Upgrading Optical Flow to 3D Scene Flow through Optical Expansion

Gengshan Yang[1]*, Deva Ramanan[1,2]
[1]Carnegie Mellon University, [2]Argo AI
{gengshay, deva}@cs.cmu.edu

1. Supplementary Material

1.1. Two-view SfM with normalized scene flow

Problem setup In Sec. 4.5, we discussed using motion-in-depth τ to estimate depth Z for a static scene when the camera pose is given. Here, we solve for the camera motion $(\mathbf{R}_c, \mathbf{t}_c)$ and the 3D shape $\mathbf{P} = (X, Y, Z)$ jointly given normalized scene flow $\hat{\mathbf{t}}$ from two views of a static scene. This can be formulated as a optimization problem, and we want to minimize the 3D re-projection error

$$\begin{aligned} L(\mathbf{R}_c, \mathbf{t}_c, \mathbf{Z}) &= \sum_k w_k d(\mathbf{R}_c\mathbf{P}_k + \mathbf{t}_c, \mathbf{P}_k + \mathbf{t}_k)^2 \\ &= \sum_k w_k ||\mathbf{R}_c\mathbf{P}_k + \mathbf{t}_c - (\mathbf{P}_k + \mathbf{t}_k)||^2 \\ &= \sum_k w_k ||\mathbf{R}_c(Z_k\tilde{\mathbf{P}}_k) + \mathbf{t}_c - (Z_k\tilde{\mathbf{P}}_k + Z_k\hat{\mathbf{t}}_k)||^2 \\ &= \sum_k w_k ||Z_k(\mathbf{R}_c\tilde{\mathbf{P}}_k - \tilde{\mathbf{P}}_k - \hat{\mathbf{t}}_k) + \mathbf{t}_c||^2, \end{aligned}$$

where k is the index of each point, w is the weight assigned to each point, and $\tilde{\mathbf{P}} = \mathbf{K}^{-1}\tilde{\mathbf{p}}$ are normalized 3D coordinates.

Solution Empirically, we find coordinate descent gives a robust solution to the above optimization problem. We alternate between pose $(\mathbf{R}_c, \mathbf{t}_c)$, and scales $\{Z_1, \ldots, Z_K\}$ as follows. Fixing the scale Z_k, the problem becomes finding a rigid transformation between two registered point sets. The solution can be described as: 1) align the center of two point clouds, 2) solve for rotation using SVD, and 3) solve for translation [5]. Fixing $(\mathbf{R}_c, \mathbf{t}_c)$, scales $\{Z_1, \ldots, Z_K\}$ can be obtained by solving a least square problem.

Results We initialize the depth to one for all 3D points $Z_k = 1, k \in \{1 \ldots, N\}$, and perform steepest descent for each sub-problem for five iterations. Qualitative results on KITTI and Blackbird [1] are shown in Fig. 2 and Fig. 3.

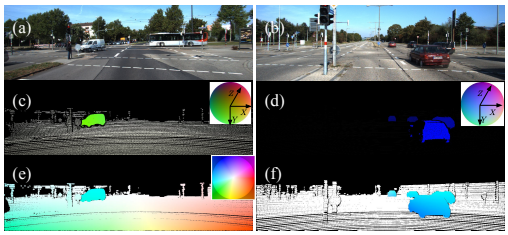


Figure 1. Normalized scene flow v.s. optical flow. (a)-(b): overlaid images of two consecutive frames. (c)-(d): visualization of normalized scene flow using negative and positive color hemispheres. Notice that normalized scene flow provides information about depth change, where in (c) the front car moving left-inwards is colored green and in (d) the car moving left-outwards is colored blue. (e)-(f): visualization of optical flow using the Middlebury color wheel [2]. In comparison, optical flow is not sensitive to change of depth, where left-moving cars are all bluish, no matter moving towards or away from the camera.

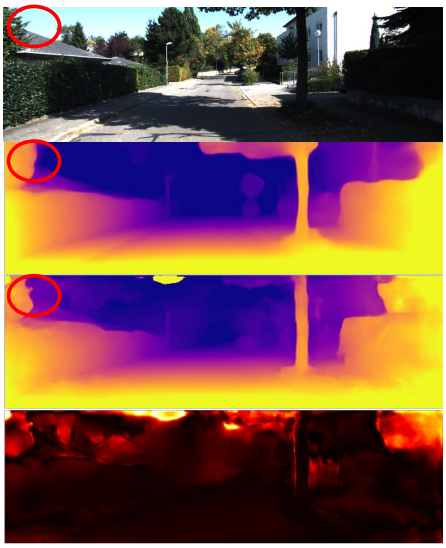
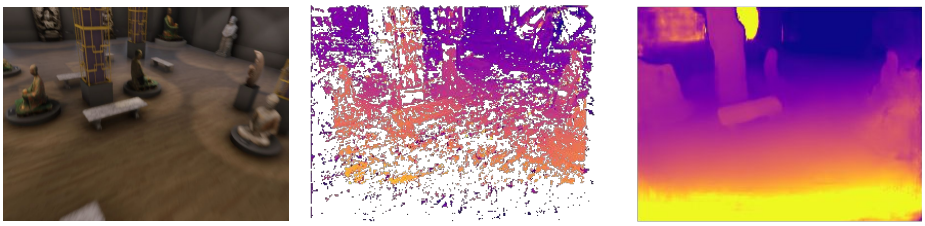
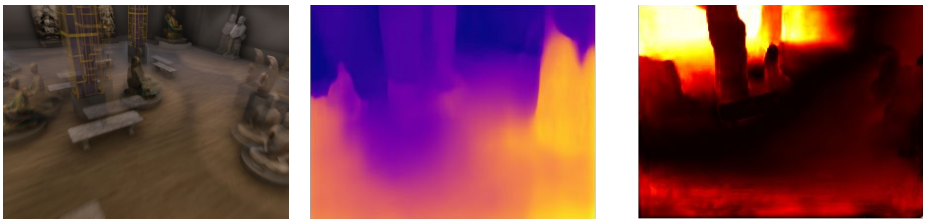


Figure 2. Results on KITTI sequence "10-03-0034" (not in the training sequences) frame 840, where the scene motion can be described by a rigid transform. From top to bottom: reference image, depth prediction from MonoDepth2 [3], our depth prediction and fitting error (uncertainty). Our method jointly estimates camera pose together with scene geometry, and predicts sharper boundaries.

*Code will be available at github.com/gengshan-y/expansion.



Frame one | Result from COLMAP (two view) | Ours



Overlaid two frames | Result from MonoDepth2 | Our-uncertainty

Figure 3. Results on Blackbird dataset [1], which is a unmanned aerial vehicle dataset for aggressive indoor flight. Prior method of estimating the scene geometry either uses sparse point correspondences, for example COLMAP [4], or rely on strong data prior, for example MonoDepth2 [3]. Our method produces dense and reliable depth without strong priors by making use of dense optical expansion.

References

[1] Amado Antonini, Winter Guerra, Varun Murali, Thomas Sayre-McCord, and Sertac Karaman. The blackbird dataset: A large-scale dataset for uav perception in aggressive flight. In *2018 International Symposium on Experimental Robotics (ISER)*, 2018. 1, 2

[2] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *IJCV*, 2011. 1

[3] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. In *ICCV*, 2019. 1, 2

[4] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2

[5] Olga Sorkine-Hornung and Michael Rabinovich. Least-squares rigid motion using svd. *Computing*, 1(1), 2017. 1