AANet: Adaptive Aggregation Network for Efficient Stereo Matching Supplementary Material

Haofei Xu Juyong Zhang*
University of Science and Technology of China
xhf@mail.ustc.edu.cn, juyong@ustc.edu.cn

In this supplementary document, we briefly review traditional cross-scale cost aggregation algorithm [1] to make this paper self-contained.

For cost volume $\boldsymbol{C} \in \mathbb{R}^{D \times H \times W}$, [1] reformulates the local cost aggregation from an optimization perspective:

$$\tilde{\boldsymbol{C}}(d, \boldsymbol{p}) = \arg\min_{z} \sum_{\boldsymbol{q} \in N(\boldsymbol{p})} w(\boldsymbol{p}, \boldsymbol{q}) \| z - \boldsymbol{C}(d, \boldsymbol{q}) \|^2, \quad (1)$$

where $\tilde{\boldsymbol{C}}(d, \boldsymbol{p})$ denotes the aggregated cost at pixel $\boldsymbol{p}$ for disparity candidate d, pixel $\boldsymbol{q}$ belongs to the neighbors $N(\boldsymbol{p})$ of $\boldsymbol{p}$, and w is the weighting function to measure the similarity of pixel $\boldsymbol{p}$ and $\boldsymbol{q}$. The solution of this weighted least square problem (1) is

$$\tilde{\boldsymbol{C}}(d, \boldsymbol{p}) = \sum_{\boldsymbol{q} \in N(\boldsymbol{p})} w(\boldsymbol{p}, \boldsymbol{q}) \boldsymbol{C}(d, \boldsymbol{q}). \quad (2)$$

Thus, different local cost aggregation methods can be reformulated within a unified framework.

Without considering multi-scale interactions, the multi-scale version of Eq. (1) can be expressed as

$$\tilde{\boldsymbol{v}} = \arg\min_{\{z^s\}_{s=1}^{S}} \sum_{s=1}^{S} \sum_{\boldsymbol{q}^s \in N(\boldsymbol{p}^s)} w(\boldsymbol{p}^s, \boldsymbol{q}^s) \| z^s - \boldsymbol{C}^s(d^s, \boldsymbol{q}^s) \|^2, \quad (3)$$

where $\boldsymbol{p}^s$ and d^s denote pixel and disparity at scale s, respectively, and $\boldsymbol{p}^{s+1} = \boldsymbol{p}^s/2$, $d^{s+1} = d^s/2$, $\boldsymbol{p}^1 = \boldsymbol{p}$ and $d^1 = d$. The aggregated cost at each scale is denoted as

$$\tilde{\boldsymbol{v}} = [\tilde{\boldsymbol{C}}^1(d^1, \boldsymbol{p}^1), \tilde{\boldsymbol{C}}^2(d^2, \boldsymbol{p}^2), \cdots, \tilde{\boldsymbol{C}}^S(d^S, \boldsymbol{p}^S)]^T. \quad (4)$$

The solution of Eq. (3) is obtained by performing cost aggregation at each scale independently:

$$\tilde{\boldsymbol{C}}^s(d^s, \boldsymbol{p}^s) = \sum_{\boldsymbol{q}^s \in N(\boldsymbol{p}^s)} w(\boldsymbol{p}^s, \boldsymbol{q}^s) \boldsymbol{C}^s(d^s, \boldsymbol{q}^s), \quad s = 1, 2, \cdots, S. \quad (5)$$

By enforcing the inter-scale consistency on the cost volume, we can obtain the following optimization problem:

$$\hat{\boldsymbol{v}} = \arg\min_{\{z^s\}_{s=1}^{S}} \left(\sum_{s=1}^{S} \sum_{\boldsymbol{q}^s \in N(\boldsymbol{p}^s)} w(\boldsymbol{p}^s, \boldsymbol{q}^s) \| z^s - \boldsymbol{C}^s(d^s, \boldsymbol{q}^s) \|^2 + \lambda \sum_{s=2}^{S} \| z^s - z^{s-1} \|^2 \right), \quad (6)$$

where λ is a parameter to control the regularization strength, and $\hat{\boldsymbol{v}}$ is denoted as

$$\hat{\boldsymbol{v}} = [\hat{\boldsymbol{C}}^1(d^1, \boldsymbol{p}^1), \hat{\boldsymbol{C}}^2(d^2, \boldsymbol{p}^2), \cdots, \hat{\boldsymbol{C}}^S(d^S, \boldsymbol{p}^S)]^T. \quad (7)$$

The optimization problem (6) is convex and can be solved analytically (see details in [1]). The solution can be expressed as

$$\hat{\boldsymbol{v}} = P \tilde{\boldsymbol{v}}, \quad (8)$$

where P is an $S \times S$ matrix. That is, the final cost volume is obtained through the adaptive combination of the results of cost aggregation performed at different scales.

Inspired by this conclusion, we design our cross-scale cost aggregation architecture as

$$\hat{\boldsymbol{C}}^s = \sum_{k=1}^{S} f_k(\tilde{\boldsymbol{C}}^k), \quad s = 1, 2, \cdots, S, \quad (9)$$

where f_k is defined by neural network layers.

References

[1] Kang Zhang, Yuqiang Fang, Dongbo Min, Lifeng Sun, Shiqiang Yang, Shuicheng Yan, and Qi Tian. Cross-scale cost aggregation for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1590–1597, 2014. 1

*Corresponding author