Chinese Journal of Electronics
vol. 33, no. 4, pp. 1010–1022, July 2024
https://doi.org/10.23919/cje.2022.00.338

RESEARCH ARTICLE

Multi-Scale Binocular Stereo Matching Based on Semantic Association

Jin ZHENG[1,2], Botao JIANG[2], Wei PENG[2], and Qiaohui ZHANG[2]

1. *State Key Laboratory of Virtual Reality Techonology and Systems, Beihang University, Beijing 100191, China*
2. *School of Computer Science and Engineering, Beihang University, Beijing 100191, China*

Corresponding author: Jin ZHENG, Email: JinZheng@buaa.edu.cn

Manuscript Received October 12, 2022; Accepted November 10, 2023

Copyright © 2024 Chinese Institute of Electronics

Abstract — Aiming at the low accuracy of existing binocular stereo matching and depth estimation methods, this paper proposes a multi-scale binocular stereo matching network based on semantic association. A semantic association module is designed to construct the contextual semantic association relationship among the pixels through semantic category and attention mechanism. The disparity of those regions where the disparity is easily estimated can be used to assist the disparity estimation of relatively difficult regions, so as to improve the accuracy of disparity estimation of the whole image. Simultaneously, a multi-scale cost volume computation module is proposed. Unlike the existing methods, which use a single cost volume, the proposed multi-scale cost volume computation module designs multiple cost volumes for features of different scales. The semantic association feature and multi-scale cost volume are aggregated, which fuses the high-level semantic information and the low-level local detailed information to enhance the feature representation for accurate stereo matching. We demonstrate the effectiveness of the proposed solutions on the KITTI2015 binocular stereo matching dataset, and our model achieves comparable or higher matching performance, compared to other seven classic binocular stereo matching algorithms.

Keywords — Binocular stereo matching, Multi-scale features, Semantic association, Depth estimation, Attention mechanism.

Citation — Jin ZHENG, Botao JIANG, Wei PENG, *et al.*, “Multi-Scale Binocular Stereo Matching Based on Semantic Association,” *Chinese Journal of Electronics*, vol. 33, no. 4, pp. 1010–1022, 2024. doi: 10.23919/cje.2022.00.338.

I. Introduction

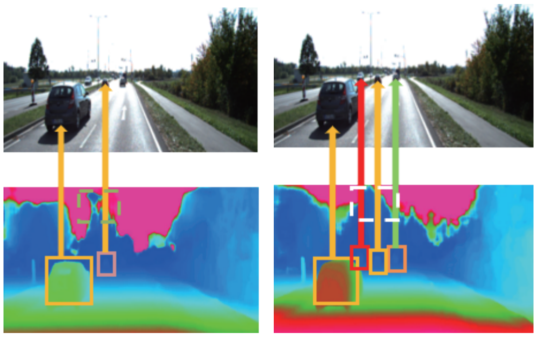
Depth estimation from RGB images has been studied for many years, and stereo matching is one of the most widely used solutions [1]–[3] because it is closely related to the human binocular vision system. Stereo matching of binocular images mainly provides dense matching pairs for the left-view and right-view images, and thus estimates the disparity of each pixel in the reference image according to the matching pairs. Then, the disparity map can be converted into the depth map, which is applied to many scenarios. It is of great significance in the fields of autonomous driving, virtual reality, 3D model reconstruction, 3D object detection and recognition, etc. [4], [5].

The existing stereo matching algorithms of binocular images are divided into traditional methods [6], [7] and deep learning methods [8]–[10]. In traditional methods, stereo matching is achieved by matching manual features in left-view and right-view images through global cost aggregation [11], [12] or local cost aggregation [13], [14]. Although a lot of research has been done, traditional methods have poor matching results in ill-posed regions, such as occlusion areas, repeated patterns, weak texture regions, and reflective surfaces [15], [16]. In addition, the processing of traditional stereo matching algorithms is complex. Recently, with the development of deep convolution neural networks, binocular stereo matching based on deep learning has been widely studied, which yields significant gains compared to traditional methods in accuracy and speed [17]–[20].

For ill-posed regions, solely applying the local pixel-based consistency constraints in support window between the matching pairs of different viewpoints is insufficient for accurate correspondence estimation, and some researchers proposed that regional support from contex-

Associate Editor: Prof. Hongkai XIONG, Shanghai Jiao Tong University.

tual information must be incorporated into the consistency constraints of stereo matching to improve the accuracy [17], [21]. However, the matching result of some extremely difficult regions, such as textureless regions and small objects, is still poor. For example, Figure 1(a) shows the disparity estimation results using PsmNet [17], which is a pyramid stereo matching network incorporating global contextual information into pixel-based image features and extending the regional support of contextual information in the computation of cost volume. Obviously, PsmNet can estimate the disparity of vehicle in near-distance, but the different parts of the vehicle at different distances from the camera have similar disparities. To make matters worse, it is not effective for small vehicles at a far distance, which poses a greater challenge for disparity estimation. The disparity of the region where the small object is located cannot be well distinguished from the disparity of background. In addition, for the textureless region, such as the sky, PsmNet cannot handle well either, and the estimated disparity boundary between the sky and the ground is not smooth, as shown in the dotted box. Improving the accuracy of stereo matching for complex scenes is still a challenge.


(a) PsmNet [17]

(b) Our method

Figure 1 Visual comparison of disparity estimation. Compared to PsmNet, our method obtains better estimation results (near vehicles) and identifies distant vehicles (three small vehicles in the far-distance).

Aiming at the problem, this paper proposes a multi-scale binocular stereo matching network based on semantic association, and attempts to use the semantic association between the easy region and difficult region to improve the stereo matching accuracy of the difficult region. Based on the analysis of image characteristics and existing problems, PsmNet, as an end-to-end learning framework, is adopted as the basic network. Then, a semantic association module is proposed, which uses the semantic segmentation result and combines the attention mechanism to obtain the semantic association among pixels. Thus, through semantic association, the disparity of those regions where the disparity is easily estimated can be used to assist the disparity estimation of relatively difficult regions. Simultaneously, a multi-scale cost volume computation module is proposed, which constructs multiple cost volumes instead of a single cost volume to distinguish the contribution of different scale features. Furthermore, the semantic association feature and multi-scale cost volume are aggregated, which fuses the high-level semantic information and the low-level local detailed information to enhance the feature representation, and is beneficial to disparity estimation of different scale objects. Finally, the effectiveness of the proposed method is illustrated. For example, as shown in Figure 1(b), the proposed method can estimate the disparity of foreground pixels better, which is very meaningful for subsequent 3D object detection tasks. In our disparity estimation results, not only the different parts of the vehicle in near-distance have different disparities, but also the far-distance small vehicles can be distinguished from the background. Although the boundary of disparity results is still not smooth in some weak texture regions, such as the junction area of sky and ground, its impact on subsequent tasks is relatively small. More generally, acting on the KITTI2015 official evaluation set, the proposed method has significant improvement compared with other state-of-the-art (SOTA) methods, such as PsmNet [17] and SGNet [22].

II. Related Work

The related work is analyzed, including the main steps of stereo matching, context and semantics embedding, attention mechanism, and multi-scale features for cost volume construction.

1. Main steps of stereo matching

A typical stereo matching algorithm consists of four steps: matching cost computation, cost aggregation, disparity computation, and disparity refinement [23].

Matching cost calculation It is used to measure the correlation between a pixel and other pixels. It is critical to choose an appropriate cost calculation method for features in multi-scale and multi-channel [24].

Cost aggregation It constructs the correlation between adjacent pixels and achieves a global optimization for accurate matching cost. Through cost aggregation, such as the aggregation constraint that the disparities of continuous pixels are also continuous, the matching cost calculation matrix is optimized, so that the cost value can be adjusted according to other adjacent pixels, and a more accurate cost value can be obtained. Cost aggregation can integrate the disparity of other pixels, especially suitable for regions with high noise and weak texture [25].

Disparity computation It carries out disparity calculation usually based on the WTA (winner takes all) algorithm. According to the cost matrices, the cost values of each pixel under all possible disparities are obtained, and then the disparity corresponding to the smallest cost is taken as the optimal result.

Disparity refinement It is a post-processing step of disparity calculation, aiming to improve and refine the disparity results. Filtering operation [26], consistency constraint check [27], and outlier removal [28] are generally used to improve the accuracy of disparity estimation.

Although the above processes can be carried out step by step, and each module can be trained independently of other ones, more attention has been paid to the end-to-end framework, which integrates these operations into a network, so as to allow an end-to-end training and return the final disparity map.

2. Context or semantics embedding

Considering that images can provide rich context and semantics, some stereo matching methods use context or semantics to improve the accuracy of stereo matching. Typical algorithms that combine contextual or semantic information to refine disparity maps include PsmNet [17], GC-Net [29], NLCA-Net [21], Displets [30], SegStereo [31], DispSegNet [32], SSPCV-Net [33], SGNet [22], PGNet [34], and so on.

Among them, PsmNet [17] proposes spatial pyramid pooling (SPP) and stacked hourglass 3D convolutional neural network (CNN) architecture. The pyramid pooling structure takes advantage of the capacity of global information by aggregating context in different scales and locations to form a cost volume, and the 3D CNN learns to regularize cost volume using stacked multiple hourglass networks in conjunction with intermediate supervision. GC-Net [29] incorporates contextual information using 3D convolutions over the cost volume, and a differentiable soft argmin operation regresses disparity values from the cost volume. NLCA-Net [21] designs a non-local context attention module to exploit the global contextual information for regularizing the cost volume, and uses a variance-based method instead of traditional concatenate operation to build the cost volume. Furthermore, Displets [30] obtains 3D models of vehicles, which are semantic embedding, to resolve matching ambiguities in reflective and textureless regions. SegStereo [31] aggregates the left segmentation feature map into a disparity branch as semantic feature embedding, and warps the right segmentation feature map to the left view for per-pixel semantic prediction with softmax [35] loss regularization, which incorporates semantic information to improve the accuracy of disparity estimation. DispSegNet [32] utilizes pyramid scene parsing (PSP) to obtain rich semantic information for auxiliary segmentation tasks, and embeds PSP as contextual information into the disparity computation module to improve the accuracy of stereo matching. SSPCV-Net [33] proposes semantic stereo matching with pyramid cost volumes, including pyramid cost volumes for describing semantic and spatial information on multiple levels. The semantic features are inferred by a semantic segmentation sub-network while the spatial features are derived by hierarchical spatial pooling. SGNet [22] considers that high-level semantic information can be helpful to handle accurate disparity estimation in low texture and illumination changes scene, and proposes semantics guided deep stereo matching. PGNet [34] proposes a panoptic parsing guided deep network, which provides valuable high-level scene clues, including semantic and instance segmentation, to tackle these challenges, such as low texture, occlusion, or large illumination changes. These methods embed semantics from the semantic features or geometric layouts.

These above networks consider the effectiveness of contextual and semantic information for binocular stereo matching, but the use of contextual semantic association is not sufficient. Actually, semantic information is not independent, and there is a semantic association among pixels, which can assist in disparity estimation, especially for ill-posed regions. How to build the contextual semantic association between easy regions and ill-posed regions, and thus, how to estimate the disparity of ill-posed regions based on the disparity of those regions where the disparity is easily estimated, is an interesting problem.

3. Attention mechanism

Another study for stereo matching is to introduce the attention mechanism into the disparity estimation network, which can extract more valid features and improve the accuracy of stereo matching. For example, MCANet [36] proposes a multi-scale context attention network with three main modules: atrous spatial pyramid pooling attention, richer convolutional features, and attention mechanism. MRDA-Net [37] uses the 2D residual dense attention network for feature extraction and the 3D convolutional attention network for matching. ACAR-Net [38] introduces a convolutional block attention module (CBAM) [39] combining spatial and channel dimensions into the binocular disparity estimation network, and uses 2D CBAM and 3D CBAM to obtain features in 2D feature extraction and 3D cost aggregation processing, respectively. The features obtained through the attention mechanism have a larger receptive field containing rich contextual feature associations. NLCA-Net [21] introduces the 3D non-local attention [40] in the 3D cost aggregation module to obtain the correlation within the features, but the 3D attention is computationally intensive and time-consuming.

Generally, the existing methods directly use the attention of feature layer to obtain the correlation of features, and do not take into account the combination of attention and semantic segmentation results to build semantic associations among different semantic categories. If the attention mechanism can be combined with semantic segmentation, it will provide more effective information for the calculation of cost volume, so as to improve the matching accuracy.

4. Multi-scale features for cost volume construction

Due to the various distances and sizes of objects, the existing methods often use multi-scale features to construct the cost volume [41], [42]. Image features are computed by deep convolutional networks, which build a multi-scale, different-spatial-resolution feature representation hierarchy layer by layer. Multi-scale means using different receptive fields to observe objects, and different-spatial-resolution means using multi-layer feature maps. The shallow layer high-resolution maps represent low-level features, and the deep layer low-resolution maps represent high-level features. Due to the high-resolution of shallow features, they are conducive to expressing local details. From the shallow layer to the deep layer, the features are more abstract and more able to express semantics. For those pixels with large disparity or in near-distance, they tend to rely on high-level features to provide more semantics; for those pixels with small disparity or in far-distance, besides semantics, they tend to rely on low-level features to provide more local details. Therefore, for the objects with different scales in far or near distance, it is necessary to make full use of multi-scale, high-level, and low-level features to provide rich semantic and detailed information.

The existing methods usually obtain the multi-scale features from the backbone network, thus, fuse these features and construct one cost volume for matching cost computation [43]. But this approach has obvious disadvantages: For multi-scale feature fusion, different scale features are upsampled to restore to the original size, which leads to the loss of detailed features; for cost volume construction, the different scale features have different contributions to disparity calculation of objects with different sizes, and the features of different scales can provide a variety of local details as well as semantic and contextual information for disparity calculation. Hence, considering multi-scale feature fusion, CFNet (cascade fusion network) [44] introduces several cascaded stages to learn multi-scale representations. Through inserting the feature integration operation into the backbone, a large proportion of the whole backbone can be utilized to fuse the multi-scale features effectively. Considering cost volume construction, reference [45] also proposes a cascade and fused cost volume (CFnet). It finds that different scale low-resolution cost volumes can cover multi-scale receptive fields and are complementary to each other, and proposes a fused cost volume representation and a cascade cost volume representation for stereo matching. Actually, only one cost volume based on the fused multi-scale features cannot effectively distinguish the contributions of different scale features. Thus, a fused or a cascade cost volume formulation [24] should be considered. Besides fused or cascaded multi-scale features, ACVNet [46] proposes attention concatenation volume (ACV), which generates attention weights based on similarity measures to filter concatenation volume. Thus, more features are introduced for cost volume construction.

Inspired by CFNet and its variants, we design a multi-scale cost volume computation module to construct multiple cost volumes, which cover multi-scale features and distinguish the contributions of different scale features. More importantly, we aggregate the semantic association feature with multi-scale cost volume to enhance the feature representation and improve the expressiveness of the cost volume, and in turn improve the performance in ill-posed regions.

III. Methodology

The existing binocular stereo matching methods still have notable weaknesses when facing ill-posed regions such as occlusion, textureless regions, and far-distance small objects. This paper analyzes the image characteristics in real outdoor scenes and introduces the semantic association module to construct semantic association among pixels. Thus, according to continuous disparities of associated regions, this paper uses the disparity of those regions where the disparity is easily estimated, such as the ground, to assist the disparity estimation of relatively difficult regions, such as small vehicle objects. At the same time, considering the influence of different scale features on binocular stereo matching, a multi-scale cost volume computation module is designed, which constructs multiple cost volumes for multi-scale features, and the features of different scales are comprehensively aggregated with semantic association feature. These features can provide richer semantic information and local detail information, thereby effectively improving the accuracy of the binocular stereo matching network.

1. Network architecture

The overall structure of the proposed multi-scale

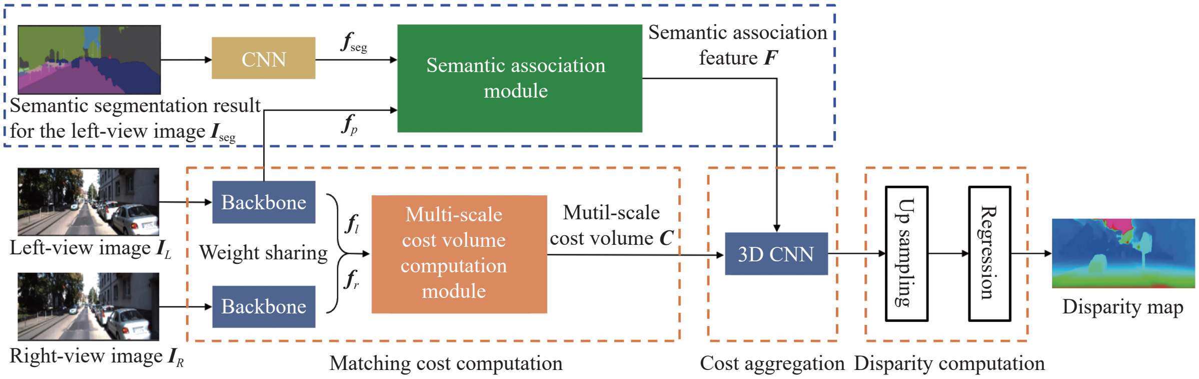

CNN
f_{seg}
Semantic association module
Semantic association feature $\boldsymbol{F}$
Semantic segmentation result for the left-view image $\boldsymbol{I}_{seg}$
f_p
Backbone
Left-view image $\boldsymbol{I}_L$
Weight sharing
f_l
f_r
Multi-scale cost volume computation module
Mutil-scale cost volume $\boldsymbol{C}$
3D CNN
Up sampling
Regression
Disparity map
Backbone
Right-view image $\boldsymbol{I}_R$
Matching cost computation
Cost aggregation
Disparity computation

Figure 2 The network architecture.

binocular stereo matching network based on semantic association is shown in Figure 2. Similar to PsmNet [17], the main process includes three steps: matching cost computation, cost aggregation, and disparity computation. The input of the network is the left-view image $\boldsymbol{I}_L$, the right-view image $\boldsymbol{I}_R$, and the semantic segmentation result $\boldsymbol{I}_{\text{seg}}$ of the left-view image.

First, CNN, as a 2D feature extraction network, is used to extract semantic features $\boldsymbol{f}_{\text{seg}}$ of $\boldsymbol{I}_{\text{seg}}$, which is the segmentation result of the left-view image. In CNN, three small convolution filters (3 × 3) are cascaded to construct a simple network with the same receptive field. The output feature map size is (1/4) × (1/4) of the input image size. The detailed parameters are shown in Table 1.

Table 1 Parameters of the proposed CNN and backbone network

Name	Layer setting	Output dimension
$\boldsymbol{I}_L$	–	$H \times W \times 3$
$\boldsymbol{I}_R$	–	$H \times W \times 3$
$\boldsymbol{I}_{\text{seg}}$	–	$H \times W \times 1$
CNN for $\boldsymbol{I}_{\text{seg}}$		
conv0_1	$3 \times 3, 32$	$\frac{1}{2}H \times \frac{1}{2}W \times 32$
conv0_2	$3 \times 3, 32$	$\frac{1}{2}H \times \frac{1}{2}W \times 32$
conv0_3	$3 \times 3, 32$	$\frac{1}{4}H \times \frac{1}{4}W \times 32$
Backbone		
conv0_1	$3 \times 3, 32$	$\frac{1}{2}H \times \frac{1}{2}W \times 32$
conv0_2	$3 \times 3, 32$	$\frac{1}{2}H \times \frac{1}{2}W \times 32$
conv0_3	$3 \times 3, 32$	$\frac{1}{2}H \times \frac{1}{2}W \times 32$
conv1_x	$\begin{bmatrix}3 \times 3, 64\\3 \times 3, 64\end{bmatrix} \times 4$	$\frac{1}{2}H \times \frac{1}{2}W \times 32$
conv2_x	$\begin{bmatrix}3 \times 3, 64\\3 \times 3, 64\end{bmatrix} \times 9$	$\frac{1}{4}H \times \frac{1}{4}W \times 64$
conv3_x	$\begin{bmatrix}3 \times 3, 128\\3 \times 3, 128\end{bmatrix} \times 4$	$\frac{1}{4}H \times \frac{1}{4}W \times 128$
conv4_x	$\begin{bmatrix}3 \times 3, 128\\3 \times 3, 128\end{bmatrix} \times 4$, dila $= 2$	$\frac{1}{4}H \times \frac{1}{4}W \times 128$
Backbone-Extract $\boldsymbol{f}_p$		
attconv	$3 \times 3, 32$	$\frac{1}{4}H \times \frac{1}{4}W \times 32$

Note: The construction of residual blocks is designated in brackets with the number of stacked blocks. Down-sampling is performed by conv0_1 (backbone), conv2_x (backbone), conv0_1 (CNN), and conv0_3 (CNN) with stride of 2. H and W denote the height and width of the input image, respectively.

Moreover, another backbone network is used for feature extraction acting on $\boldsymbol{I}_L$ and $\boldsymbol{I}_R$. Here, Resnet50, Resnet101, Vggnet, Densenet, or U-net can be used as the backbone. In our experiment, Resnet50 is adopted. The parameters of the backbone network used in our experiment are also shown in Table 1. The conv1_x, conv2_x, conv3_x, and conv4_x are the basic residual blocks for learning the unary feature extraction. For conv4_x, dilated convolution is applied to further enlarge the receptive field. Finally, a half dilation rate (1, 2) is used, like PsmNet. Thus, the multi-scale 2D features, $\boldsymbol{f}_l$ and $\boldsymbol{f}_r$, containing contextual information are obtained. In addition, $\boldsymbol{I}_L$ and $\boldsymbol{I}_R$ share the weights in the backbone network.

Second, $\boldsymbol{f}_l$ and $\boldsymbol{f}_r$ are sent to the multi-scale cost volume computation module, which is proposed in this paper. In this module, the multiple cost volumes are constructed according to the input multi-scale features ($\boldsymbol{f}_l$ and $\boldsymbol{f}_r$), and the output is the final multi-scale cost volume $\boldsymbol{C}$.

Meanwhile, $\boldsymbol{f}_{\text{seg}}$ and $\boldsymbol{f}_p$ are fed into the proposed semantic association module together to get the semantic-association feature $\boldsymbol{F}$. Here, the low-level image feature $\boldsymbol{f}_p$ is the result of applying an extra simple convolution operation on the output of the backbone, which means the input of the extra simple convolution is the result of conv4_x. The parameters of extra convolution are shown in Table 1 (Backbone-Extract $\boldsymbol{f}_p$).

Third, the subsequent 3D CNN module performs cost aggregation based on the multi-scale cost volume $\boldsymbol{C}$ and the semantic-association feature $\boldsymbol{F}$ provided by the multi-scale cost volume computation module and semantic association module, respectively. $\boldsymbol{F}$ is imposed on $\boldsymbol{C}$ as weights, which enhances the feature representations. The 3D CNN module includes a stacked hourglass encoder-decoder structure, and this structure can learn richer contextual information to refine the cost volume, and thus, the 3D CNN module outputs the refined cost volume.

Finally, for disparity computation, all the possible disparity values d for each pixel correspond to a disparity cost c_d. After the processing of the softmax layer, the disparity cost c_d is converted into a probability value. The final disparity prediction value $\hat{d}$ for each pixel is calculated as follows:

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

where d represents a specific value within the disparity range, c_d is the corresponding disparity cost when the disparity value is d, and $D_{\max}$ denotes the maximum value of disparity. $\sigma(\cdot)$ represents the softmax operation. Therefore, the disparity map is regressed.

2. Semantic association module

Most existing binocular stereo matching algorithms directly use channel attention or spatial attention at the feature level to construct the correlation among features. This paper proposes a semantic association module, which combines semantic segmentation results with multi-scale features and adopts a self-attention mechanism to obtain contextual information and semantic association among pixels. The detailed design of the proposed semantic association module is shown in Figure 3.

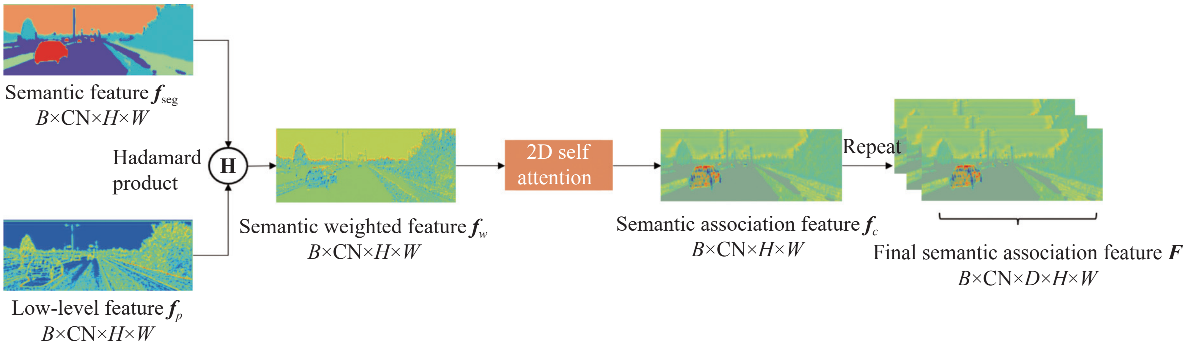

Semantic feature f_{seg}
B×CN×H×W
Hadamard product
H
Low-level feature f_p
B×CN×H×W
Semantic weighted feature f_w
B×CN×H×W
2D self attention
Semantic association feature f_c
B×CN×H×W
Repeat
Final semantic association feature F
B×CN×D×H×W

Figure 3 Semantic association module.

First, the semantic feature $\boldsymbol{f}_{\text{seg}}$ containing semantic category is obtained from the semantic segmentation result, and it is used as the input of the semantic association module together with the low-level feature $\boldsymbol{f}_p$ obtained from the left-view image. Then, Hadamard product operation is performed on $\boldsymbol{f}_{\text{seg}}$ and $\boldsymbol{f}_p$ to get the semantic weighted feature $\boldsymbol{f}_w$. The calculation formula is

$$\boldsymbol{f}_w = \boldsymbol{f}_{\text{seg}} \circ \boldsymbol{f}_p \tag{2}$$

According to the visualized feature map, the low-level image feature $\boldsymbol{f}_p$ extracted from RGB image has a strong positioning ability for details, while being relatively cluttered. The semantic feature $\boldsymbol{f}_{\text{seg}}$ is suitable for observing the outline of the objects, such as small vehicles, while being weak in identifying the internal features of the object. The Hadamard product achieves semantic weighting for image features, which embeds high-level semantics into shallow image features.

Second, 2D self-attention is adopted to get the association among the features of $\boldsymbol{f}_w$. The calculation formula is as follows:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{3}$$

where $\boldsymbol{Q} = W_q \boldsymbol{f}_w$, $\boldsymbol{K} = W_k \boldsymbol{f}_w$, and $\boldsymbol{V} = W_v \boldsymbol{f}_w$. d_k represents the feature dimension. W_q, W_k, and W_v represent different weight parameters, respectively. The above calculation can be carried out through a non-local network [40]. The detailed network structure and tensor size are shown in Figure 4. θ, ϕ, and g are three 1×1 convolutions. "$\otimes$" denotes matrix multiplication, and "$\oplus$" denotes element-wise sum.

Through the attention mechanism, an internal association is constructed for the semantic weighted feature $\boldsymbol{f}_w$, and a temporary semantic association feature $\boldsymbol{f}_c$ is obtained. Each pixel in $\boldsymbol{f}_c$ can autonomously fuse the features of other pixels that are beneficial to its disparity estimation according to the surrounding contextual and semantic information.

Finally, the temporary semantic association feature $\boldsymbol{f}_c$ is expanded in the disparity dimension through the repeat operation, and the dimension is changed from $B \times \text{CN} \times H \times W$ to $B \times \text{CN} \times D \times H \times W$. Here, B is the batch size, CN is the number of channels, and D is the maximum disparity value. W and H are the width and height of the input image, respectively. In this way, the dimension of the final semantic association feature $\boldsymbol{F}$ is consistent with the dimension of the cost volume $\boldsymbol{C}$.

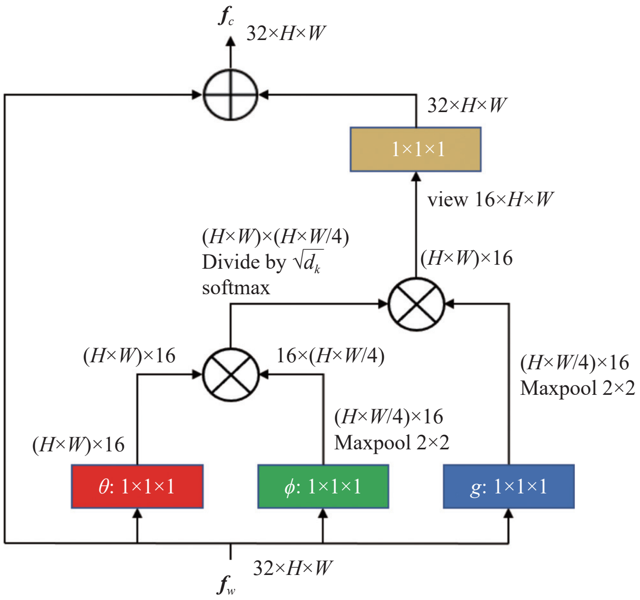


Figure 4 Non-local network.

In the semantic association module, the semantic weighted feature $\boldsymbol{f}_w$ distinguishes the different categories of objects through semantic labels, and the semantic association feature $\boldsymbol{f}_c$ establishes the association relationship between objects through the self-attention mechanism. Finally, the obtained $\boldsymbol{F}$ combines the advantages of high-level semantic features and low-level detailed features, implying explicit semantic association and accurate location. Therefore, $\boldsymbol{F}$ is used to assist the generation of refined cost volumes in the subsequent 3D CNN cost aggregation module.

3. Multi-scale cost volume computation module

The sizes of objects are varied, so it is necessary to obtain multi-scale features and consider the influence of different scale features on disparity estimation, as well as perform multi-scale cost volume computation. The proposed multi-scale cost volume computation module is shown in Figure 5.

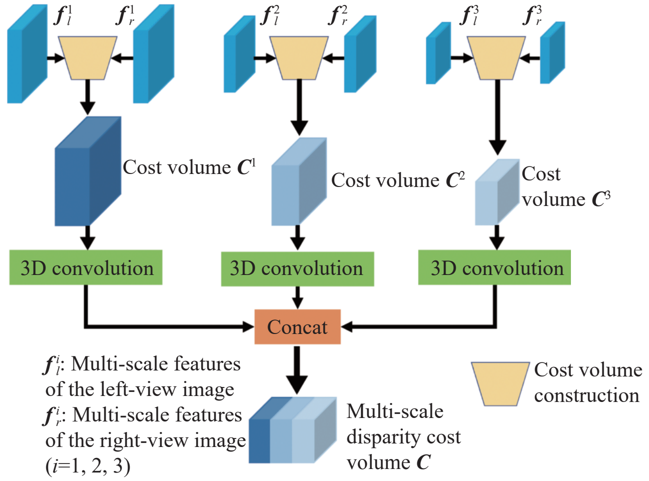


Figure 5 Multi-scale cost volume construction module.

For the multi-scale features $\boldsymbol{f}_l$ and $\boldsymbol{f}_r$ corresponding to the left-view image $\boldsymbol{I}_L$ and the right-view image $\boldsymbol{I}_R$, they both have multiple scales. Typically, they have three scales and can be denoted as $\boldsymbol{f}_l^i$ and $\boldsymbol{f}_r^i$ $(i=1,2,3)$. The cost volumes are constructed separately for different scale features. Through the cost volume construction, three cost volumes $\boldsymbol{C}^1$, $\boldsymbol{C}^2$, and $\boldsymbol{C}^3$ containing different scale features can be obtained. In order to fuse these multi-scale features, this paper uses 3D convolutions to reduce the dimension of cost volumes and then carries out concat operation along the channel dimension to obtain the final multi-scale cost volume $\boldsymbol{C}$, which combines the multi-scale features and has stronger representation ability. The calculation formula is

$$\boldsymbol{C}(x,y,D,G)=\text{Concat}\{\text{Conv3d}\{\boldsymbol{C}^i(x,y,D,g_i)\}\},\quad i=1,2,3 \tag{4}$$

Take the ith cost volume as an example, and the detailed cost volume construction process is shown in Figure 6.

First, the feature similarity is calculated in groups on all possible disparity values. The calculation formula is expressed as

$$\textbf{Corr}^i(x,y,d,g_i)=\frac{1}{(N_{\text{cn}}/Ng_i)}\langle\boldsymbol{f}_l^i(x,y),\boldsymbol{f}_r^i(x-d,y)\rangle \tag{5}$$

where x and y represent the width and height of the feature, respectively, d represents a specific value within the disparity range, g_i represents the number of groups when constructing the ith cost volume, N_{cn} represents the number of feature channels, and N_{g_i} represents the number of channels for each group of features. In addition, $\langle\boldsymbol{f}_l^i(x,y),\boldsymbol{f}_r^i(x-d,y)\rangle$ denotes the similarity computation.

Second, the 4D cost volume of $H\times W\times D\times G$ is obtained through concat operation in different disparity dimensions, and the result is

$$\boldsymbol{C}^i(x,y,D,g_i)=\text{Concat}\{\textbf{Corr}^i(x,y,1,g_i),\textbf{Corr}^i(x,y,2,g_i),\ldots,\textbf{Corr}^i(x,y,D,g_i)\} \tag{6}$$

Here, D represents the maximum disparity value and $D=192$ in the experiment.

IV. Experiments

This part shows the advantages of the proposed method through ablation experiments and comparison with other methods.

1. Experimental setting

Datasets The KITTI dataset [47] is one of the commonly used evaluation datasets for computer vision algorithms in the field of autonomous driving. In particular, KITTI2015, a binocular stereo matching dataset,

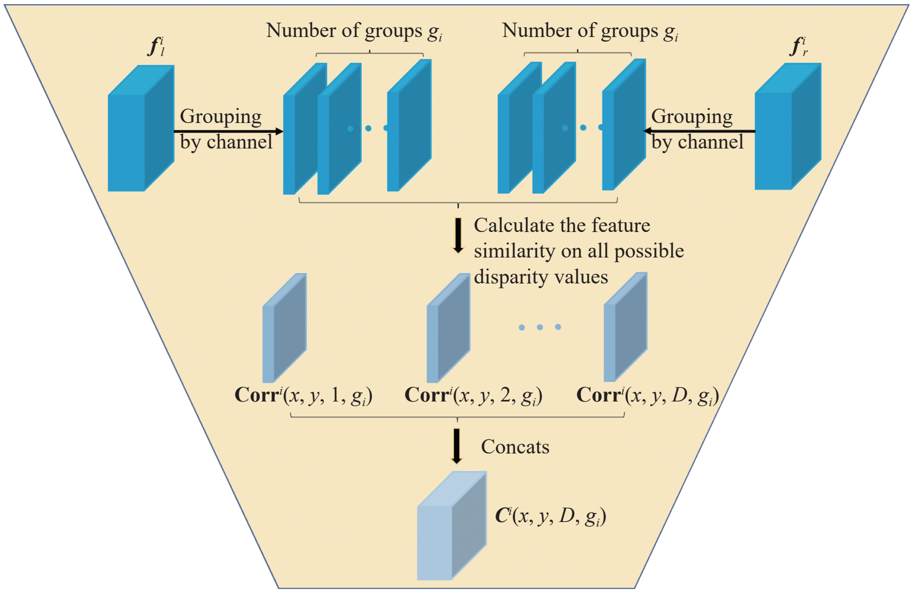
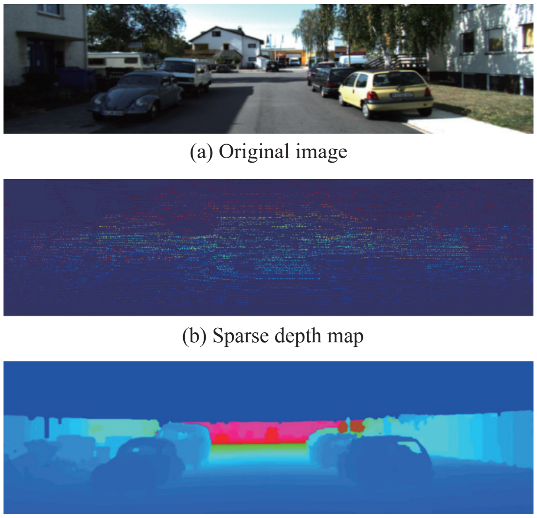

f_l^i
Number of groups g_i
Number of groups g_i
f_r^i
Grouping by channel
Grouping by channel
Calculate the feature similarity on all possible disparity values
$\textbf{Corr}^i(x, y, 1, g_i)$
$\textbf{Corr}^i(x, y, 2, g_i)$
$\textbf{Corr}^i(x, y, D, g_i)$
Concats
$C^i(x, y, D, g_i)$

Figure 6 Multi-scale cost volume construction module.

provides a total of 400 pairs of binocular images, of which 200 pairs have dense disparity map annotation. The other 200 pairs are used as the official algorithm evaluation and their annotation information is not disclosed. For the 200 pairs of binocular images with public annotation information, there are two common ways to divide the training set and the verification set. One is the division of PsmNet [17], with 160 as the training set and 40 as the verification set; the other is the division of GwcNet [19], with 180 as the training set and 20 as the verification set. In this paper, the data division way is consistent with the comparison method.

Considering insufficient image pairs in the training set of KITTI2015, this paper expands the training set by using the KITTI3D object detection dataset, which provides binocular images and corresponding laser point clouds. We project laser point clouds onto RGB images to obtain sparse depth maps. In addition, the depth completion algorithm [48] is used to complete the depth, and thus, a dense depth map is obtained. Furthermore, according to the camera parameters, the depth map can be transformed into a disparity map, which is regarded as the ground truth of disparity estimation. We processed 3712 pairs of binocular images in the KITTI3D object detection dataset and obtained 3712 disparity maps. A typical generated depth map is shown in Figure 7.


(a) Original image

(b) Sparse depth map

(c) Completed depth map

Figure 7 A depth map after depth completion.

In our experiment, the training set we used is different from PsmNet's. For our proposed network, these 3712 image pairs in KITTI3D were used for the pre-training, and 200 image pairs in KITTI2015 were used for fine-tuning. For PsmNet, the model is pre-trained with Scene Flow data and fine-tuned on the KITTI2015 training set. The reason why we choose KITTI3D instead of Scene Flow for pre-training is that PsmNet is mainly used for disparity estimation, but we believe that the ultimate purpose of disparity estimation is for downstream tasks, such as 3D object detection, especially in autonomous driving. Hence, we focus on the autonomous driving scenario, that is, we use the KITTI3D dataset for pre-training, hoping to obtain a disparity estimation model that is more suitable for the autonomous driving scenario.

Evaluation metrics The evaluation index adopts the D1 index, which is the percentage of disparity estimation outliers in the total pixels. In detail, D1-all, D1-fg, D1-bg, and D1-car represent the proportion of outliers in all pixels, in foreground pixels, in background pixels, and in vehicle pixels, respectively. For a pixel, if the absolute error of disparity estimation is greater than or equal to 3 pixels, or the relative error is greater than or equal to 5%, the pixel is regarded as an outlier. In addition, if just counting the pixels in non-occluded region as the total pixels, we can get "Noc" results. For the D1 index, the lower D1, the better matching results.

Training settings All the comparison experiments have the same hyper-parameter settings. Adam [49] is used as the optimizer, and its parameters are set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The batch size of the training is set to 12. First, 50 epochs are trained on the 3712 disparity maps, which are generated based on KITTI3D object detection dataset, and the learning rate is set to 0.001. Then, 300 epochs are fine-tuned on KITTI2015 training set. The learning rate for the first 200 epochs is set to 0.001, and the learning rate for the last 100 epochs is set to 0.0001, that is, the learning rate decreases to 1/10 of the original after 200 epochs. The input images are uniformly cropped to 256 × 256 pixels.

2. Ablation study about pre-training dataset

Although we prefer to use KITTI3D for pre-training and KITTI2015 training set for fine-tuning, for fair comparison, we also pre-train our model using Scene Flow dataset, which is consistent with PsmNet.

In detail, we pre-train our model using Scene Flow dataset for 10 epochs, and fine-tune using KITTI2015 for 300 epochs. This test uses all three datasets of Scene Flow (FlyingThings 3D, Driving, Monkaa). Since our method requires a semantic segmentation result, the pre-training uses the segmentation result provided by Mmsegmentation (OcrNet_hr48_512×1024_160k_cityscapes). All other training parameters are consistent with PsmNet's.

Finally, the stereo matching results of KITTI2015 are submitted to the official website, and the results on the official KITTI2015 evaluation set are compared to PsmNet, which are shown in Table 2.

Obviously, even though our method uses the same pre-training dataset as PsmNet, it can also get better accuracy than PsmNet, indicating the effectiveness of the proposed method. Meanwhile, our method based on "KITTI3D pre-train+KITTI2015 fine-tune" can achieve better performance than our method based on "Scene Flow pre-train+KITTI2015 fine-tune", which means KITTI3D is more effective for pre-training. Therefore, the following ablation experiment and comparison with other methods are conducted based on "pre-trained on KITTI3D and fine-tunned on KITTI2015".

Table 2 Ablation study about pre-training dataset

Method	All (%)↓			Noc (%)↓		
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all
PsmNet[1]	1.86	4.62	2.32	1.71	4.31	2.14
Ours*[2]	1.74	4.31	2.17	1.59	3.80	1.96
Ours[3]	**1.55**	**3.55**	**1.88**	**1.42**	**3.28**	**1.73**

Note: [1] PsmNet [17] (CVPR 2018).
[2] Ours*: Scene Flow pre-train+KITTI2015 fine-tune.
[3] Ours: KITTI3D pre-train+KITTI2015 fine-tune.

3. Ablation study about the semantic association and multi-scale cost volume computation

The proposed semantic association module as well as the multi-scale cost volume computation module can adapt to many existing networks. Typically, this paper conducts ablation experiments on PsmNet [17] and GwcNet [19] respectively, and obtains the experimental results on the KITTI2015 validation set, as illustrated in Table 3.

Table 3 Ablation study on KITTI2015 validation set

Experimental model	D1-car (%)↓	D1-all (%)↓
PsmNet [17] paper results	–	1.83
PsmNet reproduction results	2.484	1.851
PsmNet+semantic association	2.182	1.754
PsmNet+multi-scale computation	1.984	1.653
PsmNet+semantic association+multi-scale computation	**1.842**	**1.607**
GwcNet [19] paper results	–	1.41
GwcNet reproduction results	1.413	1.373
GwcNet+semantic association	1.222	1.229
GwcNet+multi-scale computation	1.183	1.174
GwcNet+semantic association+multi-scale computation	**1.156**	**1.158**

Note: Except for “PsmNet [17] paper results” and “GwcNet [19] paper results”, which use the training dataset mentioned in the corresponding paper, the other methods are pre-trained on KITTI3D and fine-tunned on KITTI2015.

It can be seen that the proposed semantic association module improves the overall accuracy by 0.10% based on PsmNet, and the accuracy on the foreground vehicle is improved by 0.30%; similarly, based on GwcNet, the overall improvement is 0.14%, and the accuracy on the foreground vehicle is improved by 0.19%. Obviously, the semantic association module is applicable to various networks, which can improve the overall performance of image matching, especially for the accurate matching of foreground objects.

The proposed multi-scale cost volume computation module, whether it acts on PsmNet or GwcNet, has an improvement in the overall or foreground vehicle pixels. Based on PsmNet, the overall improvement is 0.20%, and that on the foreground vehicle is improved by 0.50%; similarity, the overall improvement based on GwcNet is 0.20%, and that on the foreground vehicle is 0.23%. Although PsmNet uses SPP to obtain multi-scale features, the features are then upsampled to restore to the original size via bilinear interpolation, which leads to the loss of detailed features. In addition, the feature maps of different scales are concatenated as the final SPP feature maps; therefore, SPP uses the final SPP feature maps to construct one cost volume. This process simply aggregates the features of different scales and does not realize that the different scale features have different-level contributions to disparity estimation.

Finally, this paper also conducts a comparative experiment on the superposition of semantic association and multi-scale cost volume computation. The experimental results show that the overall improvement is 0.24% based on PsmNet, 0.64% on the foreground object vehicle, 0.22% on the basis of GwcNet, and 0.26% on the foreground object vehicle.

Thus, the two modules are universal and can be adapted to different networks, such as PsmNet and GwcNet. More importantly, the two modules proposed in this paper are effective for binocular disparity estimation, especially on foreground vehicles, which are very beneficial for subsequent 3D object detection.

Furthermore, in order to fully illustrate the role of the proposed semantic association module, we directly replace $\boldsymbol{f}_w$ and $\boldsymbol{f}_c$ in the semantic association module with $\boldsymbol{f}_p$. The results are shown in Table 4. The low-level image feature $\boldsymbol{f}_p$ is relatively cluttered, and if $\boldsymbol{f}_c$ is replaced with $\boldsymbol{f}_p$, the contextual association cannot be effectively utilized; if $\boldsymbol{f}_w$ is replaced with $\boldsymbol{f}_p$, although 2D self-attention can improve the contextual association, the lack of high-level semantics also affects feature representation. Comparatively speaking, combining semantic feature $\boldsymbol{f}_{\text{seg}}$ with low-level feature $\boldsymbol{f}_p$, and introducing 2D self-attention mechanisms to obtain semantic association feature can achieve the best results. The experimental results in Table 4 also prove this point.

4. Comparison with other methods

The comparison methods include PsmNet [17], GwcNet [19], NLCA-Net [21], SGNet [22], SegStereo [31],

Table 4 Ablation study for semantic association

Method	All (%)↓			Noc (%)↓		
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all
Ours* (replace $\boldsymbol{f}_c$ with $\boldsymbol{f}_p$)	1.78	4.48	2.23	1.64	4.04	2.04
Ours* (replace $\boldsymbol{f}_w$ with $\boldsymbol{f}_p$)	1.75	4.23	2.17	1.63	3.66	1.97
Ours	**1.55**	**3.55**	**1.88**	**1.42**	**3.28**	**1.73**

SSPCV-Net [33], and PGNet [34]. They are classic binocular stereo matching algorithms. PsmNet [17] is a basic network architecture similar to ours, and many methods adopt this architecture. GwcNet [19] is also a version of PsmNet with a similar architecture. NLCA-Net [21] adopts the self-attention mechanism. SegStereo [31], SSPCV-Net [33], SGNet [22], and PGNet [34] embed semantic constraints in the stereo matching process. In some respects, these self-attention mechanisms and semantic constraints are similar to those of our method. Therefore, the comparison with these methods is very convincing.

Table 5 shows the performance of the proposed method on the official KITTI2015 evaluation set, which is compared with other SOTA methods, including accuracy and inference time.

Table 5 Comparison results on the official KITTI2015 evaluation set

Method	All (%)↓			Noc (%)↓			Runtime[1]	Environment
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all		
PsmNet [17] (CVPR 2018)	1.86	4.62	2.32	1.71	4.31	2.14	0.41 s	Nvidia GTX Titan Xp
GwcNet [19] (CVPR 2019)	1.74	3.93	2.11	1.61	3.49	1.92	0.32 s	GPU @ 2.0 GHz (Python + C/C++)
SegStereo [31] (ECCV 2018)	1.88	4.07	2.25	1.76	3.70	2.08	0.6 s	Nvidia GTX Titan Xp
SSPCV-Net [33] (ICCV 2019)	1.75	3.89	2.11	1.61	3.40	1.91	0.9 s	1 core @ 2.5 GHz (Python)
NLCA-Net [21] (APSIPA 2020)	**1.53**	4.09	1.96	1.39	3.80	1.79	0.6 s	1 core @ 2.5 GHz (C/C++)
SGNet [22] (ACCV 2020)	1.63	3.76	1.99	1.46	3.40	1.78	0.6 s	1 core @ 2.5 GHz (Python + C/C++)
PGNet [34] (Neurocomputing 2021)	1.64	3.60	1.96	1.43	**3.21**	**1.72**	0.7 s	1 core @ 2.5 GHz (Python)
Ours	1.55	**3.55**	**1.88**	**1.42**	3.28	1.73	**0.23 s**	NVIDIA RTX 3090 (PyTorch)

Note: [1]Runtime: the inference time for a test image on a single card with batch size = 1.

It can be seen that the proposed method has significant improvement in foreground and overall indicators, effectively improving the accuracy of binocular stereo matching and disparity estimation. The semantic association module and the multi-scale cost volume computation module fully combine the semantics and the multi-scale features to obtain a more powerful feature representation, which is conducive to disparity estimation. Especially for the foreground objects, such as far-distance vehicles, the assistance of the ground semantic association and the cost volume computation at different scales make these ill-posed regions easier to estimate their disparities.

However, we also note that our method is slightly inferior to NLCA-Net on the D1-bg index. The reason is that our stereo matching mainly focuses on the semantic association and visual attention regions, which may be inclined to foreground; NLCA-Net mainly focuses on the global disparity estimation, so it performs slightly better on the D1-bg index. In addition, our method is also slightly inferior to PGNet on D1-fg and D1-all Noc index. PGNet is a panoptic parsing guided deep network, and three novel modules are designed to embed the panoptic guidance. On the one hand, it uses the semantic categories, instance layout, extra boundary, and smooth constraints from semantic and instance ground truth, thus, more information is conducive to the improvement of accuracy; on the other hand, PGNet uses Scene Flow to pre-train for 15 epochs and uses KITTI2015+KITTI2012 to finetune for 500 epochs. After that, PGNet fine-tunes on KITTI2015 dataset again for another 200 epochs when submitting to the benchmarks. PGNet needs more training epochs. Actually, in these comparison methods, SGNet adopted 700 epochs for fine-tuning, and NLCA-Net even adopted 900 epochs for fine-tuning. Comparatively, our method and PsmNet are the methods that use the least epochs in the training phase.

When we adopt "Scene Flow pre-train+KITTI2015 fine-tune", and the parameters are consistent with the parameters of PsmNet, the training process takes about 13.77 hours on Scene Flow dataset (using 3 RTX-3090GPUs with batch size equal to 12 for 10 epochs) and 2.52 hours on KITTI2015 (using 3 RTX-3090 GPUs with batch size equal to 12 for 300 epochs). It takes 0.23 s to infer one KITTI2015 testing image. When we adopt "KITTI3D pre-train+KITTI2015 fine-tune", the training process takes about 11.4 hours on KITTI3D (using 4 RTX-3090 GPUs with batch size equal to 12 for 50 epochs) and 3.7 hours on KITTI2015 (using 4 RTX-3090 GPUs with batchsize equal to 12 for 300 epochs). It also takes 0.23 s to infer one KITTI2015 testing image. Compared with other methods, the inference time of our method is also the least.

Figure 8 shows some disparity estimation results obtained by using PsmNet and our method. It can be observed that for the foreground vehicles, PsmNet can estimate the disparities of the near-distance vehicle objects, but the disparity estimation results of the far-distance vehicle objects are not good; our method can estimate the disparities of far-distant vehicle objects better, simultaneously, the edge estimation results of the near-distance vehicles are more accurate. For background pixels, such as the trees, our results are also more accurate than PsmNet's results. Actually, PsmNet cannot distinguish the trees from the sky, but our method can produce a more obvious boundary between the trees and the sky.

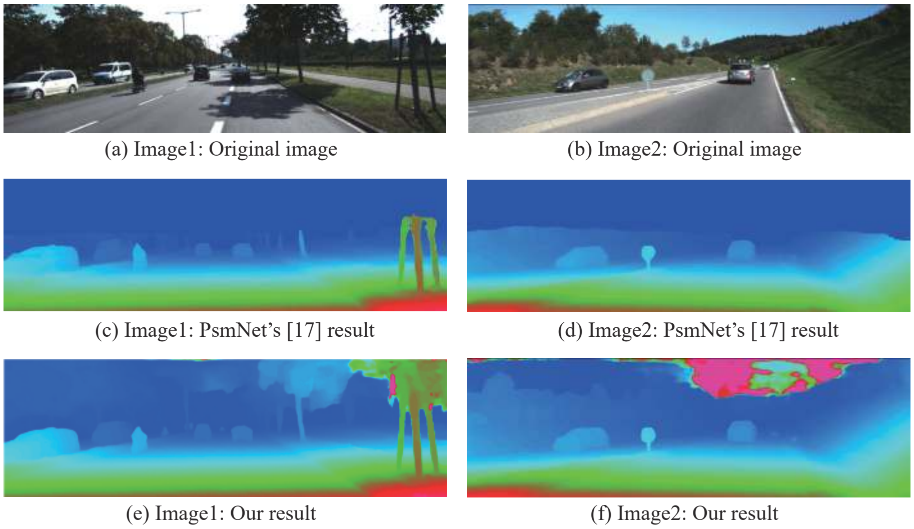

(a) Image1: Original image
(b) Image2: Original image
(c) Image1: PsmNet's [17] result
(d) Image2: PsmNet's [17] result
(e) Image1: Our result
(f) Image2: Our result

Figure 8 Visual comparison of disparity estimation results.

Similar to the results of PsmNet, the results of our method are also not good in some weak texture regions, such as the junction area of sky and ground. So, improving the performance on weak texture regions is our future work.

V. Conclusion

In this paper, the original binocular stereo matching network PsmNet is improved, and a semantic association module and a multi-scale cost volume computation module are proposed. The semantic association module introduces semantic category and attention mechanism to construct the contextual semantic association relationship among the pixels; the multi-scale cost volume computation module calculates the multiple cost volumes of features at different scales to fuse high-level semantic information and low-level local detailed information, which enhances the expressiveness of features and improves the accuracy of disparity estimation. Experimental results on KITTI2015 demonstrate our method achieves comparable or higher matching performance.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61876014).

References

[1] Z. L. Shen, X. B. Song, Y. C. Dai, *et al.*, "Digging into uncertainty-based pseudo-label for robust stereo matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 14301–14320, 2023.

[2] J. P. Jing, J. K. Li, P. F. Xiong, *et al.*, "Uncertainty guided adaptive warping for robust and efficient stereo matching," in *Proceedings of 2023 IEEE/CVF International Conference on Computer Vision*, Paris, France, pp. 3295–3304, 2023.

[3] G. W. Xu, X. Q. Wang, X. H. Ding, *et al.*, "Iterative geometry encoding volume for stereo matching," in *Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, pp. 21919–21928, 2023.

[4] C. Y. Chen, A. Seff, A. Kornhauser, *et al.*, "DeepDriving: Learning affordance for direct perception in autonomous driving," in *Proceedings of 2015 IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 2722–2730, 2015.

[5] Y. Wang, B. Yang, R. Hu, *et al.*, "PLUMENet: Efficient 3D object detection from stereo images," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Prague, Czech Republic, pp. 3383–3390, 2021.

[6] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328–341, 2008.

[7] 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)*, Hong Kong, China, pp. 15–18, 2006.

[8] W. J. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 5695–5703, 2016.

[9] J. H. Pang, W. X. Sun, J. S. Ren, *et al.*, "Cascade residual learning: A two-stage convolutional neural network for stereo matching," in *Proceedings of 2017 IEEE International Conference on Computer Vision Workshops*, Venice, Italy, pp. 878–886, 2017.

[10] M. Poggi, F. Tosi, K. Batsos, *et al.*, "On the synergies between machine learning and binocular stereo for depth estimation from images: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5314–5334, 2021.

[11] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions using graph cuts," in *Proceedings of the Eighth IEEE International Conference on Computer Vi-*

sion. ICCV 2001, Vancouver, BC, Canada, pp. 508–515, 2001.

[12] J. Sun, N. N. Zheng, and H. Y. Shum, "Stereo matching using belief propagation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 787–800, 2003.

[13] A. Hosni, C. Rhemann, M. Bleyer, *et al.*, "Fast cost-volume filtering for visual correspondence and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 504–511, 2013.

[14] K. J. Yoon and I. S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 650–656, 2006.

[15] J. Sun, Y. Li, S. B. Kang, *et al.*, "Symmetric stereo matching for occlusion handling," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, pp. 399–406, 2005.

[16] Q. Yang, L. Wang, R. Yang, *et al.*, "Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 492–504, 2009.

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

[18] F. H. Zhang, V. Prisacariu, R. G. Yang, *et al.*, "GA-Net: Guided aggregation net for end-to-end stereo matching," in *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 185–194, 2019.

[19] X. Y. Guo, K. Yang, W. K. Yang, *et al.*, "Group-wise correlation stereo network," in *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 3268–3277, 2019.

[20] Y. M. Zhang, Y. M. Chen, X. Bai, *et al.*, "Adaptive unimodal cost volume filtering for deep stereo matching," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, USA, pp. 12926–12934, 2020.

[21] Z. B. Rao, M. Y. He, Y. C. Dai, *et al.*, "NLCA-Net: A non-local context attention network for stereo matching," *APSIPA Transactions on Signal and Information Processing*, vol. 9, article no. e18, 2020.

[22] S. Y. Chen, Z. Y. Xiang, C. Y. Qiao, *et al.*, "SGNet: Semantics guided deep stereo matching," in *Proceedings of the 15th Asian Conference on Computer Vision*, Kyoto, Japan, pp. 106–122, 2020.

[23] H. Laga, L. V. Jospin, F. Boussaid, *et al.*, "A survey on deep learning techniques for stereo-based depth estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 1738–1764, 2020.

[24] X. D. Gu, Z. W. Fan, S. Y. Zhu, *et al.*, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 2492–2501, 2020.

[25] F. J. H. Wang, S. Galliani, C. Vogel, *et al.*, "PatchmatchNet: Learned multi-view patchmatch stereo," in *Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 14189–14198, 2021.

[26] H. Liu, R. Wang, Y. P. Xia, *et al.*, "Improved cost computation and adaptive shape guided filter for local stereo matching of low texture stereo images," *Applied Sciences*, vol. 10, no. 5, article no. 1869, 2020.

[27] B. L. Lu, Y. He, and H. N. Wang, "Stereo disparity optimization with depth change constraint based on a continuous video," *Displays*, vol. 69, article no. 102073, 2021.

[28] S. Gidaris and N. Komodakis, "Detect, replace, refine: Deep structured prediction for pixel wise labeling," in *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 7187–7196, 2017.

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

[30] F. Güney and A. Geiger, "Displets: Resolving stereo ambiguities using object knowledge," in *Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, pp. 4165–4175, 2015.

[31] G. R. Yang, H. S. Zhao, J. P. Shi, *et al.*, "SegStereo: Exploiting semantic information for disparity estimation," in *Proceedings of the 15th European Conference on Computer Vision*, Munich, Germany, pp. 660–676, 2018.

[32] J. M. Zhang, K. A. Skinner, R. Vasudevan, *et al.*, "DispSegNet: Leveraging semantics for end-to-end learning of disparity estimation from stereo imagery," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1162–1169, 2019.

[33] Z. Y. Wu, X. Y. Wu, X. P. Zhang, *et al.*, "Semantic stereo matching with pyramid cost volumes," in *Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*, Seoul, Korea (South), pp. 7483–7492, 2019.

[34] S. Y. Chen, Z. Y. Xiang, C. Y. Qiao, *et al.*, "PGNet: Panoptic parsing guided deep stereo matching," *Neurocomputing*, vol. 463, pp. 609–622, 2021.

[35] W. Y. Liu, Y. D. Wen, Z. D. Yu, *et al.*, "Large-margin softmax loss for convolutional neural networks," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, New York, NY, USA, pp. 507–516, 2016.

[36] H. W. Sang, Q. H. Wang, and Y. Zhao, "Multi-scale context attention network for stereo matching," *IEEE Access*, vol. 7, pp. 15152–15161, 2019.

[37] G. H. Zhang, D. C. Zhu, W. J. Shi, *et al.*, "Multi-dimensional residual dense attention network for stereo matching," *IEEE Access*, vol. 7, pp. 51681–51690, 2019.

[38] G. Y. Huang, Y. Y. Gong, Q. Z. Xu, *et al.*, "A convolutional attention residual network for stereo matching," *IEEE Access*, vol. 8, pp. 50828–50842, 2020.

[39] S. Woo, J. Park, J. Y. Lee, *et al.*, "CBAM: Convolutional block attention module," in *Proceedings of the 15th European Conference on Computer Vision*, Munich, Germany, pp. 3–19, 2018.

[40] X. L. Wang, R. Girshick, A. Gupta, *et al.*, "Non-local neural networks," in *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 7794–7803, 2018.

[41] Y. Yao, Z. Z. Luo, S. W. Li, *et al.*, "Recurrent MVSNet for high-resolution multi-view stereo depth inference," in *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 5520–5529, 2019.

[42] Q. S. Xu, W. H. Kong, W. B. Tao, *et al.*, "Multi-scale geometric consistency guided and planar prior assisted multi-view stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4945–4963, 2023.

[43] S. Cheng, Z. X. Xu, S. L. Zhu, *et al.*, "Deep stereo using adaptive thin volume representation with uncertainty awareness," in *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 2521–2531, 2020.

[44] G. Zhang, Z. Y. Li, J. M. Li, *et al.*, "CFNet: Cascade fusion network for dense prediction," *arXiv preprint*, arXiv: 2302.06052, 2023.

[45] Z. L. Shen, Y. C. Dai, and Z. B. Rao, "CFNet: Cascade and fused cost volume for robust stereo matching," in *Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp.

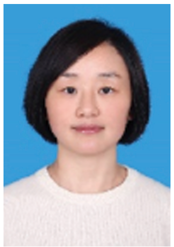
13901–13910, 2021.

[46] G. W. Xu, J. D. Cheng, P. Guo, *et al.*, “Attention concatenation volume for accurate and efficient stereo matching,” in *Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, pp. 12971–12980, 2022.

[47] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? The KITTI vision benchmark suite,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, pp. 3354–3361, 2012.

[48] J. Ku, A. Harakeh, and S. L. Waslander, “In defense of classical image processing: Fast depth completion on the CPU,” in *2018 15th Conference on Computer and Robot Vision (CRV)*, Toronto, ON, Canada, pp. 16–22, 2018.

[49] K. M. He, X. Y. Zhang, S. Q. Ren, *et al.*, “Deep residual learning for image recognition,” in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 770–778, 2016.



Jin ZHENG received the B.S. and M.S. degrees from Liaoning Technical University, Fuxin, China, in 2001 and 2004, resepectively, and the Ph.D. degree from the School of Computer Science and Engineering, Beihang University, Beijing, China, in 2009. She joined the School of Computer Science and Engineering, Beihang University, Beijing, China, in 2009. In 2014, she visited Harvard University, Cambridge, MA, USA, as a Visiting Scholar for one year. Her current research interests focus on object detection, tracking and recognition, among other similar interests.
(Email: JinZheng@buaa.edu.cn)

Botao JIANG received the B.S. degree from China University of Geosciences (Wuhan), Wuhan, China, in 2022. He is currently a postgraduate student majoring in computer technology with the School of Computer Science and Engineering, Beihang University, Beijing, China. His research interests include stereo matching, reinforcement learning, and 3D object detection.
(Email: Bert020@buaa.edu.cn)

Wei PENG received the B.S. degree in communication engineering from the Institute of Information Engineering, Hunan University, Changsha, China, in 2019. She received the M.S. degree from the School of Computer Science and Engineering, Beihang University, Beijing, China, in 2022. Her research interests include 3D object detection, tracking, and data association.
(Email: 3149169388@qq.com)

Qiaohui ZHANG received the B.S. degree from the Sino-French Engineer School, Beihang University, Beijing, China, in 2022. She is currently a postgraduate student majoring in computer technology with the School of Computer Science and Engineering, Beihang University, Beijing, China. Her research interests include depth estimation and 3D object detection.
(Email: qiaohui_zhang@buaa.edu.cn)