Dual Guided-Aggregation Network for Stereo Image Matching

Anonymous CVPR 2021 submission

Paper ID 6413

Abstract

Stereo image dense matching, which plays a key role in 3D reconstruction, still remains a challenging task in computer vision and photogrammetry. In addition to block-based matching, recent studies based on artificial neural network have achieved great progress in stereo matching by using deep convolutional networks. This study proposes a novel network called dual guided-aggregation network (Dual-GANet), which utilizes both left-to-right and right-to-left image matchings in network design and training to reduce the possibility of pixel mismatch. A flipped training with a cost volume consistentization is introduced to realize the learning of invisible-to-visible pixel matching and left-right consistency matching. In addition, a suppressed multi-regression is proposed, which suppresses unrelated information before regression and selects multiple peaks from a disparity probability distribution. The proposed dual network with left-right consistent matching scheme can be applied to existing stereo matching models. To estimate the performance, GANet that is designed based on semi-global matching is selected as the backbone with extensions and modifications on guided aggregation, disparity regression, and loss function. Experimental results on the SceneFlow and KITTI2015 datasets demonstrate the superiority of the Dual-GANet compared to related models in terms of end-point-error and pixel error rate.

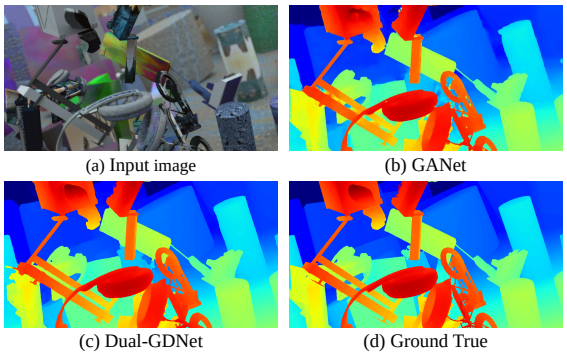
1. Introduction

The task of stereo matching is searching for the correspondences between pixels in stereo images and calculating their disparities. According to the workflow in [16], a typical stereo matching process consists of four main steps, namely, matching cost computation, cost aggregation, disparity optimization, and disparity refinement. Each step in this stereo pipeline has been extensively studied, and several advanced methods have been proposed [6,15,17]. Although this workflow performs well, the stepwise pipeline lacks an overall objective function for global optimization, and thus may suffer errors in each step [9].

Recently, deep learning techniques have shown remarkable performance in various tasks. In stereo matching, convolutional neural networks (CNNs) have been first utilized by Žbontar and Lecun [1] in matching cost computation. Instead of using similarity metrics, the authors proposed a Siamese CNN to measure the similarity between image patches. Several deep-learning-based methods have also addressed the problem of generating unary terms as similarity measurements by using CNN [3, 12]. These methods require post-processing to produce disparities. Therefore, Mayer et al. [13] proposed the integration of all steps into an end-to-end network and directly estimated the disparity from stereo images. Recently, Zhang et al. [20] proposed a guided aggregation network (GANet). Inspired by the cost aggregation in [6], a semi-global aggregation layer, which is a differentiable approximation of the semi-global matching, is introduced to capture the cost dependencies of the entire images.

Although the deep-learning models for stereo matching have excellent performance, current models still face challenges in dealing with object occlusions. Taking advantage of disparity information from both the left-to-right and right-to-left mutual disparity curves is an effective strategy to reduce the number of mismatched pixels. The traditional left-right consistency check is performed as a post-processing to remove ambiguous pixels. Instead of removing ambiguous pixels in post-processing, the main idea is to integrate left-to-right and right-to-left image matchings in the network design and training and to utilize the mutual disparity curves from dual cost volumes to reduce the possibility of pixel mismatch. In this scenario, the left-right consistency check is regarded as an optional step in post-processing for high-confident disparity extraction.

The GANet is selected as the backbone, and the proposed dual scheme can be applied to other advanced stereo matching models. To integrate the left-to-right and right-to-left image matchings in a network, a cost volume consistentization with flipped training is proposed to unify the cost volumes for both matchings and to learn invisible-to-visible matchings in addition to the general visible-to-invisible matching. In addition, a suppressed multi-regression is at-



(a) Input image (b) GANet (c) Dual-GDNet (d) Ground True

Figure 1. Performance comparison of Dual-GDNet and its backbone GA-Net.

tached with the Dual-GANet to remove unrelated information in regression and to provide support for multiple disparity candidates. Compared with related methods, the main contribution of the current study is to propose a dual guided-aggregation network for stereo image matching. Through a consistentization process on initial cost volumes, we establish a Siamese CNN attached with suppressed multi-regression, which can reduce mismatching possibility and support multiple candidate selection. As illustrated in Fig. 1, the result from the backbone GANet is improved by using the proposed dual network with flipped training. In terms of pixel error rate, which is an indicator for the measurement of mismatch possibility, an improvement of 2%–8% is obtained. Note that the backbone is a downsized version of the original GANet for the consideration of the GPU memory requirement.

2. Related Work

The related deep-learning-based stereo matching methods can be classified into three categories: *matching cost learning, end-to-end disparity learning*, and *left-right consistent learning*.

2.1. Matching cost learning

In this category, ANNs are utilized to compute the matching cost of image patches. In [1, 19], a Siamese network containing stacked convolutional layers is used to learn a similarity measurement of image patches. Similarly, Luo et al. [12] proposed a Siamese network to learn a probability distribution over all disparity values, after which they used a dot-product layer to join the branches of the network. Chen and Yuan [3] proposed a multi-scale CNN, in which the global context in down-sampled image patches is utilized to improve feature description. The matching cost learning still follows the typical workflow introduced in [16]. As mentioned, the stepwise workflow lacks an overall objective function for global optimization.

2.2. End-to-end disparity learning

The idea is integrating the four steps in the typical stereo matching workflow into a neural network and then training the network in an end-to-end manner. In Mayer et al. [13], a large synthetic dataset called Flying Chairs was created to train end-to-end deep networks for disparity estimation. Dosovitskiy et al. [5] proposed an encoder-decoder network called FlowNet to estimate disparity. Based on FlowNet, several studies have improved its performance by stacking multiple networks [7]. In Kendall et al. [10], a new network called GCNet is proposed, which utilizes 3D convolutional filters to regularize cost volumes and adopts a regression loss instead of a classification loss. Similar to GCNet, the PSMNet proposed by Shaked and Wolf [18] uses spatial pyramid pooling and 3D convolutions to incorporate contextual information in different scales. Recently, Zhang et al. (2019) [20] proposed GANet, which is inspired by the cost aggregation in [6]. A semi-global aggregation layer is presented to capture the cost dependencies of the images. The CSPN proposed by Cheng et al. [4] is a linear propagation model. The propagation is performed using recurrent convolutional operations, in which the affinity among neighboring pixels is learned through a deep CNN.

2.3. Left-right consistent learning

Inspired by the left-right consistency check, Jie et al. [8] proposed a left-right comparative recurrent model based on a convolutional long-short term memory (LSTM) network. The model contains two parallel stacked convolutional LSTM networks for left-to-right and right-to-left matchings, respectively. At each recurrent step, the model processes both views in parallel and produces error maps by performing left-right disparity comparison. Pixel mismatch generally happens in occlusion regions, because the matchings from invisible to visible pixels are missing during model training. Therefore, in the current study, a Siamese network that is unlike the LSTM-based structure in Jie et al. [8] is adopted to learn not only the matching between two visible pixels but the matching from invisible pixels to visible pixels.

3. Methodology

Fig. 2 illustrates the architecture of the Dual-GANet, which is a Siamese network containing the components of cost volume consistentization, flipped training, and suppressed regression. In the proposed network, the GANet is selected as the backbone with extensions and modifications on the semi-global guided aggregation layer, disparity regression, and loss function. A volume guided diffusion with six directions is utilized in the cost volume determination for the enlargement of receptive fields. To generate consistent cost volumes and to learn invisible-to-visible

matching, a cost volume consistentization process is performed, which consists of RS-Switch to exchange reference and support images and cost volume flipping to unify pixel positions in cost volumes. In suppressed regression, unrelated information is suppressed before the disparity regression, and multiple disparity candidates are allowed. In this section, the extensions of the GANet, cost volume consistentization, suppressed regression, and loss function are described in Sections 3.1–3.4, respectively.

3.1. GANet

The GANet proposed by Zhang et al. [20] contains initial cost volume construction, cost aggregation, and guidance network, all of which are inspired by SGM [6]. Unlike the traditional patch similarity measurement, the initial cost volume is constructed by iteratively concatenating features of the left and right images with incremental disparity shift. The features are generated by a shared-weight and stacked autoencoder network. The initial cost volume is a 4D tensor of the size $H \times W \times D_{max} \times 2F$, where H and W represent the height and width of the images, respectively; D_{max} denotes the max disparity; and F represents the number of features extracted by the autoencoder network. The initial cost volume is further refined by cost aggregation, which contains semi-global guided aggregation (SGA) layers, local guided aggregation (LGA) layers, and guidance subnets.

The SGA layer is defined based on the cost aggregation in SGM. In SGM, the aggregated cost $L_{\mathbf{r}}(\mathbf{p}, d)$ of a pixel $\mathbf{p}$ with the disparity d in the direction $\mathbf{r}$ is defined recursively as

$$L_{\mathbf{r}}(\mathbf{p}, d) = \mathbf{C}(\mathbf{p}, d) + \min \begin{cases} L_{\mathbf{r}}(\mathbf{p}-\mathbf{r}, d), \\ L_{\mathbf{r}}(\mathbf{p}-\mathbf{r}, d-1) + P_1, \\ L_{\mathbf{r}}(\mathbf{p}-\mathbf{r}, d+1) + P_1, \\ \min\limits_i L_{\mathbf{r}}(\mathbf{p}-\mathbf{r}, i) + P_2, \end{cases} \quad (1)$$

where $C(\mathbf{p}, d)$ presents the cost of the pixel $\mathbf{p}$ with the disparity d in the initial cost volume, and P_1 and P_2 are the cost penalties to prevent discontinuities and enforce smoothness in the disparity field. The discontinuity preservation is performed by adding P_1 or P_2 to the costs to penalize large changes in neighboring disparities. In GANet, the penalty parameters are defined as learnable weights in the network. Thus, the penalties are adaptive and changeable based on neighboring context information. Following the definition in SGM, the aggregated cost in SGA layer is formulated as

$$L_{\mathbf{r}}(\mathbf{p}, d) = \text{sum} \begin{cases} w_0(\mathbf{p}, \mathbf{r}) \times \mathbf{C}(\mathbf{p}, d), \\ w_1(\mathbf{p}, \mathbf{r}) \times L_{\mathbf{r}}(\mathbf{p}-\mathbf{r}, d), \\ w_2(\mathbf{p}, \mathbf{r}) \times L_{\mathbf{r}}(\mathbf{p}-\mathbf{r}, d-1), \\ w_3(\mathbf{p}, \mathbf{r}) \times L_{\mathbf{r}}(\mathbf{p}-\mathbf{r}, d+1), \\ w_4(\mathbf{p}, \mathbf{r}) \times \max\limits_i L_{\mathbf{r}}(\mathbf{p}-\mathbf{r}, i). \end{cases} \quad (2)$$
$$s.t. \sum_{i=0\ 4} w_i(\mathbf{p}, \mathbf{r}) = 1,$$

where $\{w_0(\mathbf{p}, \mathbf{r}), ..., w_4(\mathbf{p}, \mathbf{r})\}$ are the penalty weights in the network. The cost volume $C(\mathbf{p}, d)$ in Eq. (2) can be $C^R(\mathbf{p}, d)$ in the left-to-right matching or $C^{S_f}(\mathbf{p}, d)$ in the right-to-left matching. The external and internal minimal cost selections in Eq. (1) are replaced by weighted sums, which are implemented by using convolutions with strides, thus leading to an all convolutional network. Similar to SGM, the values of aggregated costs $L_{\mathbf{r}}$ increase along the path, which may lead to extremely large values. Thus, the weights $\{w_0(\mathbf{p}, \mathbf{r}), ..., w_4(\mathbf{p}, \mathbf{r})\}$ are normalized to avoid such a problem. In the implementation, the initial cost volume is sliced into D_{max} slices for each disparity, and each slice performs the cost aggregation in Eq. (2) with shared weight matrices.

The utilization of downsampling and upsampling in a stacked autoencoder may blur thin structures in images. To alleviate this problem, the LGA layer containing several guided filters is used to refine the matching costs and to recover thin structures. The local aggregation is formulated as

$$L(\mathbf{p}, d) = \text{sum} \begin{cases} \sum_{\mathbf{q} \in N_{\mathbf{p}}} w_0(\mathbf{p}, \mathbf{q}) \times \mathbf{C}(\mathbf{p}, d), \\ \sum_{\mathbf{q} \in N_{\mathbf{p}}} w_1(\mathbf{p}, \mathbf{q}) \times \mathbf{C}(\mathbf{p}, d-1), \\ \sum_{\mathbf{q} \in N_{\mathbf{p}}} w_2(\mathbf{p}, \mathbf{q}) \times \mathbf{C}(\mathbf{p}, d+1), \end{cases} \quad (3)$$
$$s.t. \sum_{\mathbf{q} \in N_{\mathbf{p}}} w_0(\mathbf{p}, \mathbf{q}) + w_1(\mathbf{p}, \mathbf{q}) + w_2(\mathbf{p}, \mathbf{q}) = 1,$$

where $\mathbf{q}$ is a pixel in the $K \times K$ neighbor region of the pixel $\mathbf{p}$. The LGA layer has three $K \times K$ filters with corresponding weights $w_0(\mathbf{p}, \mathbf{q})$, $w_1(\mathbf{p}, \mathbf{q})$, $w_2(\mathbf{p}, \mathbf{q})$ at each pixel $\mathbf{p}$ for the disparities d, $d-1$, and $d+1$, respectively.

Inspired by the concepts of data diffusion and receptive field for context learning in [4, 11], two modifications on the SGA layer are made in the current study. First, to increase the receptive field and to incorporate more context information in matching cost estimation, the 1D convolution for the slice aggregation in SGA layer is extended to 2D convolution, as illustrated in Fig. 3. Then, the aggregation is performed on 3D cost volume instead of a 2D slide of the cost volume. As a result, the number of aggregation directions is increased from four to six, including left, right, forward, backward, up, and down, thus leading to receptive field enlargement. Second, the internal maximal selection in Eq. (2), which requires the traversal in entry disparity axis, is removed in order to consider the computational cost. Considering these two modifications, Eq. (2) is reformulated as

$$L_{\mathbf{r}}(\mathbf{v}) = \text{sum} \begin{cases} w_0(\mathbf{p}, \mathbf{r}) \times \mathbf{C}(\mathbf{v}), \\ \sum_{\mathbf{q} \in N_{\mathbf{p}}} w_{\mathbf{q}}(\mathbf{p}, \mathbf{r}) \times L_{\mathbf{r}}(\mathbf{v}-\mathbf{r}+\mathbf{q}), \end{cases} \quad (4)$$
$$s.t. \sum_{i=0}^{k^2} w_i(\mathbf{p}, \mathbf{r}) = 1,$$

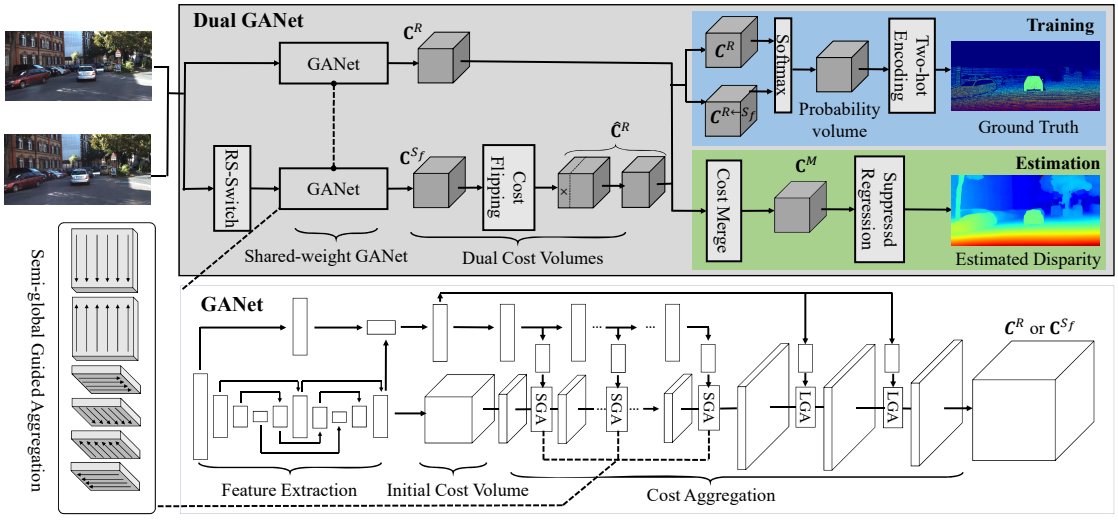


Figure 2. Overview of the Dual-GANet architecture. The Dual-GANet is a Siamese network containing the components of Siamese GANet, cost volume consistentization, flipped training (marked by blue), and suppressed regression (marked by green). The consistentization process consists of RS-Switch and cost flipping in the second GANet.

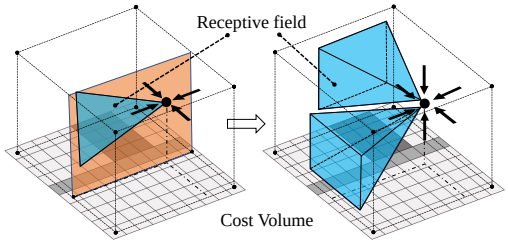


Figure 3. Illustration of the receptive field expansion in the SGA layer. Left: 4-direction aggregation with 2D receptive field in the GANet. Right: 6-direction aggregation with 3D receptive filed in Dual-GANet. The receptive fields are visualized with blue color.

where $L_{\mathbf{r}}(\mathbf{v})$ and $w_{\mathbf{q}}(\mathbf{p}, \mathbf{r})$ represent the aggregated cost of the voxel $\mathbf{v}$ and the weights for the neighbor pixel $\mathbf{q}$ in the direction r, respectively. Similarly, the weights are normalized to prevent numerical overflow.

3.2. Cost volume consistentization and flipped training

The purpose of stereo matching is to find conjugated points in a reference image based on its support image. The extracted conjugated pixel pairs can be represented as a matching set $\{(\mathbf{V}^R \rightarrow \mathbf{V}^S), (\mathbf{V}^R \rightarrow \mathbf{I}^S)\}$, where $\mathbf{V}^R$ denotes the visible pixel set in the reference image, and $\mathbf{V}^S$ and $\mathbf{I}^S$ represent the visible and invisible pixel sets, respectively, in the support image. The extraction of the matching pairs $(\mathbf{V}^R \rightarrow \mathbf{V}^S)$ using matching metrics is effective due to the visibility of the pixels and their local neighbors. However, mismatch may happen in the determination of the matching pairs $(\mathbf{V}^R \rightarrow \mathbf{I}^S)$ because of the occlusion and invisibility in the support image. The matching pairs $(\mathbf{V}^R \rightarrow \mathbf{V}^S)$ and $(\mathbf{V}^R \rightarrow \mathbf{I}^S)$ can be learned by a neural network with the aid of labeled training images, and the mismatches incurred by occlusion can be alleviated. However, the labeled training set did not contain the matching $(\mathbf{I}^R \rightarrow \mathbf{V}^S)$, that is, the matching of an invisible pixel in the reference and a visible pixel in the support image. In addition, the inference of the matching $(\mathbf{I}^R \rightarrow \mathbf{V}^S)$ from $(\mathbf{V}^R \rightarrow \mathbf{V}^S)$ and $(\mathbf{V}^R \rightarrow \mathbf{I}^S)$ is inefficient. To solve this problem, flipped training with Dual-GANet is proposed which contains left-to-right and right-to-left image matching. The left-to-right matching pairs provide the learning examples of $(\mathbf{V}^R \rightarrow \mathbf{V}^S)$ and $(\mathbf{V}^R \rightarrow \mathbf{I}^S)$, whereas the right-to-left image matching pairs support the learning of $(\mathbf{I}^R \rightarrow \mathbf{V}^S)$. In the implementation, the right-to-left image matching is realized by flipping the left-to-right matching. In other words, the matching examples $(\mathbf{I}^R \rightarrow \mathbf{V}^S)$ are created by flipping the matching $(\mathbf{V}^R \rightarrow \mathbf{I}^S)$ in the left-to-right matching to $(\mathbf{I}^S \rightarrow \mathbf{V}^R)$ in the right-to-left matching. Furthermore, a consistentization process is performed to ensure the consistency of the cost volumes generated from the left-to-right and right-to-left matching. Fig. 4 illustrates the processes consisting of RS-Switch and cost flipping.

The RS-Switch is performed before the generation of the initial cost volumes. The purpose of the RS-Switch is to exchange the reference and support images, such that the right and left images are selected as reference and support images, respectively, and the right-to-left matching can be performed. Then, the exchanged images are horizontally flipped to ensure that the search directions in the epipolar lines are consistent in the left-to-right and right-to-left matchings. However, after the processes of RS-Switch and feature extraction, the initial cost volumes from the left-to-

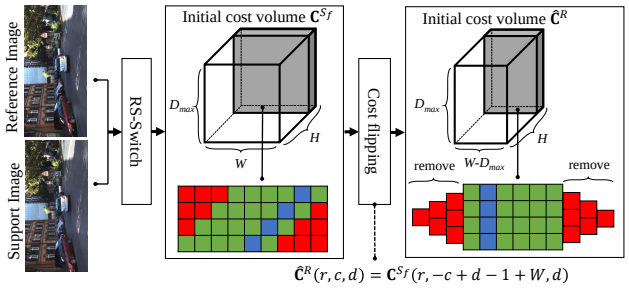


Figure 4. Illustration of initial cost volume consistentization, which consists of the steps: RS-Switch and cost flipping.

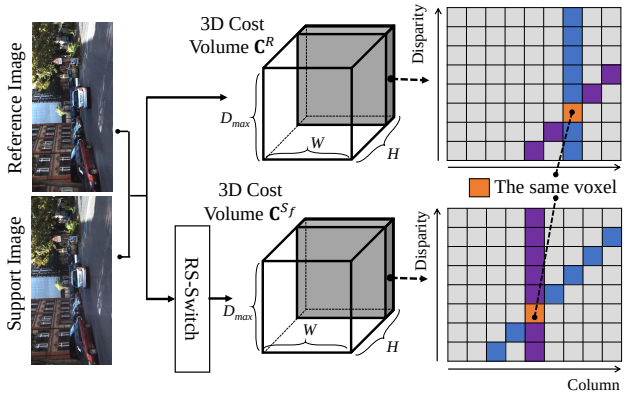


Figure 5. Illustration of cost volume inconsistency. A vertical cost pillar in $\mathbf{C}^R$ links to a tilted cost pillar in $\mathbf{C}^{S_f}$. The orange pixels in $\mathbf{C}^R$ and $\mathbf{C}^{S_f}$ are the same voxel.

right and right-to-left matchings (denoted as $\mathbf{C}^R$ and $\mathbf{C}^{S_f}$ are still inconsistent. For instance, in Fig. 5, a vertical cost pillar, which is defined as the matching costs for a pixel in different disparities, in $\mathbf{C}^R$ links to a tilted cost pillar in $\mathbf{C}^{S_f}$, and vice versa. In addition, the orange pixels in $\mathbf{C}^R$ and $\mathbf{C}^{S_f}$ are the same pixels, but they are at different locations of the cost volumes. Therefore, a cost volume transformation is performed, and the initial cost volume $\mathbf{C}^{S_f}$ is flipped according to the equation

$$\hat{\mathbf{C}}^R(r,c,d) = \mathbf{C}^{S_f}(r,-c+d-1+W,d) \tag{5}$$

where $\hat{\mathbf{C}}^R(r,c,d)$ represents the matching cost of pixel (r,c) in the reference image of the right-to-left matching. $\hat{\mathbf{C}}^R(r,c,d)$ is the dual cost volume of $\mathbf{C}^R$ which means that the pixel locations in these two cost volumes are the same, and the estimated disparities should be similar. After the cost volume flipping, the border region of width D_{max} (the red regions in Fig. 4) are removed because of the incomplete information in the disparity axis.

3.3. Suppressed regression

To take advantage of the left-to-right and right-to-left image matching, the cost volume $\mathbf{C}^R$ and its dual volume $\hat{\mathbf{C}}^R$ from Dual-GANet are merged and integrated. The merged cost volume $\mathbf{C}^M$ is defined as follows:

$$\mathbf{C}^M(r,c,d) = \begin{cases} \mathbf{C}^R(r,c,d), & \text{if } c \in [0, D_{max}-1] \\ \frac{\mathbf{C}^R(r,c,d)+\hat{\mathbf{C}}^R(r,c,d)}{2}, & \text{otherwise.} \end{cases} \tag{6}$$

As there D_{max} pixels in the dual cost volume $\hat{\mathbf{C}}^R(r,c,d)$ have incomplete information in the disparity axis, the cost merged for that pixels directly assigns $\mathbf{C}^R(r,c,d)$ to $\mathbf{C}^M(r,c,d)$.

The disparity regression in the GCNet and GANet is generally performed to estimate disparity. However, the disparity regression relies on the generation of a unimodal disparity probability distribution for a pixel in the cost volume, and the regression of disparities may also blur disparity edges. To alleviate these problems and to support multiple disparity candidates, an extension of the disparity regression called suppressed regression is proposed. As illustrated in Fig. 6, the suppressed regression consists of four main steps to approximate the pixel disparities in subpixel accuracy. By following the disparity regression in [10], the predicted costs in the merged cost volume $\mathbf{C}^M$ are converted to a probability volume by taking the negative of each value and normalizing the volume across the disparity dimension with the softmax operation $\sigma(\cdot)$. Then, the maximum of $\sigma(-\mathbf{C}^M(\mathbf{p}))$ is selected as the preliminary optimal disparity, denoted as $\hat{d}_a$, for the pixel $\mathbf{p}$. A neighboring region of $\hat{d}_a$ is determined based on the gradient of the disparity $\nabla\sigma$. The neighboring regions with $\nabla\sigma > 0$ and $\nabla\sigma < 0$ at the left-hand side and right-hand side of $\hat{d}_a$, respectively, are selected as the active regions. The areas outside of the active regions are selected as inactive regions, and the disparities in the latter regions are suppressed to zero. The probability distribution is normalized in the next step, such that the integral of the active regions is equal to one. Then, the disparity regression is applied to the normalized active region. The neighborhood of the estimated disparity with near normal distribution is used only in the disparity regression. The possibility of acquiring accurate disparities can be improved owing to the removal of unrelated information prior to the disparity regression.

If multiple selection is required, one more step is further performed. The active region is removed and then the process goes back to the first step for the selection of the next disparity candidate. The processes are terminated when all candidates are selected or when all the regions are removed.

3.4. Loss function

Most deep learning models for stereo image matching use mean squared error (MSE) as the loss function and select softmax as the activation function in the output layer to generate sharp and unimodal probability distribution for the disparity regression. However, using the MSE with

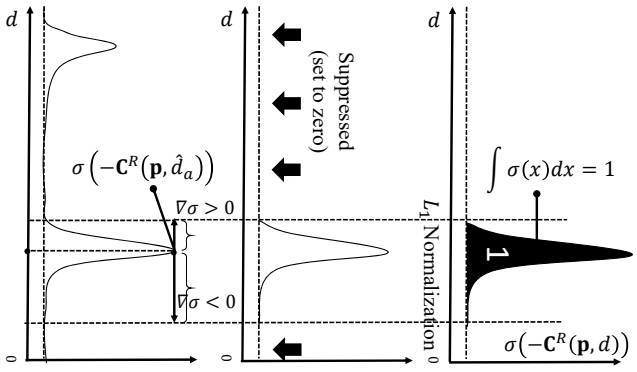


Figure 6. Illustration of the suppressed regression process. Left: defining the active regions; middle: suppressing the inactive regions; right: normalizing the active regions.

sharp and unimodal disparity probability distribution can lead to a slow convergence in model training. During back-propagation, most of the neurons in the output layer receive small gradients, because of the narrow and unimodal probability distribution. The small gradients will slow down the update of parameters. Following the design in classification, we alleviate the above-mentioned problem by using two-hot encoding with cross entropy as the loss function to replace MSE, and retaining the softmax activation function in the output layer. Due to the encoding and loss function, the small gradients can be avoided in the output layer and the parameters can be updated efficiently. In two-hot encoding, a disparity ground truth is encoded by two numerical values with weights. These two numerical values are the integers close to the disparity ground truth. The two-hot encoding is performed by using the equation

$$p(d) = \begin{cases} 1 - d^* + \lfloor d^* \rfloor, & \text{if } k = \lfloor d^* \rfloor \\ d - \lfloor d^* \rfloor, & \text{if } k = \lfloor d^* \rfloor + 1 \\ 0, & \text{otherwise} \end{cases}, \quad d \in [0, D_{max}],\ d^* \in [0, D_{max}) \tag{7}$$

where $y(d)$ is the encoded disparity probability distribution, and d^* denotes the disparity ground truth.

The cross entropy is used as the loss function to measure the similarity between the ground truth and predicted disparities. This is defined as

$$L(y, \hat{y}) = \sum_{d=0}^{D_{max}} -y(d) \ln \hat{y}(d) \tag{8}$$

where y denotes the probability distribution of the predicted disparity, which is the output of the softmax function $\sigma(\cdot)$, and $\hat{y}$ is the disparity ground truth encoded by using two-hot encoding.

4. Experimental Results

The Dual-GANet is evaluated using the SceneFlow [13] and KITTI2015 [14] datasets. The model is trained using Nvidia RTX 2080Ti. The metrics used in the model evaluation include average endpoint error (avg. EPE), average error rate (avg. ER), and average confidence level (avg. CL). The EPE is defined as the absolute distance between an estimated disparity and its corresponding ground truth. The error rate is defined with a threshold k, denoted as ER $> k$ px, to represent the percentage of pixels with EPEs larger than x pixels. The confidence level is defined as the distance between the cost volume $\mathbf{C}^R$ and its dual cost volume $\hat{\mathbf{C}}^R$, which is used to measure the consistency of the left-to-right and right-to-left image matching. In addition, the proposed model allows the selection of two disparity candidates, and the evaluation metric is applied to the best candidate.

4.1. Evaluation using the Scene Flow Dataset

The input tensor is set to 192 (maximal disparity) $\times$ 512 (height) $\times$ 960 (width). Considering the limited GPU memory, the proposed model uses a downsized GANet, denoted as GANet_small, as the backbone. In GANet_small, the down-sampling factor in the bottleneck layers of SGA is set to 3 instead of 4 in the GANet. Specifically, the tensors of the bottleneck layers in SGA are $(\frac{D_{max}}{4}, \frac{W}{4}, \frac{H}{4})$ and $(\frac{D_{max}}{8}, \frac{W}{8}, \frac{H}{8})$ instead of the original tensors $(\frac{D_{max}}{3}, \frac{W}{3}, \frac{H}{3})$ and $(\frac{D_{max}}{6}, \frac{W}{6}, \frac{H}{6})$, respectively. In addition, the number of layers in the feature extraction is reduced, and one-fourth of the original layers is removed.

The disparity estimation results using Dual-GANet are shown in Fig. 7. The evaluations of the 6-direction 2D aggregation, suppressed regression, and multi-candidate selection are provided in Table 1. In the comparison of Model #3 and Model #2, the results indicated that the avg. EPE and avg. ER in Model #3 are improved by 10.0% and 10.8%, respectively. This shows that the enlargement of the receptive field in the SGA layer can improve the recognition of objects and the estimation of disparities. In Model #4, the avg. ER is improved from 10.31% to 7.37% compared with the result in Model #3, thereby demonstrating that the suppressed regression can reduce the error rate in the disparity estimation. In addition, in Model #6, the avg. ER is further improved from 7.37% to 6.65% with the aid of Dual-GANet. This shows that the flipped training for the learning of invisible-to-visible pixel matching and left-right consistency matching scheme can reduce error rate and mismatching. Furthermore, with the aid of multi-candidate selection, the avg. EPE and avg. ER in Model #7 can decrease to 0.418 px and 5.81%, respectively, which are better than those obtained using GANet.

4.2. Evaluation on the KITTI2015 Dataset

In the evaluation of the KITTI2015 dataset, the input tensor is set to 192 $\times$ 256 $\times$ 1024, and the 6-direction aggregation in the SGA layers and suppressed regression are used in both the GANet_small and Dual-GANet. The compar-

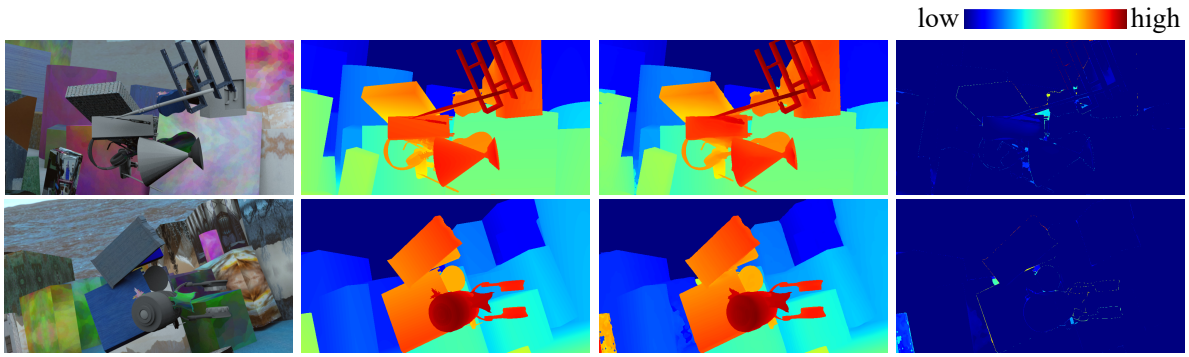


Figure 7. Disparity estimation using the SceneFlow dataset. $1^{st} - 4^{th}$ columns: input reference image, disparity ground truth, Dual-GANet results, disparity error map.

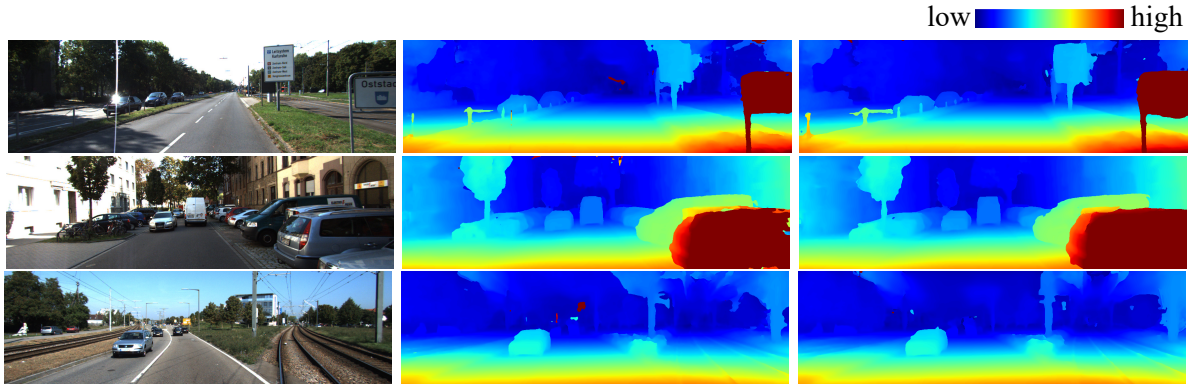


Figure 8. Disparity estimation using KITTI2015. Left: input reference image; middle: disparity results generated by GANet; right: disparity results generated by Dual-GANet.

isons between the disparities estimated by GANet and Dual-GANet are shown in Fig. 8. As shown in the visual comparisons, several mismatches occurred in the homogenous sky regions in GANet_small. Most of these mismatches can be fixed by using the Dual-GANet with the ability of left-right consistency matching.

To further evaluate the proposed model, the related and recent methods and models, including SGM [6], GCNet [10], PSMNet [2], GANet [20], and CSPN [4], are compared. The comparison results are shown in Table 2. After comparing the Dual-GANet (with dual scheme) with GANet_small (without dual scheme), results indicate that the avg. ER is improved from 2.32% to 1.76%. In addition, the Dual-GANet using a small version of the GANet as backbone is slightly better, in terms of avg. ER than the original GANet. Furthermore, the Dual-GANet with 2-candidate selection can reduce the avg. ER to 0.95%, which is much better than the results of the related models. Note that the dual network scheme can be applied to other advanced models. The disparity estimation can be further improved when the original GANet or CSPN is used as the backbone.

4.3. Evaluation of Flipped Training

To evaluate the dual network scheme and flipped training, a comparison of cost volumes generated by the GANet (without dual scheme) and Dual-GANet (with dual scheme) is conducted. The results are shown in Fig. 9. The two points in the stereo image selected for evaluation, which are denoted as $\mathbf{p}_1$ and $\mathbf{p}_2$, are located at a homogenous sky region and an occlusion region, respectively. Their disparity search ranges along the epipolar lines are displayed as blue dots in the images. These two kinds of regions often cause mismatches because of the poor context information. In Fig. 9, the disparity probability distributions in $\mathbf{C}^R$, $\hat{\mathbf{C}}^R$, and $\mathbf{C}^M$ are visualized by blue, orange, and green colors, respectively. For fair comparison, in the GANet, the cost volumes $\mathbf{C}^R$ from the left-to-right image matching and the cost volume $\hat{\mathbf{C}}^R$ from the right-to-left image matching are generated separately. Then, the cost volume $\mathbf{C}^M$ is obtained by merging $\mathbf{C}^R$ and $\hat{\mathbf{C}}^R$. In Dual-GANet, the cost volumes $\mathbf{C}^R$, $\hat{\mathbf{C}}^R$, and $\mathbf{C}^M$ are generated by using the workflow shown in Fig. 2. The pixel disparities are obtained by using the

Table 1. Evaluation and comparison using the SceneFlow dataset. The number of aggregation directions is denoted as # of A.D..

No.	Model	# of A.D.	Disparity regression	Suppressed regression	2-candidate	avg. EPE (px)	avg. ER > 1px
1	GANet	4	✓			0.780	8.7%
2	GANet_small (backbone)	4	✓			0.995	11.56%
3	GANet_small	6	✓			0.895	10.31%
4	GANet_small	6		✓		0.865	7.37%
5	GANet_small	6		✓	✓	0.440	6.56%
6	Dual-GANet	6		✓		0.862	6.65%
7	Dual-GANet	6		✓	✓	0.418	5.81%

Table 2. Comparisons of the related methods and the proposed models using the KITTI2015 dataset.

Model	avg. EPE	avg. ER>3px
SGM [6]	–	6.38%
GCNet [10]	–	6.16%
PSMNet [2]	–	2.32%
GANet [20]	–	1.81%
CSPN [4]	–	1.74%
GANet_small	0.790	2.32%
Dual-GANet (single candidate)	0.712	1.76%
Dual-GANet (two candidates)	0.589	0.95%

merged cost volume.

In the case of point $\mathbf{p}_1$, the disparity probability distributions in $\mathbf{C}^R$, $\hat{\mathbf{C}}^R$, and $\mathbf{C}^M$ of the GANet are inconsistent because of poor context information in the homogenous region, which results in the unstable estimation of disparity. The predicted disparity is 158.92, which is far from the ground truth (0.0). By contrast, with the aids of dual network and flipped training, the disparity probability distributions generated by the Dual-GANet are near consistent, and the estimated disparity (4.92) is close to the ground truth. In the case of point $\mathbf{p}_2$, the disparity probability distributions in $\mathbf{C}^R$, $\hat{\mathbf{C}}^R$, and $\mathbf{C}^M$ of the GANet and Dual-GANet are consistent. However, the confident level (CL) in the Dual-GANet is 2.31, which is better than that in the GANet (CL=9.8). This result demonstrates that the proposed dual network and flipped training are able to stabilize the disparity estimation.

5. Conclusions and Future Works

A Dual-GANet for stereo image dense matching is proposed in this work. By integrating the left-right consistent matching and utilizing the invisible-to-visible pixel learning in network design and training, the possibility of pixel mismatches can be reduced. In addition, the suppressed regression and multi-candidate selection can refine the disparity estimation and improve the probability of selecting the correct disparity. In the implementation, a downsized version of the GANet is used as the backbone with certain modifications on the loss function and aggregation. The uses of cross entropy with two-hot encoding and the six-direction aggregation with 3D receptive field in the LGA layers can reduce matching error rate. The experiment results on the SceneFlow and KITTI2015 datasets demonstrate the feasibility and practicality of the proposed method. To further improve the performance, the original GANet or other advanced stereo image matching models can be selected as the backbone. For our future work, we are interested in extending and applying the Dual-GANet to multi-image matching to further improve the accuracy of the estimated disparities.

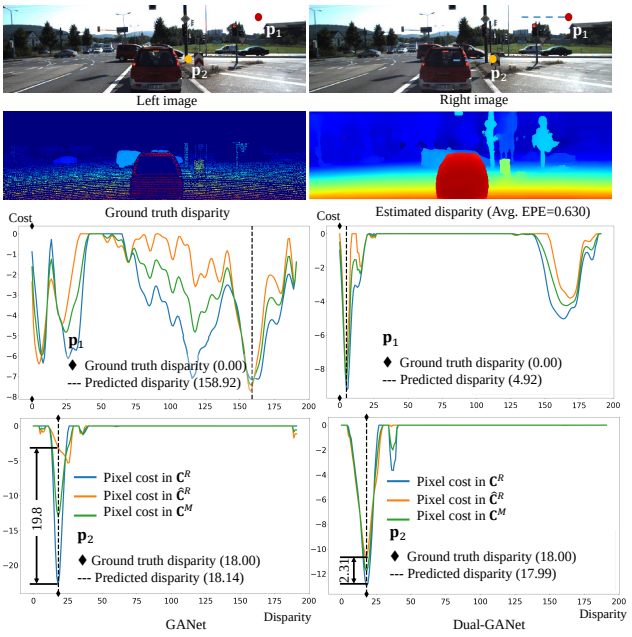


Figure 9. Comparison of the disparity probability distributions of pixels in the cost volumes of the GANet (left) and Dual-GANet (right).

References

[1] Jure Žbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *CoRR*, abs/1510.05970, 2015. 1, 2

[2] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. *CoRR*, abs/1803.08669, 2018. 7, 8

[3] Jiahui Chen and Chun Yuan. Convolutional neural network using multi-scale information for stereo matching cost computation. pages 3424–3428, 09 2016. 1, 2

[4] Xinjing Cheng, Peng Wang, and Ruigang Yang. Learning depth with convolutional spatial propagation network. *IEEE transactions on pattern analysis and machine intelligence*, 42:2361–2379, 2020. 2, 3, 7, 8

[5] Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip Häusser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. *CoRR*, abs/1504.06852, 2015. 2

[6] Heiko Hirschmüller. Stereo Processing by Semi-Global Matching and Mutual Information. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 30(2):328–341, 2008. 1, 2, 3, 7, 8

[7] Eddy Ilg, Tonmoy Saikia, Margret Keuper, and Thomas Brox. Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation. *CoRR*, abs/1808.01838, 2018. 2

[8] Zequn Jie, Pengfei Wang, Yonggen Ling, Bo Zhao, Yunchao Wei, Jiashi Feng, and Wei Liu. Left-right comparative recurrent model for stereo matching. *CoRR*, abs/1804.00796, 2018. 2

[9] Junhua Kang, Lin Chen, Fei Deng, and Christian Heipke. Context pyramidal network for stereo matching regularized by disparity gradients. *ISPRS Journal of Photogrammetry and Remote Sensing*, 157, 09 2019. 1

[10] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. *CoRR*, abs/1703.04309, 2017. 2, 5, 7, 8

[11] Sifei Liu, Shalini De Mello, Jinwei Gu, Guangyu Zhong, Ming-Hsuan Yang, and Jan Kautz. Learning affinity via spatial propagation networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1520–1530. Curran Associates, Inc., 2017. 3

[12] W. Luo, A. G. Schwing, and R. Urtasun. Efficient deep learning for stereo matching. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5695–5703, 2016. 1, 2

[13] Nikolaus Mayer, Eddy Ilg, Philip Häusser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. *CoRR*, abs/1512.02134, 2015. 1, 2, 6

[14] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3061–3070, 2015. 6

[15] Mikhail Mozerov and Joost Weijer. Accurate stereo matching by two-step energy minimization. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 24, 01 2015. 1

[16] D. Scharstein, R. Szeliski, and R. Zabih. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001)*, pages 131–140, 2001. 1, 2

[17] Mozhdeh Shahbazi, Gunho Sohn, and Jérôme Théau. High-density stereo image matching using intrinsic curves. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146:373–388, 12 2018. 1

[18] Amit Shaked and Lior Wolf. Improved stereo matching with constant highway networks and reflective confidence learning. *CoRR*, abs/1701.00165, 2017. 2

[19] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. *CoRR*, abs/1504.03641, 2015. 2

[20] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. GA-Net: Guided Aggregation Net for End-to-end Stereo Matching. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019. 1, 2, 3, 7, 8