GA-Net: Guided Aggregation Net for End-to-end Stereo Matching

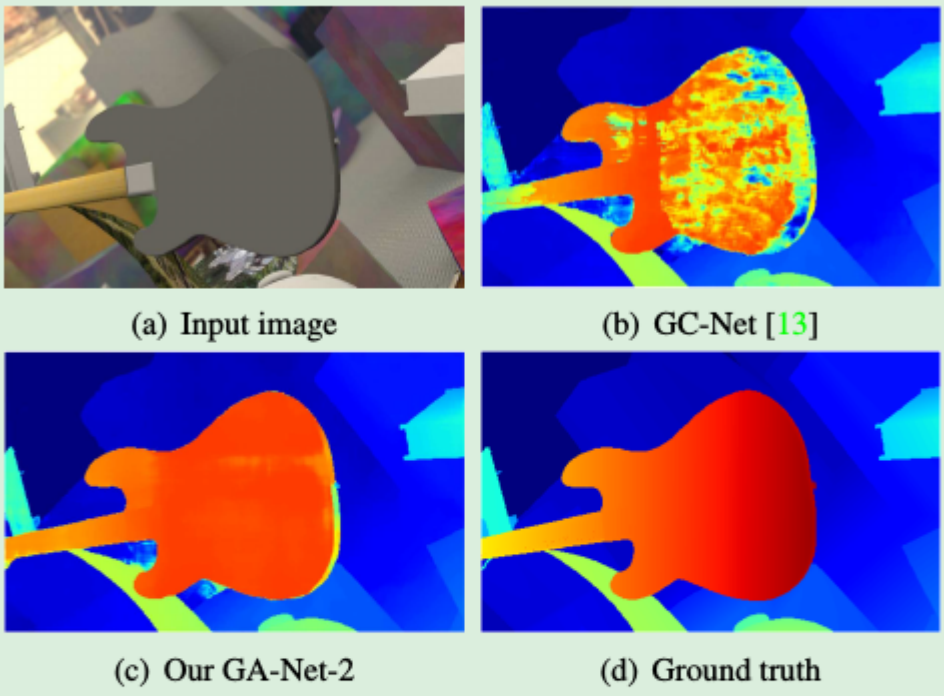
Abstract

> The propose two novel neural net layers, aimed at capturing local and the whole-image cost dependencies respectively.
> The first is a semi-global aggregation layer which is a differentiable approximation of the semi-global matching, the second is the local guided aggregation layer which follows a traditional cost filtering strategy to refine the thin structures.
> These two layers can be used to replace the widely used 3D concolutional layer which is computationally costly and memory-consuming as it has cubic computational/memory complexity.

1. Introduction

> Traditionally, stereo reconstruction is decomposed into three important steps: feature extraction (for matching cost computation), matching cost aggregation and disparity prediction. And cost aggregation is a key step needed to obtain accurate disparity estimations in challenging regions.
> In this work, the proposed solution considerably increase accuracy, while decreasing both memory and computation costs.
>
> - First, we introduce a semi-global guided aggregation layer (SGA) which implements a differentiable approximation of semi-global matching (SGM) and aggregates the matching cost in different directions over the whole image. This enables accurate estimations in occluded regions or large textureless/reflective regions.
> - Second, we introduce a local guided aggregation layer (LGA) to cope with thin structures and object edges in order to recover the loss of details caused by down-sampling and up-coming layers.
>
> A cost aggregation block with only two GA layers and two 3D convolutional layers easily outperforms the state-of-the-art GC-Net, which has nineteen 3D convolutional layers. More

importantly, one GA layer has only 1/100 computational complextity in terms of FLOPs (floating-point operations) as that of a 3D convolution.



(a) Input image (b) GC-Net [13]

(c) Our GA-Net-2 (d) Ground truth

Figure 1: Performance illustrations. (a) a challenging input image. (b) Result of the state-of-the-art method GC-Net [13] which has nineteen 3D convolutional layers for matching cost aggregation. (c) Result of our GA-Net-2, which only uses two proposed GA layers and two 3D convolutional layers. It aggregates the matching information into the large textureless region and is an order of magnitude faster than GC-Net. (d) Ground truth.

2. Related Work

2.1 Deep Neural Networks for Stereo Matching

Recently, end-to-end deep neural network models have become popular.
DispNet for disparity estimation
Pang built a two-stage convolutional neural network to first estimate and then refine the disparity maps.
GC-Net incorporated the feature extraction, matching cost aggregation and disparity estimation into a single end-to-end deep neural model to get state-of-the-art accuracy on several benchmarks.
PSMNet

2.2 Cost Aggregation

2.2.1 Local Cost Aggregation

> The cost volume C is formed of matching costs at each pixel's location for each candidate disparity value d.

2.2.2 Semi-Global Matching

> When enforcing (semi-)global aggregation, the matching cost and the smoothness constraints formulated into one energy function with the disparity map of the input image as D.

3. Guided Aggregation Net

> In this section, we describe our proposed guided aggregation network (GA-Net), including the guided aggregation (GA) layers and the improved network architecture.

3.1 Guided Aggregation Layers

3.1.3 Efficient Implementation

> We use several 2D convolutional layers to build a fast guidance subnet. Ut uses the reference image as input and outputs the aggregation weights w.

3.2 Network Architecture

> The GA-Net consists of four parts:
>
> - the feature extraction block, we use a stacked hourglass network which is densely connected by concatenations between different layers.
> - the cost aggregation for the 4D cost volume,
> - the guidance subnet to produce the cost aggregation weights
> - the disparity regression.

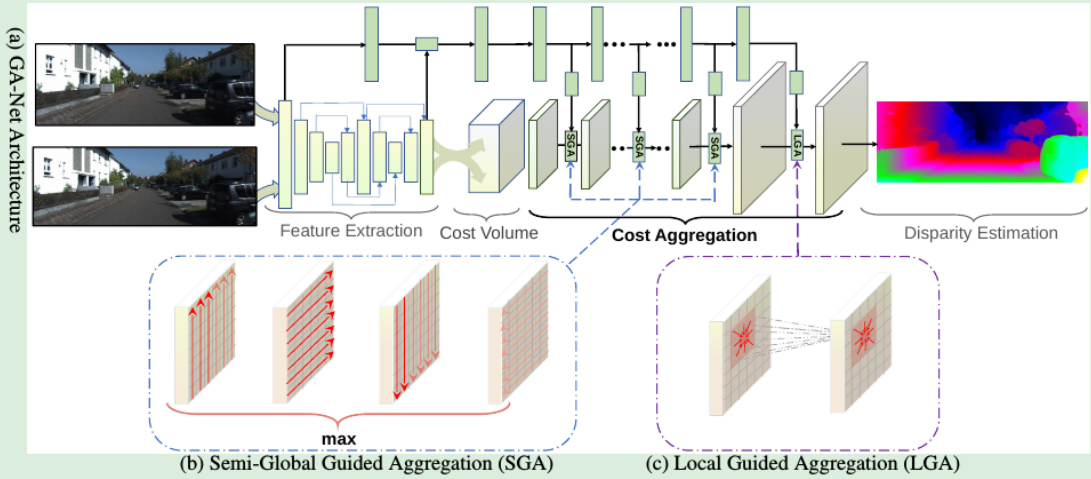


Figure 2: (a) Architecture overview. The left and right images are fed to a weight-sharing feature extraction pipeline. It consists of a stacked hourglass CNN and is connected by concatenations. The extracted left and right image features are then used to form a 4D cost volume, which is fed into a cost aggregation block for regularization, refinement and disparity regression. The guidance subnet (green) generates the weight matrices for the guided cost aggregations (SGA and LGA). (b) SGA layers semi-globally aggregate the cost volume in four directions. (c) The LGA layer is used before the disparity regression and locally refines the 4D cost volume for several times.

3.3 Loss Function

4. Experiments

4.1 Ablation Study

> We evaluate the performance of GA-Nets with different settings, including different architectures and different number (0-4) of GA layers.
> The guided aggregation models significantly outperform the baseline setting which only has 3D convolutional layers for cost aggregation.

4.2 Effects of Guided Aggregations

> In this section, we compare the **guided aggregation strategies** with other **matching cost aggregation methods**. We also analyze the effects of the GA layers by observing the post-softmax probabilities output by different models.
> Firstly, the proposed GA-Nets are compared with the cost aggregation archirectures in GC-Net and PSMNet. GA-Nets have fewer parameters, run at a faster speed and achieve better accuracy.
> We also study the effects of the GA layers by comparing with the same archirectures without GA steps. These baseline models “GA-Nets” have the same network architectures and all

other settings except the there is no GA layer implemented. For all these models, GA layers have significantly improved the model's accuracy (by 0.5-1.0 pixels in average EPE).
Finally, in order to observe and analyze the effects of GA layers, we plot the post-softmax probabilities with respect to a range of candidate disparities. These probabilities are directly used for disparity estimation and **can reflect the effectiveness of the cost aggregation strategies**.
For large textureless regions, there would be a lot of noise since there is no any distinctive features in these regions for correct matching. The SGA layers successfully suppress these noise in the probabilities by aggregating surrounding matching information. The LGA layer further concentrates the probability peak on the ground truth value.

4.3 Comparisons with SGMs and 3D Convolutions

The SGA layer is a differentiable approximation of the SGM. But it produces far better result compared with both the original SGM with handcrafted features and MC-CNN with CNN based features. This is because 1) SGA does not have any user-defined parameters that all learned in an end-to-end fashion. 2) The aggregation of SGA is fully guided and controlled by the weight matrices. The guidance subnet learns effective geometrical and contextual knowledge to control the directions, scopes and strengths of the cost aggregations.
Our SGA layer is also more efficient and effective than the 3D convolutional layer. This is because 3D convolutional layer could only aggregate in a local region restricted by the kernel size. As a result, as series of 3D convolutions along with encode and decode archirectures are indispensable in order to achieve good results.
As a comparison, SGA layer aggregates semi-globally in a single layer which is more efficient. Another advantage of the SGA is that the aggregation's direction, scope and strength are fully guided by variable weights according to different geometrical and contextual information in different locations.

4.4 Complexity and Real-time Models

The SGA layer are much faster and more effective than 3D convolutions. This allows us to build an accurate real-time model.

4.5 Evaluations on Benchmarks

5. Conclution

> In this paper, we developed much more efficient and effective guided matching cost aggregation (GA) strategies, including the semi-global aggregation (SGA) and the local guided aggregation (LGA) layers for end-to-end stereo matching.
> The GA layers significantly improve the accuracy of the disparity estimation in challenging regions, such as occlusions, large textureless/reflective regions and thin structures.
> The GA layers can be used to replace computationally costly 3D convolutions and get better accuracy.