Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 7386280, 11 pages
https://doi.org/10.1155/2021/7386280

Research Article

Group-Based Atrous Convolution Stereo Matching Network

Qijie Zou,[1] **Jing Yu**,[1] **Hui Fang**,[2] **Jing Qin**,[1] **Jie Zhang**,[1] **and Shengkai Liu**[1]

[1]*Department of Information Engineering Faculty, Dalian University of China, 116622, China*
[2]*School of Computer Science, Loughborough University, LE113TU, UK*

Correspondence should be addressed to Jing Yu; dl_yujing163@163.com

Received 21 August 2021; Revised 11 October 2021; Accepted 2 November 2021; Published 28 November 2021

Academic Editor: Chi-Hua Chen

Stereo matching is the key technology in stereo vision. Given a pair of rectified images, stereo matching determines correspondences between the pair images and estimate depth by obtaining disparity between corresponding pixels. The current work has shown that depth estimation from a stereo pair of images can be formulated as a supervised learning task with an end-to-end frame based on convolutional neural networks (CNNs). However, 3D CNN puts a great burden on memory storage and computation, which further leads to the significantly increased computation time. To alleviate this issue, atrous convolution was proposed to reduce the number of convolutional operations via a relatively sparse receptive field. However, this sparse receptive field makes it difficult to find reliable corresponding points in fuzzy areas, e.g., occluded areas and untextured areas, owing to the loss of rich contextual information. To address this problem, we propose the Group-based Atrous Convolution Spatial Pyramid Pooling (GASPP) to robustly segment objects at multiple scales with affordable computing resources. The main feature of the GASPP module is to set convolutional layers with continuous dilation rate in each group, so that it can reduce the impact of holes introduced by atrous convolution on network performance. Moreover, we introduce a tailored cascade cost volume in a pyramid form to reduce memory, so as to meet real-time performance. The group-based atrous convolution stereo matching network is evaluated on the street scene benchmark KITTI 2015 and Scene Flow and achieves state-of-the-art performance.

1. Introduction

Some complex advanced visual tasks depend on depth perception [1], such as robot control and navigation [2], three-dimensional measurement [3], unmanned aerial vehicles (UAVs), virtual reality, and microoperating system parameter detection, showing the significance of distance information acquisition for vision works. Stereo matching is one of visual tasks, which computes the disparity of each pixel when given a pair of rectified images. Extensive work has been proposed for the task, including conventional methods and newly deep learning (DL) methods. Nevertheless, in practical application, the size of the target object is diverse. For larger objects (such as indoor walls and tables and outdoor sky and ground), we may ignore fine details; for smaller objects (such as pedestrians and vehicles), it will be a lack of global information. These make it difficult in dealing with ill-posed regions which still need to be calculated accurately, such as weakly textured regions, occluded areas, and reflective surfaces. Moreover, CNN-based algorithms are computationally expensive. These deficiencies make the research based on vision develop continuously.

Stereo matching estimates depth by matching pixels from a rectified image pair captured by two cameras, in which the goal is to obtain distance and contextual information from disparity quickly and accurately. The performance depends on robustly segmenting objects at multiple scales. Features provide contextual information for the stereo matching process to compute disparity.

Conventional stereo matching can be expressed as a multi-level optimization problem, which generally includes four steps: matching cost calculation, cost aggregation, disparity calculation, and disparity optimization [4]. The method achieves a trade-off between the computational complexity and the quality of the results obtained. However, the performance of the traditional stereo matching method is heavily limited by the

handcrafted features, such as graph cut [5], belief propagation [6], BM (Block Matching), and DP (Dynamic Programming) [7], which are adopted by cost functions, with poor robustness [8–11].

At present, stereo matching algorithms have become a deep learning task resorting to the development of CNN. CNN is introduced to replace one or more components in the legacy stereo pipeline. Some CNN-based algorithms have shown their stronger feature extraction ability than traditional methods. MC-CNN [12] first utilizes a convolutional neural network to learn how to match corresponding points in the matching cost computation process. End-to-end disparity estimation network as one of the CNN-based algorithms, which integrates all steps in the stereo matching pipeline for concatenating optimization, produces dense disparity maps from stereo images directly. Stereo matching networks with end-to-end approach are able to generate highly accurate depth estimation from stereo image pairs. However, they require huge memory and computation consumption. Meanwhile, it is difficult to infer reliable correspondences in ill-posed with limited receptive field and lack of contextual information. Atrous Spatial Pyramid Pooling (ASPP) explores an incoming convolutional feature layer with filters at multiple sampling rates and effective fields-of-views, thus capturing objects as well as rich image context at multiple scales [13].

In particular, we consider that the challenge of stereo matching is how to reduce memory usage while ensuring that the contextual information is fully utilized. Therefore, in this work, we design a network which could make use of rich contextual information as well as minimize GPU memory occupation. Specifically, we propose a small dilation atrous convolution group and a light tailored cascade cost volume, which boost the accuracy and performance by extracting features with more contextual information and lessen consumed memory and time.

Our main contributions are listed as follows:

(1) A group-based atrous convolution pyramid module is presented in this paper. The module ensures multiscale context information captured from various receptive fields when reducing the network size significantly

(2) The tailored cascade cost volume is constructed by changing the output channels and utilizing pyramid construction of cascade structure leading to efficiency to calculate the disparity

It is worth noting that the above two properties of GASPP make the training error converge faster in matching, and the accuracy of disparity regression is higher. Compared with the classic CNN-based model, PSM-Net, the average time of each iteration of our model is shortened by about 30%. In addition, the model is less dependent on the batch size.

2. Background Knowledge

2.1. Stereo Matching Based on Deep Learning. Currently, stereo matching algorithms can be treated as a deep learning task. For convenience, we classify stereo matching based on deep learning algorithms into two categories: non-end-to-end learning algorithm and end-to-end algorithm. An end-to-end algorithm learns to map an image pair directly to a disparity map using a deep learning network.

CNN-based algorithms use a convolutional neural network for geometry learning by making full use of the contextual information in stereo images. It is crucial to make utmost use of environmental contextual information, which includes local and global features, for further preserving subtle details.

GC-Net [14] is presented as an end-to-end supervised learning model that expresses the correspondences between stereo image patches by constructing a cost volume. To fully explore global contextual information, multiscale 3D convolution and 3D deconvolution are used to cost volume regularization. PSM-Net [15] introduces the Spatial Pyramid Pooling (SPP) [16] module, which extends pixel-level features to regional-level features of different scales, combing global and local feature cues to form a reliable cost volume [15, 16]. In addition, to gather more contextual features, Song et al. [17] propose a multitask network named EdgeStereo, which is composed of backbone subnetworks and edge subnetworks, to integrate edge cues by embedding edge perception smoothness loss regularization to improve matching performance.

2.2. Atrous Convolution. Atrous convolution is a method to expand the receptive field, known as "dilated convolution" due to contain dilation filters [18]. The receptive field can be understood as the size of the receptive range of neurons in the network to the original image, and it also refers to the size of the pixel point on the original image of the output feature map of each layer. Another pooling method can also extend the receptive field through downsampling. Downsampling has to unavoidably downsample the resolution of the image. When performing a pooling operation each time, the spatial resolution will be sacrificed. Thus, downsampling causes the loss of information and affects the performance of the disparity regression. An atrous kernel can be dilated by inserting zeros into suitable positions. In other words, atrous convolution has the ability to implement a larger receptive field size without sacrificing spatial resolution. Compared with a traditional convolution operator, the feature map generated by atrous convolution has the same size of its input while representing features from a larger receptive field, which means that higher-level semantics can be encoded.

In contrast, ordinary convolution can obtain a larger receptive field size by choosing a large size convolution kernel, connecting several convolutional layers with small size or other methods. However, these approaches will increase the number of parameters and calculation cost, which cannot guarantee real-time processing performance [19]. To this end, atrous convolution is developed to gather feature information. It inserts "holes" (that is, zero) into the encoding part of the convolution kernel, which can obtain a larger receptive field size without increasing the number of kernel parameters. This approach ensures that it can gain more accurate predictions while keeping the same computation cost. However, the atrous convolution framework has an inherent problem, which is the gridding issue, which incurs information loss for feature extraction. The architecture is shown in Figure 1.

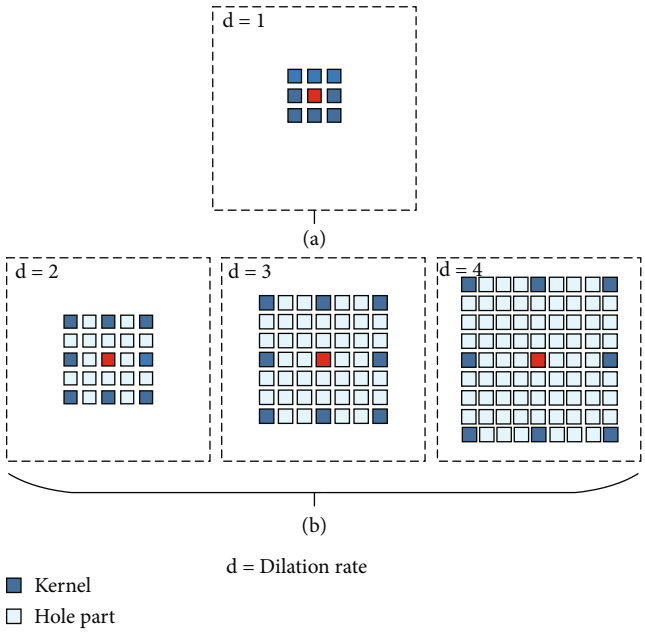

Figure 1: The comparison of ordinary convolution and hole convolution: (a) schematic diagram of ordinary convolution; (b) schematic diagram of atrous convolution.

The receptive field size of a convolution kernel can be formulated as the following equation:

$$R = (d - 1) \times (K - 1) + K, \tag{1}$$

where d represents the dilation rate and K represents the size of a kernel. R is the size of the receptive field.

2.3. Cost Volume. In stereo matching tasks, the cost volume performs matching cost calculation, whose purpose is to measure the correlation between the pixels to be matched and the candidate pixels [20–23]. Whether two pixels are homonymy points, the matching cost can be calculated by the matching cost function. The cost is smaller, inversely represents greater correlation, and also implies that the probability of being homonymy points is greater. Thus, cost volume is also equivalent to a space similarity measurement.

As shown in Figure 2, a standard cost volume usually consists of $W \times H \times D \times F$, which contains four dimensions, where $W \times H$ is regarded as the spatial resolution, D is the number of disparity planes, and F represents the number of channels of the feature map. A four-dimension cost volume retains feature dimensions and integrates it into the cost volume.

3. Related Work

3.1. Atrous Convolution-Based Methods. In the stereo matching task, contextual information refers to the relationship between an object and its surroundings or subregions [24], for instance, the relationship between a vehicle and its subareas (windshield, door). The size of the receptive field

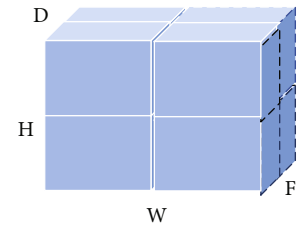

Figure 2: Standard cost volume.

indicates where the contextual information comes from. The atrous convolutional layer is typically used to solve problems related to semantic segmentation, which resolves the contradiction between resolution and the receptive field [25, 26]. In our paper, we apply it to an end-to-end stereo matching network for implementing multiscale information feature extraction.

Although atrous convolution solves the contradiction between feature map resolution and a larger receptive field [27–30], it has unicity when generating semantic information from the feature map. Specifically, all neurons in the feature map are generated by the atrous convolution with the same receptive field, which implies that only a single-scale feature is applied to complete the semantic information generation process. ASPP [13] uses atrous convolution that increases the dilation rate layer by layer to extract features and organizes the atrous convolutional layers in a parallel manner to obtain multiscale information. Multiscale information is a significant factor in segmentation of varying scales and fuzzy pixels; it requires gaining diverse ranges of contextual information to provide richer information for subsequent cost aggregation. However, when the dilation rate increases to a certain extent, the number of effective filtering parameters decreases gradually. In the extreme, when the sizes of the receptive field and the feature map are the same, the convolution does not capture the contextual information of the entire image but degenerates into a $1 * 1$ convolution, gradually losing its modeling ability.

Thus, to design a network architecture that encodes multiscale information while keeping a large receptive field sufficiently, Yang et al. [29] propose a method called DenseASPP that stacks atrous convolutional layers in parallel and in cascade. The cascade structure is mainly composed of multilevel atrous convolutions with a gradually increasing dilation rate. The parallel structure deals the same input feature with a sequential of atrous convolutional layers with various dilation rates, followed by connecting the results together and obtaining an output feature that is the input multiscale sampling feature. In DenseASPP, each atrous convolution makes full use of an atrous filter with a reasonable dilation rate. Through a series of atrous convolutions, the neurons in the later layers possess a larger receptive field without encountering the issue of kernel degradation. Extracting features by the cascade method implements a very large coverage and denser features, improving the recognition ability of the algorithm when the target ratio changes, so as to improve the robustness of matching.

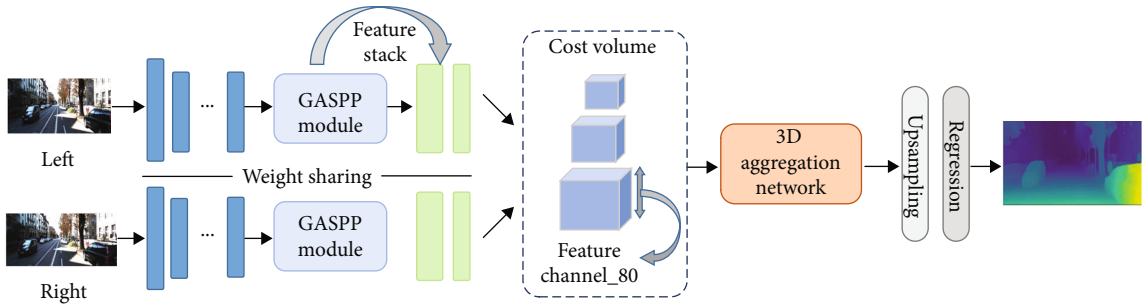

Figure 3: The architecture of proposed group-based atrous convolution stereo network.

3.2. Cost Volume Optimization. GC-Net [14] firstly uses unary features to form the cost volume to compute the stereo matching cost. This allows the network to learn incorporating context which can operate over features. But the large space in the cost volume also leads huge computation cost, which requires more active research effort to deal with.

Cascade cost volume [31] is constructed based on the idea of a feature pyramid to realize disparity estimation from coarse to fine level. A lower resolution disparity map is constructed by a smaller cost volume to complete the first estimate. In the following stage, it can reduce the disparity hypothesis range of the current scale and adjust the disparity map output by utilizing the previous stage, which can reduce the memory usage. GWCNet [32] adopts a group-related strategy, and group multichannel features are mapped with the disparity channel to form a group-based cost volume, which is comprised by concatenated volume and group-related volume. The concatenated volume is the same as that of PSM-Net but contains fewer channels, alleviating the need for parameters. Based on GC-Net [14], Lu et al. [33] propose the Sparse Cost Volume (SCV) to introduce the concept of step size (the distance that right feature moves) in the process of gathering the features of the image pair to form the cost volume to compress adjacent pixels to the center pixel to reduce redundant calculations. The Convolutional Attention Residual Network (CAR-Net) [34] combines residual network and attention mechanism to extract features, with simplifying network parameters while retaining the semantic information of the feature map.

Although methods based on cost volume remarkably boost the performance, they still rely on interpolation operation and downsampled cost volumes, which need high GPU storage. Our tailored cascade volume modifies feature channels to improve real-time performance and GPU memory efficiency.

4. Group-Based Atrous Convolution Stereo Network

4.1. Motivation. In semantic image segmentation, atrous convolutions can realize large receptive fields and meet the requirements of multiscale feature information through various dilation operations, which makes an effective method to deal with the challenging scale changing problem of objects. However, when the dilation rate increases, the spacing between the atrous convolution kernels also increases and prompts some local information loss, which can be understood as a gridding problem. Atrous convolution becomes more and more ineffective with increased dilation rate, and the number of parameters becomes higher. Consequently, these issues can influence the segmentation accuracy. Thus, we expect to design a method that enables large receptive fields utilizing small dilation convolutional layers and is more suitable for the stereo matching model. Based on the above ideas, we present group-based atrous convolutional layers to solve the contradiction between the receptive field and the gridding problem.

Moreover, the stereo matching model requires high real-time performance. At present, the deep learning model occupies a lot of memory and necessarily takes a long time to train; this obeys the real-time thought. Therefore, we introduce a tailored cascading cost volume. On the one hand, the structure is designed as a pyramid to match the previous GASPP. On the other hand, it enables reducing the dependence on batch size and the consumption of computer memory to enhance real-time performance.

4.2. Network Architecture. We present an end-to-end group-based atrous convolution stereo network. The architecture is shown in Figure 3. Our stereo matching model is composed of four parts: feature extraction, cost volume, cost aggregation, and disparity regression.

In the feature extraction part, a group-based atrous spatial pyramid module is introduced to generate features to reduce the loss of local information. The cost volume is tailored on the basis of the cascaded cost volume, and we adjust the number of channels to reduce the memory usage and accelerate computing speed, so as to cooperate with the group-based atrous pyramid module. Cost aggregation uses a three-dimensional stacked hourglass network; at last, we use soft-argmin function [13] to complete the disparity regression.

4.3. Group-based Atrous Convolution Spatial Pyramid Pooling Module. A group-based atrous pyramid structure is proposed for feature extraction. Feature extraction is the prerequisite first step for correct disparity estimation. In PSM-Net, the SPP module is used to extend pixel-level features to regional-level features with different scales, then utilizes global contextual information for stereo matching [14].

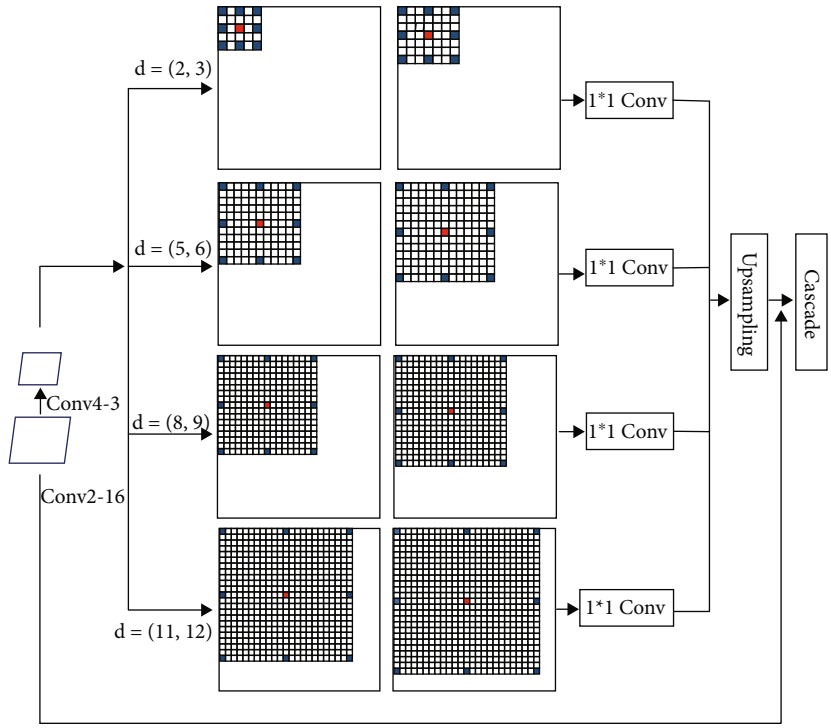

Figure 4: The structure of GASPP.

It is crucial to acquire contextual information for stereo matching, and the amount of contextual information is closely related to the size of its receptive field.

In the atrous convolution kernel, zero is filled between two pixels, so the receptive field only covers a gridding region, which implies that only nonzero positions are sampled and some adjacent information is lost. When the dilation rate increases, the gridding problem becomes more serious. For instance, ASPP uses an atrous convolutional layer with dilation rates of 6, 12, 18, and 24; the interval is 6, which can implement the purpose of extracting multiscale information. However, since the kernel of atrous convolution is not continuous, its calculation formula is similar to the gridding format as shown in Figure 1. As the dilation rate increases, the receptive field becomes larger and the distance between the kernels also increases. Then, there is less correlation between the obtained information, and more local information will be lost, which will affect the performance of the disparity regression later. DenseASPP obtains wider and denser feature information by cascading, but it still has a gridding problem.

To incorporate large context and compute feature maps more densely, whilst reducing the loss of local information and improving the accuracy of the disparity map, we propose the Group-based Atrous Convolution Spatial Pyramid Pooling (GASPP). The model is shown in Figure 4. The basic idea is to substitute an original atrous convolution with a large dilation rate with two continuous atrous convolutions which have a smaller dilation rate. The purpose is to use a smaller dilation rate but maintain the size of receptive field at the same time. In other words, the correlation between the information gets strengthened and the hole area smaller. Consequently, it improves the disparity regression performance in those ill-posed areas, e.g., occluded regions. We design four groups of atrous convolutional layers in the GASPP module and assign two atrous convolutional layers with continuous increased dilation rate to each group. The designed dilation rate between each group is gradually increased, forming{(2,3) (5,6) (8,9) (11,12)} four parallel branches, which can provide feature maps of spatial information with different scales, and the four branches complement each other. Ultimately, the output is accumulated to obtain a feature map that contains multiscale information, with wider and denser receptive fields.

To simplify notations, we use $G^n_{K,d}(x)$ to represent a group-based atrous convolution and consequently write GASPP y as the following equation:

$$y = \begin{cases} G^1_{3,2}(x) + G^1_{3,3}(x) + G^2_{3,5}(x) + G^2_{3,6}(x), \\ +G^3_{3,8}(x) + G^3_{3,9}(x) + G^4_{3,11}(x) + G^4_{3,12}, \end{cases} \tag{2}$$

where d is the dilation rate, K denotes the kernel size, and n is the group number. The receptive field size of GASPP can be formulated as the following equation:

$$\begin{cases} d_2 = d_1 + 1, \\ R_g = (2d_1 - 1) \times (K - 1) + 2K, \end{cases} \tag{3}$$

where d_1 and d_2 represent the size of a kernel of a group. Otherwise, R_g contributes to the size of the group-based atrous convolution receptive field.

In this manner, if only one atrous convolution with a dilation rate of 6 and kernel of 3 is used, the size of the receptive field is $5 * 2 + 3 = 13$, while if two atrous convolutions, one with a dilation rate of 3 and kernel of 3 and another with a dilation rate of 4 and kernel of 3, are used, the size of the receptive field is $(2 * 2 + 3) + (3 * 2 + 3) = 16$. Otherwise, if we chose two atrous convolutions, one with a dilation rate of 2 and kernel of 3 and another with a dilation rate of 3 and kernel of 3, the size of the receptive field is $(1 * 2 + 3) + (2 * 2 + 3) = 12$. Ultimately, we choose the $(2, 3)$ combination that has the most similar receptive field with dilation rate of 6. This same setting is applied to all other GASSP groups.

After feature extraction, we take some measures to make the cost volume be appropriate for the GASPP model. In the following section, we will show the formulation about the tailored cascade cost volume.

4.4. Tailored Cascade Cost Volume. Gu et al. [31] introduce the cascade cost volume. The algorithm is based on the idea of the feature pyramid, which processes the cost volume in stages by utilizing a gradually refined scale (forecast output from coarse to fine). The formation of the cost volume mainly goes through three steps: firstly, the space of the cost volume is determined, which is the disparity plane; secondly, what to do is to warp the features extracted from the stereo image pair to the disparity plane to construct feature volume; ultimately, the feature volume is fused to build the cost volume.

The cascaded cost volume is divided into two stages in the stereo matching task. For the processing of the feature volume, the initial number of channels is changed from 32 to 320 through a convolutional layer, which is consistent with channels for feature extraction. We utilize two 2D convolutional layers to tailor the cascade cost volume. One 2D convolutional layer is used to reduce the number of channels to 160, and another one is used to cut down the number of feature channels to 80 as shown in Figure 3. In this way, the characteristic channels are {1/2, 1/4}, respectively, and the number of final channels is fewer, which occupy less space during training process. This design is similar to a layered structure, which reduces layer by layer, losing less information and ensuring the performance and effect of the network.

4.5. Output Module and Loss Function. In the output module, we use two 3D convolutions to generate a 4D volume with 1 channel, followed by the upsampling operation and transforming it into a probability volume with a softmax function along with a disparity dimension. The probability of each pixel is calculated from the predicted cost through softmax operation $\sigma(\bullet)$. Due to the higher the cost, the probability of matching is lower, so we take prediction negative cost $-c_d$. At last, the predicted disparity $\bar{d}$ is obtained by calculating the weighted probability sum of each disparity d; the disparity regression function is defined as the following equation:

$$\bar{d} = \sum_{d=0}^{\max} d \times \sigma(-c_d), \tag{4}$$

where d and $\sigma(-c_d)$ denote a possible level and the corresponding probability. The final loss is given by the following equation:

$$L = \sum_{k=0}^{3} \lambda_k \bullet \text{Smooth}_{L_1}\left(\bar{d}_k - \hat{d}\right), \tag{5}$$

where $\lambda_k$denotes the coefficient of the k_{th} disparity prediction and $\hat{d}$denotes ground-truth disparity maps. The predicted disparity map in the last four output modules are represented by $\bar{d}_0$, $\bar{d}_1$, $\bar{d}_2$, and $\bar{d}_3$. The Smooth_{L_1} loss function is defined as the following equation:

$$\text{Smooth}_{L_1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1, \\ |x| - 0.5, & \text{otherwise.} \end{cases} \tag{6}$$

5. Experiment

This section introduces the experimental settings and results. We evaluate the key components of our model on the Scene Flow [35] and KITTI 2015 [36] datasets. In addition, we also compare our method with advanced stereo matching methods on the KITTI benchmark.

5.1. Experiment Details

5.1.1. Experiment Environment Settings. The end-to-end network is implemented in Windows environment and runs under the PyTorch deep learning framework. In terms of hardware facilities, we apply NVIDIA 1070Ti GPU to train our model, and the batch size is set to 2.

For all datasets, we set the resolution of the training stereo image pair to 512×256, the RGB values of all images are normalized in [-1, 1], and the initial maximum disparity $D_{\max}$ is set to192. The model is trained with the Adam optimizer, and several optimization parameters are set to $\beta_1 = 0.9$ and $\beta_2 = 0.99$. In addition, the four output coefficients are adjusted, respectively, as

$$\lambda_0 = 0.5, \lambda_1 = 0.5, \lambda_2 = 0.7, \lambda_3 = 1.0. \tag{7}$$

Commonly, end-to-end stereo networks are pretrained from scratch on the Scene Flow dataset, then further to optimize the network on a smaller target dataset such as KITTI 2015. The Scene Flow dataset is more effective when moving to KITTI. Therefore, for the Scene Flow dataset, we train 15 epochs with a fixed learning rate of 0.001. For the KITTI 2015 dataset, we employ the pretrained model on the Scene Flow for further optimization. The KITTI optimization training is set for 300 epochs. The learning rate for the first 200 epochs is 0.001, and that of the other 100 epochs is 0.0001.

5.1.2. Datasets

(1) Scene Flow. The dataset is a large synthetic dataset, which includes 35454 pairs of stereo images for training and 4370

Table 1: Experimental evaluation of different components of our network.

Network model							>3px	EPE	Run time (s)
The dilation of GASPP				Channel of cost volume					
(2,3)	(5,6)	(8,9)	(11,12)	40	80	160			
✓	✓						4.21	0.87	1.63
	✓	✓			✓		3.40	0.81	1.36
✓	✓	✓			✓		3.02	0.73	1.23
	✓	✓	✓	✓			3.21	0.75	1.32
	✓	✓	✓			✓	3.18	0.72	1.28
✓	✓	✓	✓	✓			3.12	0.71	1.26
✓	✓	✓	✓			✓	2.92	0.69	1.19
✓	✓	✓	✓		✓		**2.81**	**0.64**	**1.15**

The bold data represents the optimal data result under this column.

pairs of stereo images for testing. It also provides dense and exhaustive ground truth disparity maps and camera parameter information for each pair, in which the resolution of all images is 960×540.

The Scene Flow subset comprise three scenes. Among them, FlyingThings3D is a scene with random-type objects including extensive floating objects with rich details, the Driving dataset is a street scene captured in the process of simulating car driving, and Monkaa is a scene that includes monkeys in a deep forest environment, which involves closer objects and means more areas with larger disparity values.

(2) KITTI 2015. This is a dataset collected from a real street scene, including 200 pairs of stereo images for training and 200 pairs of stereo images for testing. The dataset provides a sparse disparity map collected by LiDAR as the ground truth value.

For the Scene Flow dataset, we select End-Point-Error (EPE) as the evaluation metric, which is defined as the mean average disparity error in pixels, calculating the Euclidean distance between the prediction error of each pixel and true error, followed by taking the average value. The rate between error and the accuracy of matching is the inverse proportion, so when the error is smaller, the matching accuracy is inversely higher. The calculation method of EPE is defined as follows:

$$\mathrm{EPE} = \frac{1}{N} \sum_{m \in N} \sqrt{(d_m - d\wedge_m)^2}, \tag{8}$$

where N denotes the total number of pixels, d_m represents the ground truth at the m_{th} pixel, and $\widehat{d}_m$ represents the predicted disparity value at the m_{th} pixel.

For the KITTI 2015 dataset, we apply a 3 px error as the predicted error, which differs from the label by a threshold of three pixels. In other words, the 3 px error refers to the ratio of the number of pixels whose absolute value between the predicted disparity value and the ground truth value exceeds 3 to the entire image.

5.2. Ablation Experiment. In this section, to understand the impact of each model components on the final performance, we conducted a comprehensive ablation study by adding each component individually and showed how the components make an impact on the performance. The overall comparative experimental results of our network are shown in Tables 1 and 2. Moreover, we compare our proposed method with one of the most advanced stereo matching methods, PSM-Net, to evaluate our network jointly on the Scene Flow and KITTI 2015 datasets.

5.2.1. Ablation Experiment for GASPP Module. We first conduct ablation experiments on the GASPP structure. GASPP is a group of two consecutive convolutional layers with a continuous dilation rate, which consists of four groups. This part mainly compares the pooling module, GASPP module, original SPP module, ASPP module, and DenseASPP module.

Combining the experimental results in evaluation of different components of our network in Tables 1 and 3 indicated that the GASPP module improves the effect and performance of disparity estimation to a certain extent. The application of the group-based atrous convolutional layer reduces EPE from 0.62 to 0.57. The analysis shows that for the complex KITTI 2015 dataset, the SPP module loses more spatial information in the process of pooling and upsampling, resulting in slightly worse effect of the network, while GASPP can collect contextual information more closely and preserve the salient local information.

Inspired by DenseASPP, we propose the GASPP module based on the idea of grouping. Among them, the size of the dilation rate within the group and the interval of the dilation rate between each group need to be manually set. In order to achieve the optimization, the study conducted multigroup experiments on the dilation rate and interval between groups of the GASPP module, setting a total of 7 sets of parameters. The interval between the first three groups is 1, and the initial dilation rate is different. The last four groups are set with different numbers of interval, and the dilation rate of the initial group is the same for testing. The experimental results are shown in Table 2. It can be seen from the results in Table 2 that the best result can be obtained when the interval is 2.

Table 2: Evaluation of experimental effects of GASPP module with various dilation rates.

Serial number	Interval of dilation rate	Dilation group				>3px	EPE	Run time (s)
		Group 1	Group 2	Group 3	Group 4			
1	1	(2.3)	(4.5)	(6.7)	(8.9)	3.22	0.83	1.23
2	1	(3.4)	(5.6)	(7.8)	(9.10)	3.15	0.87	1.34
3	1	(5.6)	(7.8)	(9.10)	(10.11)	3.21	0.85	1.28
4	2	(2.3)	(5.6)	(8.9)	(11.12)	**2.91**	**0.68**	**1.15**
5	3	(2.3)	(6.7)	(10.11)	(14.15)	3.02	0.70	1.20
6	4	(2.3)	(7.8)	(13.14)	(18.19)	3.11	0.72	1.25
7	5	(2.3)	(8.9)	(14.15)	(20.21)	3.18	0.85	1.28

The bold data represents the optimal data result under this column.

Table 3: Experimental evaluation of GASPP module.

Pooling	EPE	>3px
SPP	0.62	2.62
ASPP	0.60	2.63
DenseASPP	0.59	**2.60**
GASPP	**0.57**	2.61

The bold data represents the optimal data result under this column.

5.2.2. Ablation Experiment for the Tailored Cost Volume. In this paper, we add the tailored cost volume to our model and explore to limit the computational cost by adding a 2D convolutional layer at the end to design the number of feature channels. We provide a comparison between the results of using the 2D convolution layers for the proposed tailored cascade cost volume and the original cascade cost volume in the first row in Table 1. In order to explore the effect of different feature channel numbers on the cascaded cost volume, we set up three groups of feature numbers for verification, which are 40, 80, and 160. The experimental results are shown in Figures 5 and 6.

Figure 5 shows the comparison of the number of network parameters and the run time between PSM-Net and the cascade cost volume of GASPP with different feature channels. The number of network parameters of PSM-Net, Cascade-40, and Cascade-160 decreases by 36.2%, 9.12%, and 3.21%, respectively. Figure 6 displays the correspondence between the number of iterations and the 3 px error. Cascade-80 of GASPP implements the minimum error and the fastest convergence speed at the same time.

The combined relationship is depicted in Figures 5 and 6; with the increase of the number of feature channels, the accuracy of stereo matching is improved. From Cascade-40 to Cascade-160, the number of feature channels has quadrupled, but the error is increased instead. Therefore, the increase of feature channels leads to expensive computational cost but insignificant performance improvement. From the results, one may conclude that the Cascade-80 is the most suitable channel for our network model.

5.2.3. Benchmark Performance on KITTI 2015 Dataset. In this section, we conduct a detailed study of Figure 7 disparity

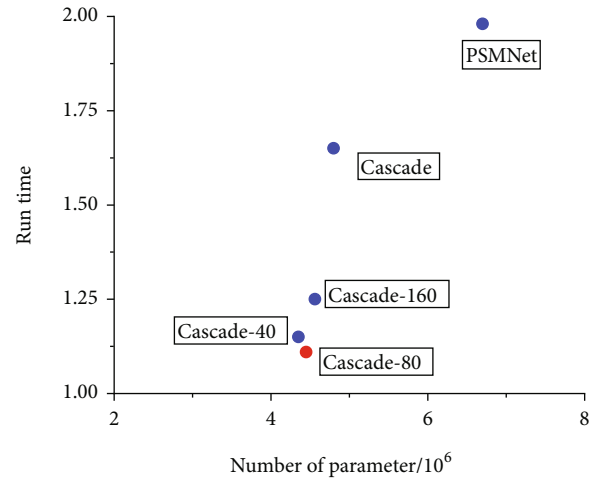

Figure 5: Correspondence between the number of network parameters and run time.

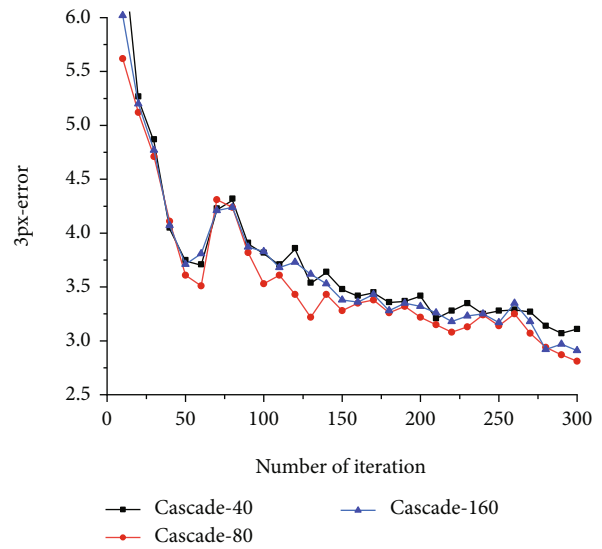

Figure 6: Correspondence between the number of iterations and 3 px error.

maps trained by PSM-Net and the cascade cost volume and contrasted them. We use a rectangular box to mark the areas where our network has different matching effects with PSM-

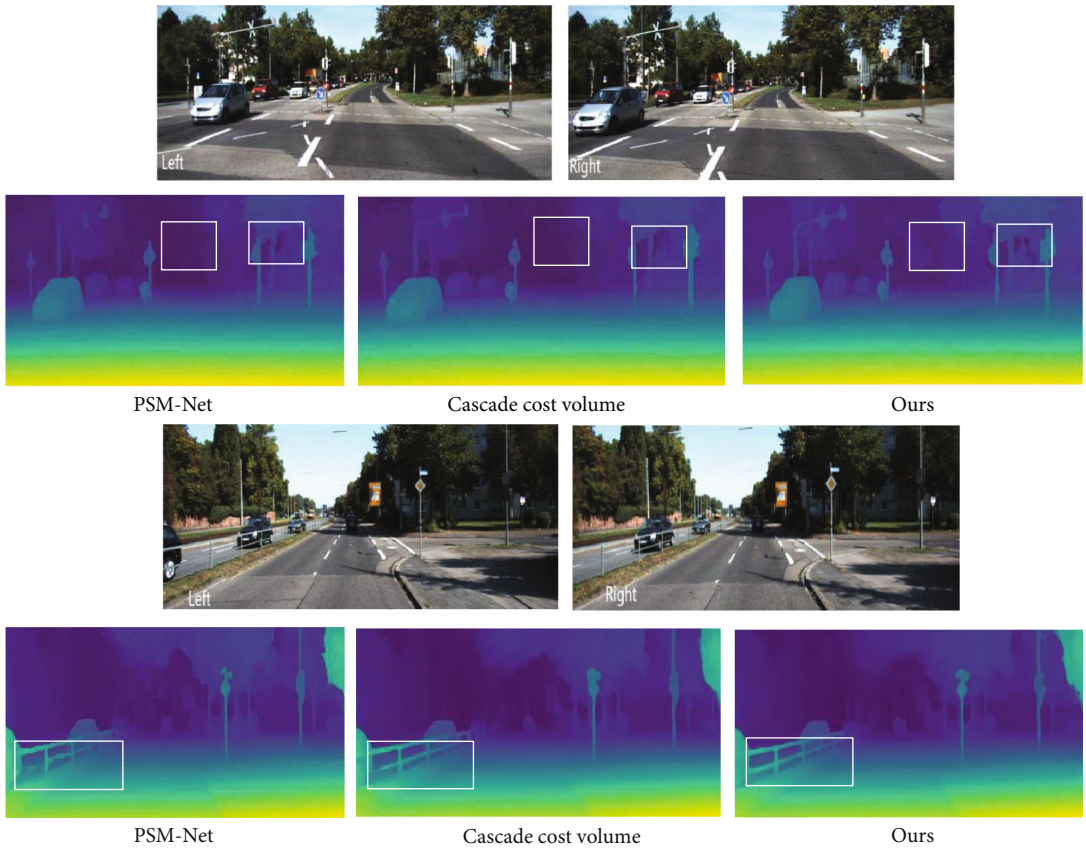

Figure 7: Results of disparity estimation for KITTI 2015 images. The top panel shows the input stereo image pair. For each input image, the disparity maps were obtained by PSM-Net, cascade cost volume, and our network.

Table 4: KITTI 2015 stereo matching network test results.

Model	EPE	>3px	Run time (s)
PSM-Net	0.88	**2.80**	1.53
GWCNet	0.74	3.20	1.43
Cascade cost volume	0.72	3.04	1.81
GC-Net	1.02	2.97	2.53
Ours	**0.68**	2.91	**1.06**

The bold data represents the optimal data result under this column.

Net and the Cascade Network. The matching results are reported by the KITTI evaluation server, and the best result can be obtained when the interval is 2. The GASPP network realizes robust results with less running time. It implies that in practical application such as mobile robots, our network will display fine real-time performance. Table 4 shows the comparison between our network and other stereo matching networks on the KITTI 2015 datasets.

6. Conclusions

This paper presents a group-based atrous convolution pyramid pooling module, which uses a densely atrous convolution to form multiscale receptive fields, reduce the loss of local information, and improve the matching accuracy. The dilation rate of GASPP is continuous to lower the hole part generated in an original atrous convolution operation. The dilation rate between groups is gradually increased to ensure the aggregation of multiscale contextual information to improve the accuracy of disparity estimation. Furthermore, we propose a tailored cascaded cost volume and add a 2D convolution to form a layered structure, to reduce memory usage and speed up the execution efficiency of our network. The experimental results demonstrate the effectiveness of the proposed method in terms of both disparity estimation accuracy and running efficiency. In our future work, as most of the end-to-end algorithms rely on datasets for learning, training, and verification, without using the images in real life, we would further investigate our test on data collected from real-world applications. To date, weak-supervised and multitask fusion networks have developed rapidly. Weak-supervised networks do not depend on large-scale datasets with ground truth values, which makes network training easier. The multitask fusion network is extraordinarily helpful for matching accuracy in small or ill-posed areas, which will be the focus on follow-up research.

Data Availability

The data used to support the findings of this study are available from the authors upon request.

Conflicts of 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 work was supported by the Dalian University Research Platform Project Funding: Dalian Wise Information Technology of Med and Health Key Laboratory, and the National Natural Science Foundation of China (No. 11701061): Research on SA Algorithm for Nonconvex Stochastic Semidefinite Programming.

References

[1] J. D. S. Solak and E. D. Bolat, "A new hybrid stereovision-based distance-estimation approach for mobile robot platforms," *Computers & Electrical Engineering*, vol. 67, pp. 672–689, 2018.

[2] L. I. Xiu-Juan, W. Liu, and L. I. Shan-Hong, "Robust control algorithm of bionic robot based on binocular vision navigation," *Computer Science*, vol. 21, no. 44, pp. 318–322, 2017.

[3] T. Trzcinski, M. Christoudias, P. Fua, and V. Lepetit, "Boosting binary keypoint descriptors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, CVPR2013 Portland, American, 2013.

[4] Z. Kun, M. Xiangxi, and B. Cheng, "Review of stereo matching algorithms based on deep learning," *Computational Intelligence and Neuroscience*, vol. 2020, 12 pages, 2020.

[5] H. Wang, M. Wu, Y. Zhang, and L. Zhang, "Effective stereo matching using reliable points based graph cut," in *2013 Visual Communications and Image Processing (VCIP)*, pp. 1–6, Kuching, Malaysia, 2013.

[6] A. Klaus, M. Sormann, and K. Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," in *18th International Conference on Pattern Recognition (ICPR'06)*, pp. 15–18, Hong Kong, China, 2006.

[7] Y. Ji, Y. Li, X. Sun, S. Yan, and N. Guo, "Stereo matching algorithm based on binocular vision," in *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, pp. 843–847, Hefei, China, 2020.

[8] S. Mohammad and T. Morris, "Binary robust independent elementary feature features for texture segmentation," *Advanced Science Letters*, vol. 23, no. 6, pp. 5178–5182, 2017.

[9] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Minneapolis, MN, USA, 2007.

[10] K.-J. Yoon and I.-S. Kweon, "Locally adaptive support-weight approach for visual correspondence search," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, pp. 924–931, San Diego, CA, USA, 2005.

[11] J. Yang, D. Xing, Z. Hu, and T. Yao, "A two-branch network with pyramid-based local and spatial attention global feature learning for vehicle re-identification," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 1, pp. 46–54, 2021.

[12] J. Žbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1592–1599, CVPR 2015 Boston, MA, USA, 2015.

[13] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[14] A. Kendall, H. Martirosyan, S. Dasgupta et al., "End-to-end learning of geometry and context for deep stereo regression," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 66–75, ICCV2017 Venice, Italy, 2017.

[15] J. Chang and Y. Chen, "Pyramid stereo matching network," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5410–5418, CVPR2018 Salt Lake City, UT, USA, 2018.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[17] X. Song, X. Zhao, L. Fang, H. Hu, and Y. Yu, "EdgeStereo: an effective multi-task learning network for stereo matching and edge detection," *International Journal of Computer Vision*, vol. 128, no. 4, pp. 910–930, 2020.

[18] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2016, https://arxiv.org/abs/1511.07122.

[19] H. Dai, X. Zhang, Y. Zhao, H. Sun, and N. Zheng, "Adaptive disparity candidates prediction network for efficient real-time stereo matching," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, p. 1, 2021.

[20] J. Pang, W. Sun, J. S. Ren, C. Yang, and Q. Yan, "Cascade residual learning: a two-stage convolutional neural network for stereo matching," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 878–886, ICCVW2017 Venice, Italy, 2017.

[21] F. Zhang, V. Prisacariu, R. Yang, and P. H. S. Torr, "GA-Net: guided aggregation net for end-to-end stereo matching," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 185–194, CVPR2019 Long Beach, CA, USA, 2019.

[22] X. Jia, W. Chen, and Z. Liang, "Bidirectional stereo matching network with double cost volumes," *IEEE Access*, vol. 9, pp. 19651–19658, 2021.

[23] Z. Liang, Y. Guo, Y. Feng et al., "Stereo matching using multi-level cost volume and multi-scale feature constancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 300–315, 2021.

[24] G. Y. Nie, M. M. Cheng, Y. Liu et al., "Multi-level context ultra-aggregation for stereo matching," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3278–3286, CVPR2019 Long Beach, CA, USA, 2019.

[25] G. Yang, J. Manela, M. Happold, and D. Ramanan, "Hierarchical deep stereo matching on high-resolution images," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5510–5519, CVPR2019 Long Beach, CA, USA, 2019.

[26] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, *Rethinking atrous convolution for semantic image segmentation*, 2018.

[27] Q. Du, R. Liu, Y. Pan, S. Sun, S. Sun, and Z. Zheng, "Depth estimation with multi-resolution stereo matching," in *2019 IEEE Visual Communications and Image Processing (VCIP)*, pp. 1–4, Sydney, NSW, Australia, 2019.

[28] P. Wang, P. Chen, Y. Yuan et al., "Understanding convolution for semantic segmentation," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1451–1460, Lake Tahoe, NV, USA, 2018.

[29] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3684–3692, CVPR2018 Salt Lake City, UT, USA, 2018.

[30] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[31] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2492–2501, CVPR2020 Seattle, WA, USA, 2020.

[32] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3268–3277, CVPR2019 Long Beach, CA, USA, 2019.

[33] C. Lu, H. Uchiyama, D. Thomas, A. Shimada, and R.-i. Taniguchi, "Sparse cost volume for efficient stereo matching," *Remote Sensing*, vol. 10, no. 11, p. 1844, 2018.

[34] G. Huang, Y. Gong, Q. Xu, K. Wattanachote, K. Zeng, and X. Luo, "A convolutional attention residual network for stereo matching," *IEEE Access*, vol. 8, pp. 50828–50842, 2020.

[35] N. Mayer, E. Ilg, P. Hausser et al., "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4040–4048, CVPR2016 Las Vegas, NV, USA, 2016.

[36] M. Menze, C. Heipke, and A. Geiger, "Joint 3d estimation of vehicles and scene flow," *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences*, vol. II-3/W5, pp. 427–434, 2015.