Fast Multi-frame Stereo Scene Flow with Motion Segmentation
— Supplementary Material —

Tatsunori Taniai
RIKEN AIP

Sudipta N. Sinha
Microsoft Research

Yoichi Sato
The University of Tokyo

In the supplementary material we present details of our SGM stereo and flow implementations (used in Sec. 4.1 and Sec. 4.5) as well as the segmentation ground prior (used in Sec. 4.7) that were omitted from the main paper due to the limit on page length. We also discuss parameter settings and their effects on our method. Note that we review the SGM algorithm as proposed by Hirschmuller [3], but describe the algorithm using our own notation to be consistent with the main paper. We also provide additional qualitative results and comparisons with state-of-the-art methods in the supplementary video.

A. SGM Stereo

In the binocular and epipolar stereo stages (Sec. 4.1 and 4.3), we solve stereo matching problems using the semi-global matching (SGM) algorithm [3]. Here, stereo matching is cast as a discrete labeling problem, where we estimate the disparity map $\mathcal{D}_\mathbf{p} = \mathcal{D}(\mathbf{p}) : \Omega \to D$ (where $D = \{D_{\text{min}}, \cdots, D_{\text{max}}\}$ is the disparity range) that minimizes the following 2D Markov random field (MRF) based energy function.

$$E_{\text{stereo}}(\mathcal{D}) = \sum_{\mathbf{p}\in\Omega} C_\mathbf{p}(\mathcal{D}_\mathbf{p}) + \sum_{(\mathbf{p},\mathbf{q})\in N} cV_{\mathbf{pq}}(\mathcal{D}_\mathbf{p}, \mathcal{D}_\mathbf{q}). \tag{A1}$$

Here, $C_\mathbf{p}(\mathcal{D}_\mathbf{p})$ is the unary data term that evaluates photo-consistencies between the pixel $\mathbf{p}$ in the left image I^0 and its corresponding pixel $\mathbf{p}' = \mathbf{p} - (\mathcal{D}_\mathbf{p}, 0)^T$ at the disparity $\mathcal{D}_\mathbf{p}$ in the right image I^1. $V_{\mathbf{pq}}(\mathcal{D}_\mathbf{p}, \mathcal{D}_\mathbf{q})$ is the pairwise smoothness term defined for neighboring pixel pairs $(\mathbf{p}, \mathbf{q}) \in N$ on the 8-connected pixel grid. In SGM, this term is usually defined as

$$V_{\mathbf{pq}}(\mathcal{D}_\mathbf{p}, \mathcal{D}_\mathbf{q}) = \begin{cases} 0 & \text{if } \mathcal{D}_\mathbf{p} = \mathcal{D}_\mathbf{q} \\ P_1 & \text{if } |\mathcal{D}_\mathbf{p} - \mathcal{D}_\mathbf{q}| = 1 \\ P_2 & \text{otherwise} \end{cases}. \tag{A2}$$

Here, P_1 and P_2 ($0 < P_1 < P_2$) are smoothness penalties. The coefficient c in Eq. (A1) is described later.

While the exact inference of Eq. (A1) is NP-hard, SGM decomposes the 2D MRF into many 1D MRFs along 8 cardinal directions $\mathbf{r}$ and minimizes them using dynamic programming [3]. This is done by recursively updating the following cost arrays $L_\mathbf{r}(\mathbf{p}, d)$ along 1D scan lines in the directions $\mathbf{r}$ from the image boundary pixels.

$$L_\mathbf{r}(\mathbf{p}, d) = C_\mathbf{p}(d) + \min_{d'\in D} \left[L_\mathbf{r}(\mathbf{p} - \mathbf{r}, d') + V_{\mathbf{pq}}(d, d')\right] - \min_{d'\in D} L_\mathbf{r}(\mathbf{p} - \mathbf{r}, d'). \tag{A3}$$

Here, by introducing the following normalized scan-line costs

$$\bar{L}_\mathbf{r}(\mathbf{p}, d) = L_\mathbf{r}(\mathbf{p}, d) - \min_{d'\in D} L_\mathbf{r}(\mathbf{p}, d'), \tag{A4}$$

the updating rule of Eq. (A3) is simplified as follows.

$$L_{\mathbf{r}}(\mathbf{p}, d) = C_{\mathbf{p}}(d) + \min_{d' \in D} \left[\bar{L}_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, d') + V_{\mathbf{pq}}(d, d')\right] \tag{A5}$$

$$= C_{\mathbf{p}}(d) + \min\{\bar{L}_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, d), \bar{L}_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, d-1) + P_1, \bar{L}_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, d+1) + P_1, P_2\} \tag{A6}$$

Then, the scan-line costs by the 8 directions are aggregated as

$$S(\mathbf{p}, d) = \sum_{\mathbf{r}} L_{\mathbf{r}}(\mathbf{p}, d), \tag{A7}$$

from which the disparity estimate at each pixel $\mathbf{p}$ is retrieved as

$$\mathcal{D}_{\mathbf{p}} = \underset{d \in D}{\operatorname{argmin}}\, S(\mathbf{p}, d). \tag{A8}$$

Recently, Drory *et al*. [2] showed that the SGM algorithm is a variant of message passing algorithms such as belief propagation and TRW-T [9] that approximately optimize Eq. (A1). Here, the coefficient c in Eq. (A1) is a scaling factor that accounts for an overweighting effect on the data term during SGM ($c = 1/8$ when using 8 directions) [2].

Drory *et al*. [2] also proposed an uncertainty measure $\mathcal{U}$ that is computed as

$$\mathcal{U}(\mathbf{p}) = \min_{d} \sum_{\mathbf{r}} L_{\mathbf{r}}(\mathbf{p}, d) - \sum_{\mathbf{r}} \min_{d} L_{\mathbf{r}}(\mathbf{p}, d). \tag{A9}$$

$\mathcal{U}(\mathbf{p})$ is lower-bounded by 0, and becomes 0 when minimizers of 8 individual scan-line costs agree. Since the first and second term in Eq. (A9) are respectively computed in Eqs. (A8) and (A3), the computation of $\mathcal{U}(\mathbf{p})$ essentially does not require computational overhead.

In our implementation of SGM, we use the data term $C_{\mathbf{p}}(\mathcal{D}_{\mathbf{p}})$ defined using truncated normalized cross-correlation in Eq. (5) in the main paper. The smoothness penalties P_1 and P_2 are defined as follows.

$$P_1 = \lambda_{\text{sgm}} / |\mathbf{p} - \mathbf{q}| \tag{A10}$$

$$P_2 = P_1 \left(\beta + \gamma w^{\text{col}}_{\mathbf{pq}}\right) \tag{A11}$$

Here, $w^{\text{col}}_{\mathbf{pq}}$ is the color edge-based weight used in Eq. (15) and we use parameters of $(\lambda_{\text{sgm}}, \beta, \gamma) = (200/255, 2, 2)$. The disparity range is fixed as $\{D_{\text{min}}, \cdots, D_{\text{max}}\} = \{0, \cdots, 255\}$ for the original image size of KITTI (since we downscale the images by a factor of 0.65, the disparity range is also downscaled accordingly). We also set the confidence threshold τ_{u} for the uncertainty map $\mathcal{U}$ to 2000 by visually inspecting $\mathcal{U}(\mathbf{p})$.

B. SGM Flow

We have extended the SGM algorithm for our optical flow problem in Sec. 4.5. Here, we estimate the flow map $\mathcal{F}_{\mathbf{p}} = \mathcal{F}(\mathbf{p}) : \Omega \rightarrow R$ (where $R = ([u_{\text{min}}, u_{\text{max}}] \times [v_{\text{min}}, v_{\text{max}}])$ is the 2D flow range) by minimizing the following 2D MRF energy.

$$E_{\text{flow}}(\mathcal{F}) = \sum_{\mathbf{p} \in \Omega} C'_{\mathbf{p}}(\mathcal{F}_{\mathbf{p}}) + \sum_{(\mathbf{p}, \mathbf{q}) \in N} c V'_{\mathbf{pq}}(\mathcal{F}_{\mathbf{p}}, \mathcal{F}_{\mathbf{q}}). \tag{A12}$$

Similarly to SGM stereo, we use the NCC-based matching cost of Eq. (5) for the data term $C'_{\mathbf{p}}(\mathcal{F}_{\mathbf{p}})$ to evaluate matching photo-consistencies between I^0_t and I^0_{t+1}. We also define the smoothness term as

$$V'_{\mathbf{pq}}(\mathcal{F}_{\mathbf{p}}, \mathcal{F}_{\mathbf{q}}) = \begin{cases} 0 & \text{if } \mathcal{F}_{\mathbf{p}} = \mathcal{F}_{\mathbf{q}} \\ P_1 & \text{if } 0 < \|\mathcal{F}_{\mathbf{p}} - \mathcal{F}_{\mathbf{q}}\| \leq \sqrt{2} \\ P_2 & \text{otherwise} \end{cases} . \tag{A13}$$

Since we use integer flow labels, the second condition in Eq. (A13) is equivalent to saying that the components of the 2D vectors $\mathcal{F}_{\mathbf{q}} = (u_{\mathbf{q}}, v_{\mathbf{q}})$ and $\mathcal{F}_{\mathbf{p}} = (u_{\mathbf{p}}, v_{\mathbf{p}})$ can at-most differ by 1. We use the same smoothness penalties $\{P_1, P_2\}$ and the parameter settings with SGM stereo.

The optimization of Eq. (A12) is essentially the same with SGM stereo, but the implementation of updating scan-line costs in Eq. (A3) was extended to handle the new definition of the pairwise term $V'_{\mathbf{pq}}$. Therefore, Eq. (A6) is modified using a flow label $\mathbf{u} = (u, v) \in R$ as follows.

$$L_{\mathbf{r}}(\mathbf{p}, \mathbf{u}) = C_{\mathbf{p}}(\mathbf{u}) + \min\{\bar{L}_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, \mathbf{u}), \bar{L}_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, \mathbf{u} + \Delta_{\pm 1}) + P_1, P_2\} \tag{A14}$$

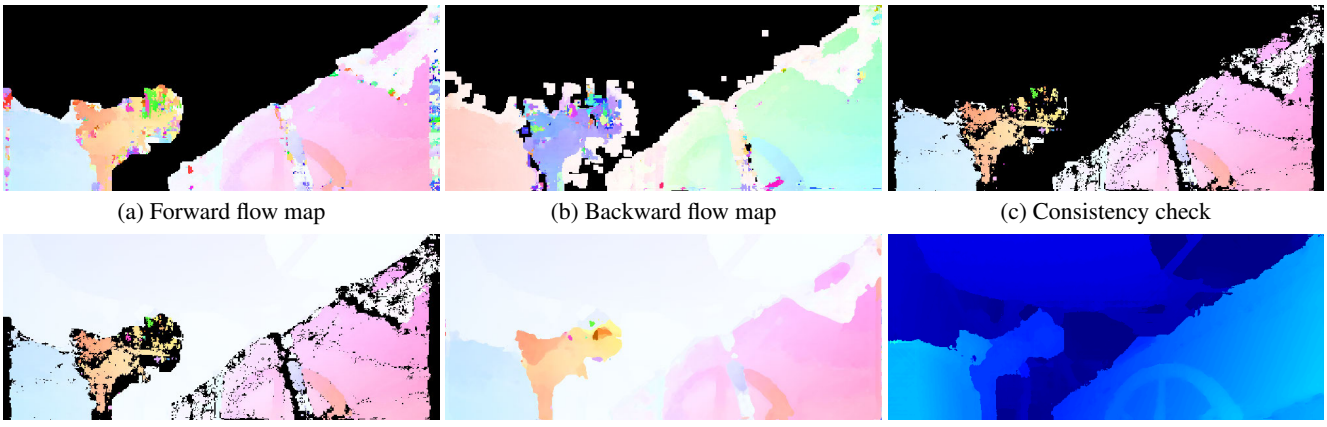
Here, $(\mathbf{u} + \Delta_{\pm 1})$ is enumeration of 8 labels neighboring to $\mathbf{u}$ in the 2D flow space.

C. Refinement of Flow Maps

In the optical flow stage of Sec. 4.5, we refine flow maps using consistency check and weighted median filtering. Similar schemes are commonly employed in stereo and optical flow methods such as [10, 4, 5]. Below we explain these steps.

We first estimate the forward flow map $\mathcal{F}^0$ (from I^0_t to I^0_{t+1}) by SGM for only the foreground pixels of the initial segmentation $\tilde{\mathcal{S}}$ such as shown in Fig. A1 (a). Then, using this flow $\mathcal{F}^0$ and the mask $\tilde{\mathcal{S}}$, we compute a mask in the next image I^0_{t+1} and estimate the backward flow map $\mathcal{F}^1$ (from I^0_{t+1} to I^0_t) for those foreground pixels. This produces a flow map such as shown in Fig. A1 (b). We filter out outliers in $\mathcal{F}^0$ using bi-directional consistency check between $\mathcal{F}^0$ and $\mathcal{F}^1$ to obtain a flow map with holes (Fig. A1 (c)), whose background is further filled by the rigid flow $\mathcal{F}_{\text{rig}}$ (see Fig. A1 (d)). Finally, weighted median filtering is applied for the hole pixels followed by median filtering for all foreground pixels to obtain the non-rigid flow estimate such as shown in Fig. A1 (e).

At the final weighted median filtering step, the filter kernel $\omega^{\text{geo}}_{\mathbf{pq}} = e^{-d_{\mathbf{pq}}/\kappa_{\text{geo}}}$ is computed using geodesic distance $d_{\mathbf{pq}}$ on



(a) Forward flow map (b) Backward flow map (c) Consistency check

(d) Background filling (e) Weighted median filtering (f) Disparity map as guidance image

Figure A1. Process of flow map refinement.

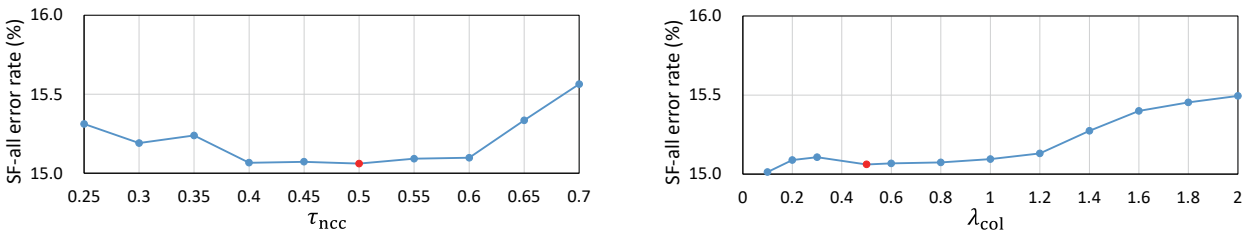


Figure A2. Profiles of scene flow accuracies with reference to parameters τ_{ncc} (left) and λ_{col} (right). The error rates are evaluated on 200 training sequences from KITTI. The scores with the default parameter settings are colored by red.

the disparity map $\mathcal{D}$ (Fig. A1 (f)) as the guidance image. For this, we define the distance between two adjacent pixels as

$$\text{dist}(\mathbf{p}_1, \mathbf{p}_2) = |\mathcal{D}(\mathbf{p}_1) - \mathcal{D}(\mathbf{p}_2)| + \|\mathbf{p}_1 - \mathbf{p}_2\|/100. \tag{A15}$$

The geodesic distance $d_{\mathbf{pq}}$ is then computed for the pixels in the filter window $\mathbf{q} \in W_{\mathbf{p}}$ as the cumulative shortest-path distance from $\mathbf{q}$ to the center pixel $\mathbf{p}$. This is efficiently computed using an approximate algorithm [8]. We use the filter window $W_{\mathbf{p}}$ of 31×31 size and $\kappa_{\text{geo}} = 2$. The subsequent (constant-weight) median filtering further reduces outliers [7], for which we use the window of 5×5 size.

D. Segmentation Ground Prior

The segmentation ground prior term mentioned in Sec. 4.7 is computed as follows. First, we detect the ground plane from the disparity map $\mathcal{D}(\mathbf{p})$. We use RANSAC to fit a disparity plane $[d = au + bv + c]$ defined on the 2D image coordinates. Here, we assume that the cameras in the stereo rig are upright. Therefore, during RANSAC we choose disparity planes whose b is positive and high and $|a|$ is relatively small. Then, we compute the disparity residuals between $\mathcal{D}$ and the ground plane as $r_{\mathbf{p}} = |\mathcal{D}_{\mathbf{p}} - (a\mathbf{p}_u + b\mathbf{p}_v + c)|$, where (a, b, c) are the obtained plane parameters. Our ground prior as a cue of background is then defined as follows.

$$C^{\text{gro}}_{\mathbf{p}} = \lambda_{\text{gro}} \Big(\min(r_{\mathbf{p}}, \tau_{\text{gro}})/\tau_{\text{gro}} - 1 \Big) \tag{A16}$$

When $r_{\mathbf{p}} = 0$, $C^{\text{gro}}_{\mathbf{p}}$ strongly favors background, and when $r_{\mathbf{p}}$ increases to τ_{gro}, it becomes 0. The thresholding value $\tau_{\mathbf{p}}$ is set to $0.01 \times D_{\text{max}}$. We use $\lambda_{\text{gro}} = 10$.

E. Parameter Settings

In this section, we explain our strategy of tuning parameters and also show effects of some parameters. Most of the parameters can be easily interpreted and tuned, and our method is fairly insensitive to parameter settings.

For example, the effects of the threshold τ_{u} for the uncertainty map $\mathcal{U}$ (Sec. 4.1), the threshold τ_w for the patch-variance weight $\omega^{\text{var}}_{\mathbf{p}}$ (Sec. 4.4), and κ_3 of the image edge-based weight $\omega^{\text{str}}_{\mathbf{pq}}$ (Sec. 4.4) can be easily analyzed by direct visualization as shown in Figure 3 (b), Figures 4 (b) and (e).

The parameters of SGM (discussed in Sec. A) can be tuned independently from the whole algorithm.

For the weights $(\lambda_{\text{ncc}}, \lambda_{\text{flo}}, \lambda_{\text{col}}, \lambda_{\text{potts}})$ in Sec. 4.4, we first tuned $(\lambda_{\text{ncc}}, \lambda_{\text{flo}}, \lambda_{\text{potts}})$ on a small number of sequences. Since the ranges of the NCC appearance term (Eq. (11)) and flow term (Eq. (12)) are limited to $[-1, 1]$, they are easy to interpret. Then, we tuned λ_{col} of the color term (Eq. (14)). Here, $\lambda_{\text{potts}}/\lambda_{\text{col}}$ is known to be usually around 10 - 60 from previous work [6, 1].

Even though we fine-tuned τ_{ncc} and λ_{col} for Sintel, they are insensitive on KITTI image sequences. We show the effects of these two parameters for KITTI training sequences in Figure A2. The threshold τ_{ncc} for NCC-based matching costs was adjusted for Sintel because its synthesized images have lesser image noise compared to real images of KITTI. Also, the

weight λ_{col} was adjusted for Sintel, to increase the weight on the prior color term (Sec. 4.7). For Sintel sequences, sometimes moving objects stop moving on a few frames and become stationary momentarily. In such cases, increasing λ_{col} improves the temporal coherence of the motion segmentation results. In the future we will improve the scheme for online learning of the prior color models, which will improve temporal consistency of motion segmentation and also will make λ_{col} more insensitive to settings.

References

[1] Y. Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, volume 1, pages 105–112, 2001.

[2] A. Drory, C. Haubold, S. Avidan, and F. A. Hamprecht. Semi-global matching: a principled derivation in terms of message passing. *Pattern Recognition*, pages 43–53, 2014.

[3] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 30(2):328–341, 2008.

[4] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 35(2):504–511, 2013.

[5] C. R. Michael Bleyer and C. Rother. Patchmatch stereo - stereo matching with slanted support windows. In *Proc. of British Machine Vision Conf. (BMVC)*, pages 14.1–14.11, 2011.

[6] C. Rother, V. Kolmogorov, and A. Blake. "grabcut": Interactive foreground extraction using iterated graph cuts. *ACM Trans. on Graph.*, 23(3):309–314, 2004.

[7] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2439, 2010.

[8] P. J. Toivanen. New geodesic distance transforms for gray-scale images. *Pattern Recogn. Lett.*, 17(5):437–450, 1996.

[9] M. Wainwright, T. Jaakkola, and A. Willsky. Map estimation via agreement on (hyper)trees: Message-passing and linear programming approaches. *IEEE Trans. on Information Theory*, 51:3697–3717, 2002.

[10] Q. Zhang, L. Xu, and J. Jia. 100+ times faster weighted median filter (wmf). In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2830–2837, 2014.