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

Feihu Zhang, Victor Prisacariu, Ruigang Yang, Philip H.S. Torr
University of Oxford, Baidu Research

Code: https://github.com/feihuzhang/GANet

1. Key Steps of Stereo Matching:

- Feature Extraction
  - Patches [Zbontar et al. 2015], Pyramid [Chang et al. 2018].
  - Encoder-decoder [Kendall et al. 2017], etc.
- **Matching Cost Aggregation**
  - **Feature based matching cost is often ambiguous.**
  - **Wrong matches easily have a lower cost than correct ones.**
- Disparity Estimation
  - Classification loss, Disparity regression [Kendall et al. 2017].

2. Problem and Target:

Matching Cost Aggregation:

- Current Deep Neural Networks:
  - Only 2D/3D convolutions
- Traditional Methods:
  - Geometric, Optimization
  - SGM [Hirschmuller. 2008], CostFilter [Hosni et al. 2013], etc.

Formulate Traditional Geometric & Optimization into Neural Networks.

3. Contributions:

- **Semi-Global Aggregation (SGA) Layer**
  - Differentiable SGM.
  - Aggregate Matching Costs Over Whole Image.
- **Local Guided Aggregation (LGA) Layer**
  - Learn Guided Filtering.
  - Refine Thin Structures and Edges.
  - Recover Loss of Accuracy Caused by Down-sampling.

4. Energy Function and SGM:

$$E(D) = \sum_{\mathbf{p}} \{ C_{\mathbf{p}}(D_{\mathbf{p}}) + \sum_{\mathbf{q}\in N_{\mathbf{p}}} P_1 \cdot \delta(|D_{\mathbf{p}} - D_{\mathbf{q}}| = 1) + \sum_{\mathbf{q}\in N_{\mathbf{p}}} P_2 \cdot \delta(|D_{\mathbf{p}} - D_{\mathbf{q}}| > 1) \}.$$

Sum of Matching Costs

Smoothness Penalties

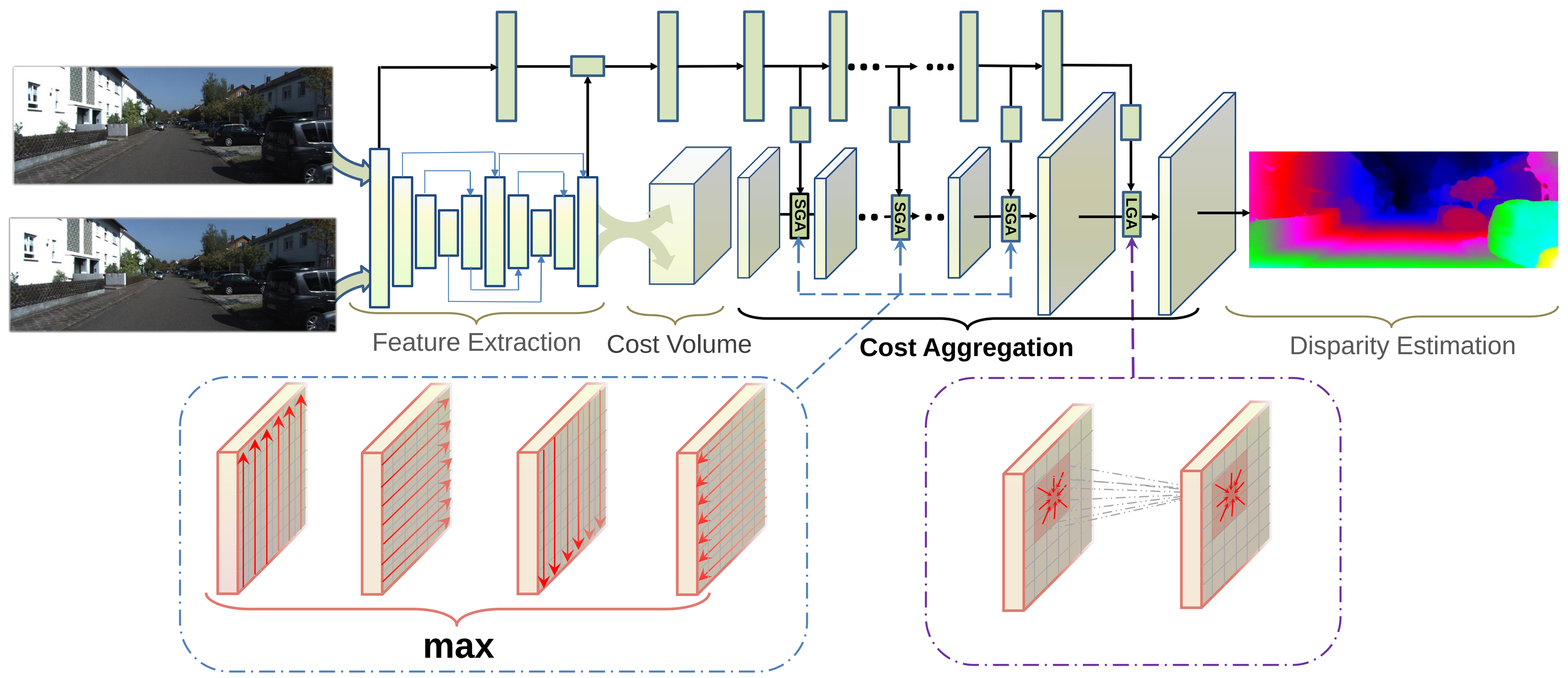
- **Approximate Solution: SGM**

$$C_{\mathbf{r}}^A(\mathbf{p},d) = C(\mathbf{p},d) + \min \begin{cases} C_{\mathbf{r}}^A(\mathbf{p}-\mathbf{r},d), \\ C_{\mathbf{r}}^A(\mathbf{p}-\mathbf{r},d-1) + P_1, \\ C_{\mathbf{r}}^A(\mathbf{p}-\mathbf{r},d+1) + P_1, \\ \min_i C_{\mathbf{r}}^A(\mathbf{p}-\mathbf{r},i) + P_2. \end{cases}$$

Cannot be Used in Deep Neural Networks

- Produce only zeros.
- Not immediately differentiable.
- Produce fronto-parallel surfaces.

- User-defined parameters.
- Hard to tune.
- Fixed for all locations.



SGA Layer **LGA Layer**

5. SGM to SGA Layer:

- User-defined param (P_1, P_2) --> learnable weights (W_1, … W_4):
  - *Learnable and adaptive in different scenes and locations.*
- Second/internal "min" --> "max" selection:
  - *Maximize the probability at the ground truth labels.*
  - *Avoid zeros and negatives, more effective.*
- First "min" --> weighted "sum":
  - *Proven effective in [Springenberg, et al, 2014], no loss of accuracy.*
  - *Reduce fronto-parallel surfaces in large textureless regions.*
  - *Avoid zeros and negatives.*

$$C_{\mathbf{r}}^A(\mathbf{p},d) = C(\mathbf{p},d) + \min \begin{cases} C_{\mathbf{r}}^A(\mathbf{p}-\mathbf{r},d), \\ C_{\mathbf{r}}^A(\mathbf{p}-\mathbf{r},d-1) + P_1, \\ C_{\mathbf{r}}^A(\mathbf{p}-\mathbf{r},d+1) + P_1, \\ \min_i C_{\mathbf{r}}^A(\mathbf{p}-\mathbf{r},i) + P_2. \end{cases}$$

SGM Equation

$$C_{\mathbf{r}}^A(\mathbf{p},d) = C(\mathbf{p},d) + \mathrm{sum} \begin{cases} \mathbf{w}_1(\mathbf{p},\mathbf{r}) \cdot C_{\mathbf{r}}^A(\mathbf{p}-\mathbf{r},d), \\ \mathbf{w}_2(\mathbf{p},\mathbf{r}) \cdot C_{\mathbf{r}}^A(\mathbf{p}-\mathbf{r},d-1), \\ \mathbf{w}_3(\mathbf{p},\mathbf{r}) \cdot C_{\mathbf{r}}^A(\mathbf{p}-\mathbf{r},d+1), \\ \mathbf{w}_4(\mathbf{p},\mathbf{r}) \cdot \max_i C_{\mathbf{r}}^A(\mathbf{p}-\mathbf{r},i). \end{cases}$$

SGA Layer

6. LGA Layer:

- Learn guided $3 \times k \times k$ filtering kernel for each location/pixel.
- Locally refine thin structures and edges.
- Recover loss of accuracy caused by down-sampling.

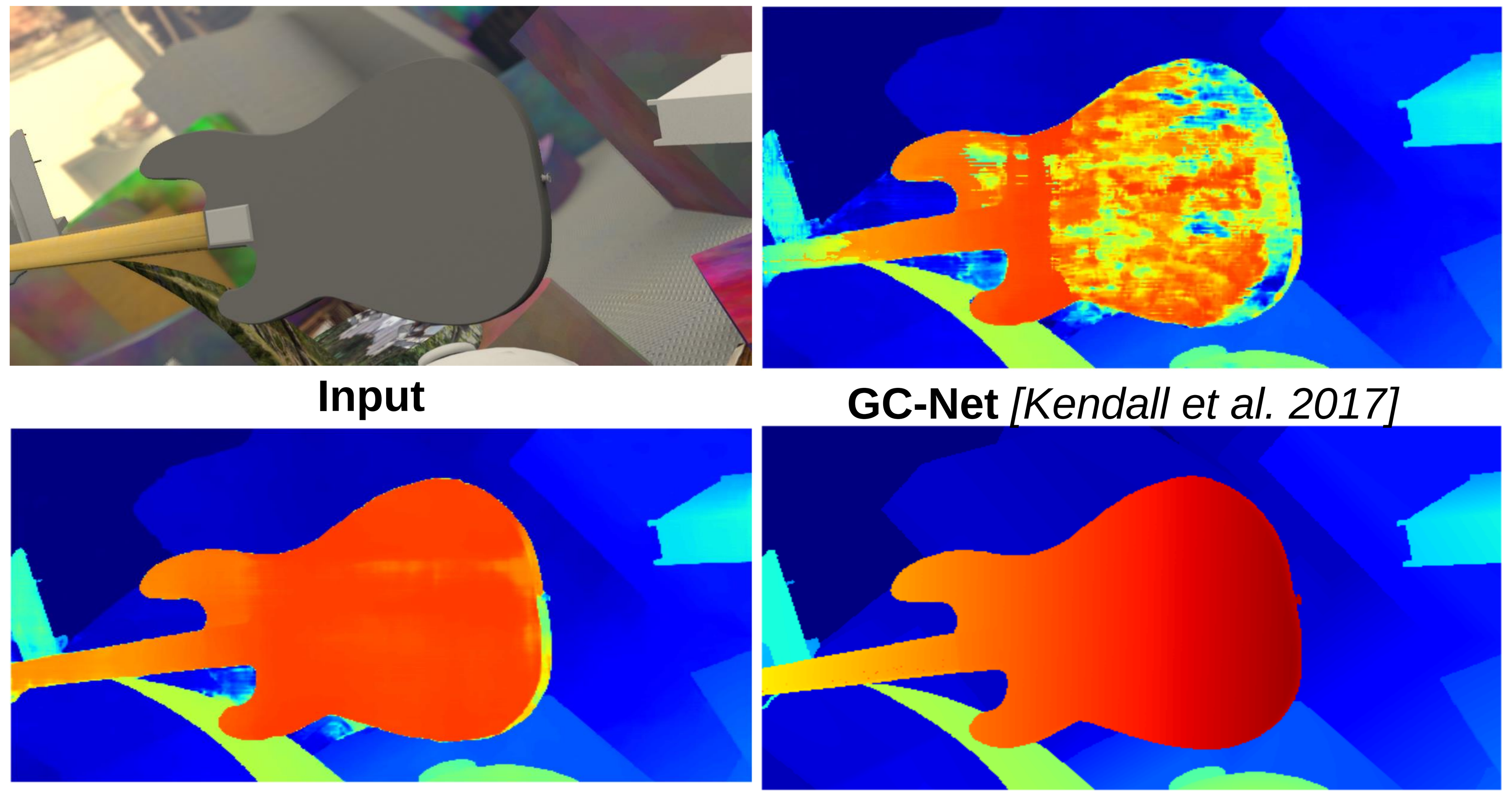
$$C^A(\mathbf{p},d) = \mathrm{sum} \begin{cases} \sum_{\mathbf{q}\in N_{\mathbf{p}}} \omega_0(\mathbf{p},\mathbf{q}) \cdot C(\mathbf{q},d), \\ \sum_{\mathbf{q}\in N_{\mathbf{p}}} \omega_1(\mathbf{p},\mathbf{q}) \cdot C(\mathbf{q},d-1), \\ \sum_{\mathbf{q}\in N_{\mathbf{p}}} \omega_2(\mathbf{p},\mathbf{q}) \cdot C(\mathbf{q},d+1). \end{cases}$$

$$s.t. \sum_{\mathbf{q}\in N_{\mathbf{p}}} \omega_0(\mathbf{p},\mathbf{q}) + \omega_1(\mathbf{p},\mathbf{q}) + \omega_2(\mathbf{p},\mathbf{q}) = 1$$

7. Experimental Results:

- **Evaluation and Comparisons on SceneFlow Dataset**

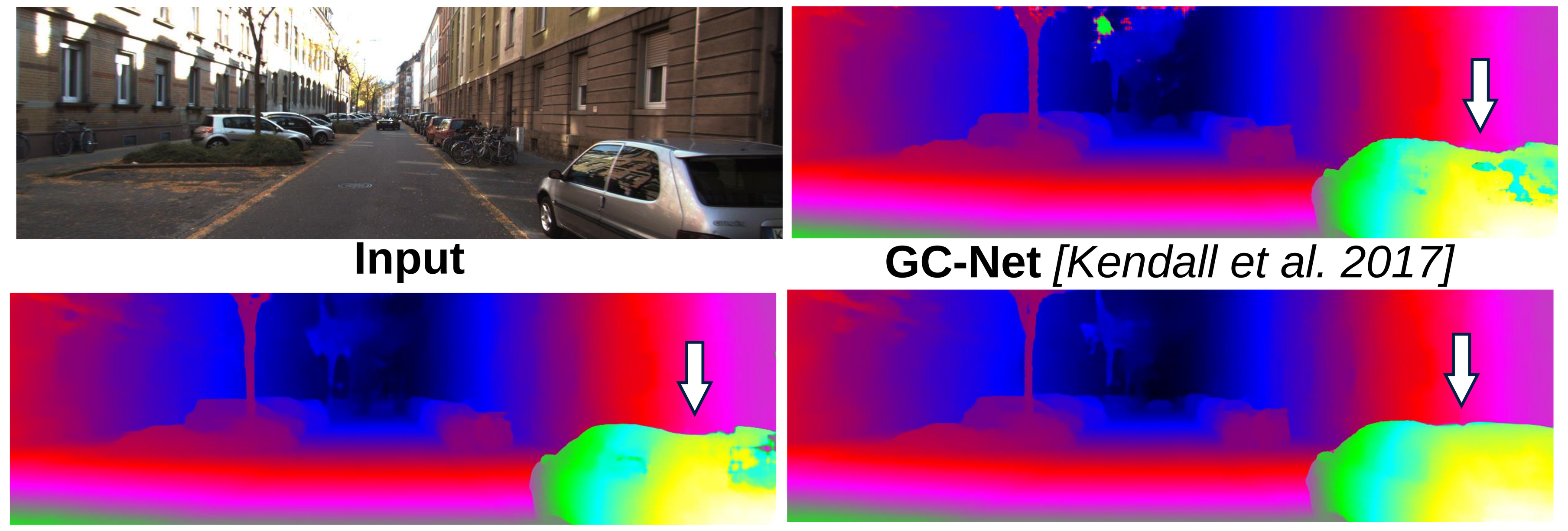
Models	3D conv layers	GA layers	Avg. EPE (pixel)	Error rate (%)
GC-Net	19	-	1.80	15.6
PSMNet	35	-	1.09	12.1
GANet-15	15	5	0.84	9.9
GANet-deep	22	9	**0.78**	**8.7**



Input — **GC-Net** *[Kendall et al. 2017]* — **Our GANet-2** — **Ground Truth**

- **Evaluation and Comparisons on KITTI Benchmarks**

Models	KITTI 2012 benchmark		KITTI 2015 benchmark	
	Non-Occluded	All Area	Non-Occluded	All Area
GC-Net	1.77	2.30	2.61	2.87
PSMNet	1.49	1.89	2.14	2.32
GANet-15	1.36	1.80	1.73	1.93
GANet-deep	**1.19**	**1.60**	**1.63**	**1.81**



Input — **GC-Net** *[Kendall et al. 2017]* — **PSMNet** *[Chang et al. 2018]* — **Our GANet-15**