superior performance on many metrics for both the KITTI 2012 and 2015 datasets. For the KITTI 2012 dataset, for the 3-all error metric, the proposed network outperformed IGEV-Stereo and CREStereo by 9.59% and 8.33%, respectively. In the KITTI 2015 comparison, the proposed network exceeded IGEV-Stereo and CREStereo by 5.6% and 11.24%, respectively,

for the D1-all metric. It is evident that the proposed algorithm has shown significant improvements after optimization. However, it is slightly deficient in speed, which is exactly the area we will continue to improve in the next step.

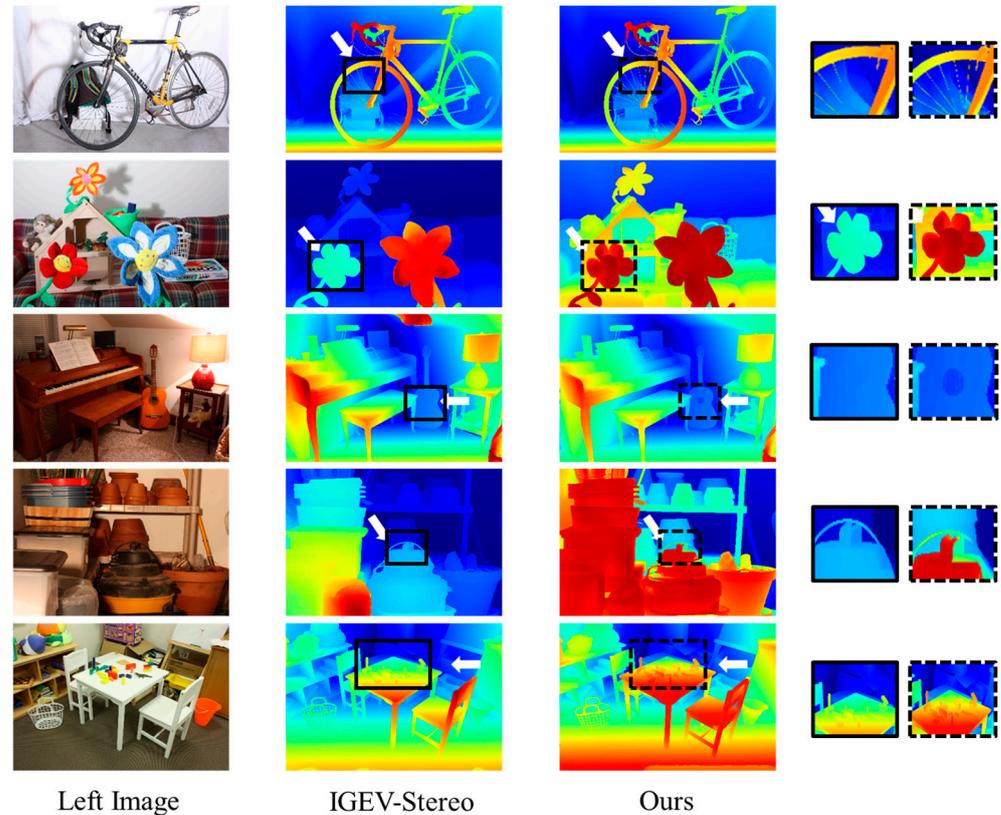


Figure 7. Qualitative results for Middlebury. The first column shows the original images (left images) in the dataset, and the second and third columns show the results of IGEV-Stereo and the AKC-Stereo network proposed in this article, respectively. The proposed network exhibits better results for detailed parts as well as regions with a background.

Table 3. Quantitative evaluation on KITTI 2012 and KITTI 2015. The AKC-Stereo network proposed in this paper runs 16 updates at inference. Bold: best.

Model	KITTI 2012				KITTI 2015			Runtime (s)
	3-noc (%)	3-all (%)	EPE (px) noc	EPE (px) all	D1-bg (%)	D1-fg (%)	D1-all (%)	
PSMNet [50]	1.49	1.89	0.5	0.6	1.86	4.62	2.32	0.41
GwcNet [28]	1.32	1.70	0.5	0.5	1.74	3.93	2.11	0.32
CFNet [51]	1.23	1.58	0.4	0.5	1.54	3.56	1.88	0.18
ACVNet [29]	1.13	1.47	0.4	0.5	1.37	3.07	1.65	0.20
LEAStereo [13]	1.13	1.45	0.5	0.5	1.40	2.91	1.65	0.30
NMRF [52]	1.01	1.35	0.4	0.4	1.28	3.13	1.59	0.09
DMIO [53]	1.14	1.48	0.4	0.4	1.45	2.61	1.64	—
CREStereo [33]	1.14	1.46	0.4	0.5	1.45	2.86	1.69	0.41
RAFT-Stereo [15]	1.30	1.66	0.4	0.5	1.58	3.05	1.82	0.38
IGEV-Stereo [35]	1.12	1.44	0.4	0.4	1.38	2.67	1.59	0.18
Ours	1.03	1.32	0.4	0.4	1.23	2.48	1.50	0.52

Figure 8 shows the qualitative comparison results of the proposed network in the 2015 KITTI dataset, both after 50 k steps of training under the condition of batch size 2. It can be seen that the AKC-Stereo network performs very well in regions with light reflection regions and thin structures, with sharper contours and edges.

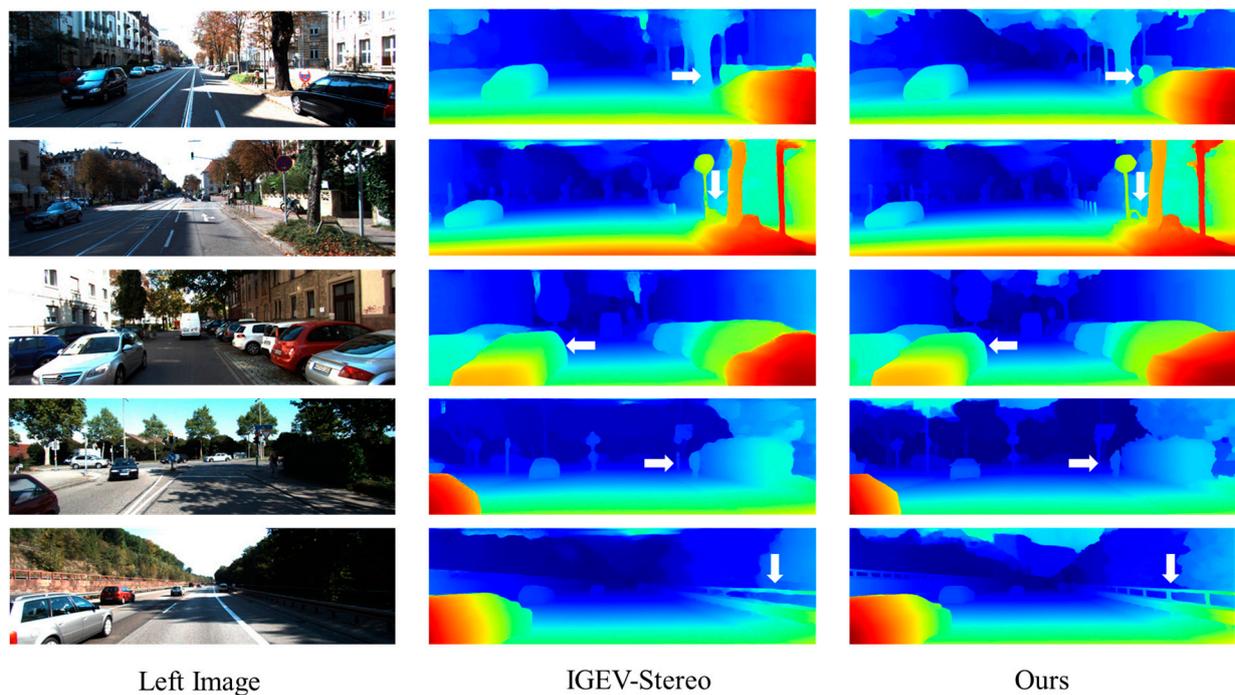


Figure 8. Qualitative results on the KITTI2015 dataset. The first column shows the left image of the original image in the dataset, the second column shows the operation results of the baseline IGEV-Stereo, and the third column shows the operation results of the AKC-Stereo network. From the direction indicated by the arrow, it can be seen that the AKC-Stereo network proposed in this article performs exceptionally well in areas with high light reflection such as signs and areas with thin structures such as railing.

3.5. Generalization Experiment

Since the real world is never a perfect dataset with balanced class distribution, no noise, no outliers, and uniform data distribution, the ability to generalize is therefore essential for a network to be able to adapt to new and unknown situations, rather than just performing well on the training data. The strength of the generalization ability directly reflects whether the patterns learned by the model are universal and applicable to a broader range of situations.

In this section, the generalization ability of the AKC-Stereo network proposed in this paper for unknown scenarios will be studied. The Sceneflow dataset is large and contains images under a variety of scenes and lighting conditions, while the Middlebury and ETH3D datasets are synthetic datasets and real scene datasets, respectively, which can more widely reflect complex situations in the real world. At present, most networks choose to use the Middlebury and ETH3D datasets for generalization ability testing. Therefore, in order to better compare with other advanced network models, the AKC-Stereo network model is also trained on Sceneflow, and directly evaluated on the training sets of Middlebury2014 and ETH3D. The quantitative results are shown in Table 4, and it can be seen that the proposed network achieves better results with more stable performance under the same settings.

Table 4. Generalization experiments. All models are trained on Sceneflow. The 2-pixel error rate is used for Middlebury 2014, and the 1-pixel error rate for ETH3D.

Model	Middlebury2014		ETH3D
	Half	Quater	
PSM [50]	25.8	14.2	23.8
GwcNet [28]	18.1	—	9.0
DSMNet [54]	13.8	8.1	6.2
RAFT-Stereo [15]	8.7	7.3	3.2
IGEV-Stereo [35]	7.1	6.2	3.6
Ours	6.4	5.8	3.3

4. Discussion and Conclusions

This paper proposes a novel adaptive kernel convolution stereo matching recurrent network for stereo matching. This network incorporates an adaptive multi-scale feature extraction module (KAP) and a new matching attention module (MAR) to perform multi-scale feature extraction and obtain a more reliable correlation volume. Through the KAP module, the learnable offset is used to form an adaptive search window to extract multi-scale image features, and the multi-scale features are fused by kernel convolution normalization. Through the new matching attention MAR module, the initial correlation volume is initially refined by self-attention in the two directions of disparity and epipolar line, which provides a more accurate basis for the subsequent GRU iterative update to calculate disparity.

The accuracy of the improved network AKC-Stereo is better than that of the basic network IGEV-Stereo. It outperforms most advanced algorithms developed in recent years across various metrics, particularly excelling in handling pathological regions in images, such as highly reflective surfaces and thin structures. Although the method proposed in this paper has some shortcomings in calculation speed, this is precisely the direction of further research and improvement in the next step.

Author Contributions: Conceptualization, J.W. and H.S.; methodology, J.W.; software, J.W.; validation, J.W.; formal analysis, H.S.; investigation, J.W.; resources, J.W. and H.S.; data curation, J.W.; writing—original draft preparation, J.W.; writing—review and editing, H.S. and P.J.; visualization, J.W.; supervision, P.J.; project administration, H.S.; funding acquisition, H.S.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (No. 62475255).

Data Availability Statement: Data are contained within the article. The data presented in this study are available on request from the corresponding author. The data are not publicly available due to restrictions of privacy.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Scharstein, D.; Szeliski, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vis.* **2002**, *47*, 7–42. [[CrossRef](#)]
- Zeng, J.; Yao, C.; Wu, Y.; Jia, Y. Temporally consistent stereo matching. In Proceedings of the European Conference on Computer Vision, Milan, Italy, September 29–4 October 2024; Springer: Cham, Switzerland, 2024; pp. 341–359.
- Zhang, G.; Lee, J.H.; Lim, J.; Suh, I.H. Building a 3-D line-based map using stereo SLAM. *IEEE Trans. Robot.* **2015**, *31*, 1364–1377. [[CrossRef](#)]
- Wang, Y.; Chao, W.L.; Garg, D.; Hariharan, B.; Campbell, M.; Weinberger, K.Q. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8445–8453.
- Liu, C.W.; Chen, Q.; Fan, R. Playing to Vision Foundation Model’s Strengths in Stereo Matching. *arXiv* **2024**, arXiv:2404.06261. [[CrossRef](#)]
- Shi, T.; Boutry, N.; Xu, Y.; Géraud, T. Local intensity order transformation for robust curvilinear object segmentation. *IEEE Trans. Image Process.* **2022**, *31*, 2557–2569. [[CrossRef](#)]

7. Zhuang, S.; Zhang, X.; Tu, D.; Ji, Y.; Yao, Q. A dense stereo matching method based on optimized direction-information images for the real underwater measurement environment. *Measurement* **2021**, *186*, 110142. [[CrossRef](#)]
8. Yang, A.; Zhang, C.; Chen, Y.; Zhuansun, Y.; Liu, H. Security and privacy of smart home systems based on the Internet of Things and stereo matching algorithms. *IEEE Internet Things J.* **2019**, *7*, 2521–2530. [[CrossRef](#)]
9. Zenati, N.; Zerhouni, N. Dense stereo matching with application to augmented reality. In Proceedings of the 2007 IEEE International Conference on Signal Processing and Communications, Dubai, United Arab Emirates, 24–27 November 2007; IEEE: Piscataway, NJ, USA, 2007; pp. 1503–1506.
10. Butler, D.J.; Wulff, J.; Stanley, G.B.; Black, M.J. A naturalistic open source movie for optical flow evaluation. In Proceedings of the Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Proceedings, Part VI 12. Springer: Berlin/Heidelberg, Germany, 2012; pp. 611–625.
11. Mayer, N.; Ilg, E.; Hausser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; Brox, T. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4040–4048.
12. Tremblay, J.; To, T.; Birchfield, S. Falling things: A synthetic dataset for 3d object detection and pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2038–2041.
13. Cheng, X.; Zhong, Y.; Harandi, M.; Dai, Y.; Chang, X.; Drummond, T.; Li, H.; Ge, Z. Hierarchical neural architecture search for deep stereo matching. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 22158–22169.
14. Tankovich, V.; Hane, C.; Zhang, Y.; Kowdle, A.; Fanello, S.; Bouaziz, S. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14362–14372.
15. Lipson, L.; Teed, Z.; Deng, J. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In Proceedings of the 2021 International Conference on 3D Vision (3DV), London, UK, 1–3 December 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 218–227.
16. Hirschmuller, H. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *30*, 328–341. [[CrossRef](#)]
17. Heo, Y.S.; Lee, K.M.; Lee, S.U. Robust stereo matching using adaptive normalized cross-correlation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 807–822.
18. Fife, W.S.; Archibald, J.K. Improved census transforms for resource-optimized stereo vision. *IEEE Trans. Circuits Syst. Video Technol.* **2012**, *23*, 60–73. [[CrossRef](#)]
19. Hosni, A.; Rhemann, C.; Bleyer, M.; Rother, C.; Gelautz, M. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 504–511. [[CrossRef](#)] [[PubMed](#)]
20. Sun, J.; Zheng, N.N.; Shum, H.Y. Stereo matching using belief propagation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 787–800.
21. Klaus, A.; Sormann, M.; Karner, K. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 20–24 August 2006; IEEE: Piscataway, NJ, USA, 2006; Volume 3, pp. 15–18.
22. Hirschmuller, H. Accurate and efficient stereo processing by semi-global matching and mutual information. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; IEEE: Piscataway, NJ, USA, 2005; Volume 2, pp. 807–814.
23. Seki, A.; Pollefeys, M. Sgm-nets: Semi-global matching with neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 231–240.
24. Feng, M.; Cheng, J.; Jia, H.; Liu, L.; Xu, G.; Yang, X. Mc-stereo: Multi-peak lookup and cascade search range for stereo matching. In Proceedings of the 2024 International Conference on 3D Vision (3DV), Davos, Switzerland, 18–21 March 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 344–353.
25. Xu, H.; Zhang, J. Aanet: Adaptive aggregation network for efficient stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1959–1968.
26. Kendall, A.; Martirosyan, H.; Dasgupta, S.; Henry, P.; Kennedy, R.; Bachrach, A.; Bry, A. End-to-end learning of geometry and context for deep stereo regression. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 66–75.
27. Zhang, F.; Prisacariu, V.; Yang, R.; Torr, P.H. Ga-net: Guided aggregation net for end-to-end stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 185–194.
28. Guo, X.; Yang, K.; Yang, W.; Wang, X.; Li, H. Group-wise correlation stereo network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3273–3282.
29. Xu, G.; Cheng, J.; Guo, P.; Yang, X. Attention concatenation volume for accurate and efficient stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12981–12990.
30. Zheng, D.; Wu, X.M.; Liu, Z.; Meng, J.; Zheng, W.S. Diffuvolume: Diffusion model for volume based stereo matching. *arXiv* **2023**, arXiv:2308.15989.

31. Teed, Z.; Deng, J. Raft: Recurrent all-pairs field transforms for optical flow. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part II 16. Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 402–419.
32. Wang, F.; Galliani, S.; Vogel, C.; Pollefeys, M. Itermv: Iterative probability estimation for efficient multi-view stereo. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8606–8615.
33. Li, J.; Wang, P.; Xiong, P.; Cai, T.; Yan, Z.; Yang, L.; Liu, J.; Fan, H.; Liu, S. Practical stereo matching via cascaded recurrent network with adaptive correlation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16263–16272.
34. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv* **2014**, arXiv:1406.1078.
35. Xu, G.; Wang, X.; Ding, X.; Yang, X. Iterative geometry encoding volume for stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 21919–21928.
36. Wang, H.; Fan, R.; Cai, P.; Liu, M. PVStereo: Pyramid voting module for end-to-end self-supervised stereo matching. *IEEE Robot. Autom. Lett.* **2021**, *6*, 4353–4360. [[CrossRef](#)]
37. Nasirigerdeh, R.; Torkzadehmahani, R.; Rueckert, D.; Kaissis, G. Kernel normalized convolutional networks. *arXiv* **2022**, arXiv:2205.10089.
38. Menze, M.; Geiger, A. Object scene flow for autonomous vehicles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3061–3070.
39. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [[CrossRef](#)] [[PubMed](#)]
40. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9308–9316.
41. Ioffe, S. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
42. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
43. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 603–612.
44. Xu, P.; Xiang, Z.; Qiao, C.; Fu, J.; Pu, T. Adaptive Multi-Modal Cross-Entropy Loss for Stereo Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 5135–5144.
45. Zhang, Z.R.; Fan, D.Z.; Ji, S.; Dong, Y.; Li, D.Z.; Liu, J. Stereo matching of satellite remote sensing images based on attention mechanism. *J. Spatio-Temporal Inf.* **2024**, *31*, 41–49.
46. Schops, T.; Schonberger, J.L.; Galliani, S.; Sattler, T.; Schindler, K.; Pollefeys, M.; Geiger, A. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3260–3269.
47. Scharstein, D.; Hirschmüller, H.; Kitajima, Y.; Krathwohl, G.; Nešić, N.; Wang, X.; Westling, P. High-resolution stereo datasets with subpixel-accurate ground truth. In Proceedings of the Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, 2–5 September 2014; Proceedings 36. Springer International Publishing: Berlin/Heidelberg, Germany, 2014; pp. 31–42.
48. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 3354–3361.
49. Guo, X.; Zhang, C.; Lu, J.; Duan, Y.; Wang, Y.; Yang, T.; Zhu, Z.; Chen, L. Openstereo: A comprehensive benchmark for stereo matching and strong baseline. *arXiv* **2023**, arXiv:2312.00343.
50. Chang, J.R.; Chen, Y.S. Pyramid stereo matching network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5410–5418.
51. Shen, Z.; Dai, Y.; Rao, Z. Cfnet: Cascade and fused cost volume for robust stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13906–13915.
52. Guan, T.; Wang, C.; Liu, Y.H. Neural Markov Random Field for Stereo Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 5459–5469.
53. Shi, Y. Rethinking Iterative Stereo Matching from Diffusion Bridge Model Perspective. *arXiv* **2024**, arXiv:2404.09051.
54. Zhang, F.; Qi, X.; Yang, R.; Prisacariu, V.; Wah, B.; Torr, P. Domain-invariant stereo matching networks. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part II 16; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 420–439.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.