Supplementary Materials for UnOS: Unified Unsupervised Optical-flow and Stereo-depth Estimation by Watching Videos

Yang Wang[1] Peng Wang[1] Zhenheng Yang[2] Chenxu Luo[3] Yi Yang[1] Wei Xu[1]
[1]Baidu Research [2] University of Southern California [3] Johns Hopkins University
{wangyang59, wangpeng54, yangyi05, wei.xu}@baidu.com zhenheny@usc.edu chenxuluo@jhu.edu

1. Implementation Details

The whole system was implemented using Tensorflow [1]. When minimizing Eq. 5 in the RDVO module, we used the 3D representation described as in the last line of Eq. 6. When calculating the rigid-aware flow consistency loss term in Eq. 9, the 2D representation in the first line of Eq. 6 was adopted.

In the first stage of training, the smoothness loss of optical flow was applied across the entire image, *i.e.* $\mathcal{L}_{fs} = \mathcal{L}_s(\mathbf{F}_{t\to s}, 1, 2)$. In the third stage of training, the smoothness loss of optical flow was only applied on the moving region, *i.e.* $\mathcal{L}_{fs} = L_s(\mathbf{F}_{t\to s}, 1 - \mathbf{M}_t, 2)$

2. Scene Flow Evaluation

With the estimated disparity, optical flow, camera motion and motion segmentation, we can evaluate our method on the KITTI 2015 scene flow benchmark. The disparity information of second image pair mapped into the reference frame is called D2 in the scene flow benchmark. The static region of D2 can be obtained by transforming D1 using the estimated camera motion, while the moving region of D2 can be obtained by warping the disparity of the second image pair back to the reference frame using the estimated optical flow. The quantitative results are shown in Table. 1. As expected, due to the high quality of our estimated stereo depth and optical flow, our results are significantly better than that presented in EPC [5]. Here in Table. 1, we do not show the performance of EPC since they did not evaluate their method on test server of KITTI. However, the number shown in Tab.3 of the EPC paper on validation set is much worse than ours.

3. Ablation Study

We performed the ablation study by removing the RDVO module and show the corresponding results in Table. 2 and Table. 3, where we can see that removing RDVO hurts the performance.

4. More Qualitative Examples

More qualitative examples of depth, optical flow and motion mask can be found in Fig. 1, Fig. 2 and Fig. 3 respectively.

5. Evaluation on Monkaa Dataset

In order to illustrate the generalizability of our method, we further trained and evaluated our method on the Monkaa dataset [3]. We randomly split the Monkaa dataset into 95% training and 5% validation. We took the models trained on the KITTI dataset and then finetuned on the Monkaa dataset in a pure unsupervised manner. The quantitative results can be found in Table. 4. We compared our full method ”UnOS (Full)” to our baseline methods ”UnOS (Flownet-only)” and ”UnOS(Stereo-Only)” on the tasks of optical flow and disparity estimations. We can see that our full method ”UnOS(Full)” improves over the baseline method on the optical flow estimation task significantly, while performs relatively similar on the disparity estimation task. Qualitative results are shown in Fig. 4.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, page 7, 2017.

[3] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016.

[4] Simon Meister, Junhwa Hur, and Stefan Roth. UnFlow: Unsupervised learning of optical flow with a bidirectional census loss. In *AAAI*, New Orleans, Louisiana, Feb. 2018.

Method	D1-bg	D1-fg	D1-all	D2-bg	D2-fg	D2-all	Fl-fg	Fl-bg	Fl-all	SF-bg	SF-fg	SF-all
ISF	4.12 %	6.17 %	4.46 %	4.88 %	11.34 %	5.95 %	5.40 %	10.29 %	6.22 %	6.58 %	15.63 %	8.08 %
OSF	4.54 %	12.03 %	5.79 %	5.45 %	19.41 %	7.77 %	5.62 %	18.92 %	7.83 %	7.01 %	26.34 %	10.23 %
Ours	5.10 %	14.55 %	6.67 %	9.61 %	24.28 %	12.05 %	16.93 %	23.34 %	18.00 %	19.70 %	35.43 %	22.32 %

Table 1. Scene flow evaluation on the 200 test images of KITTI 2015 benchmark.

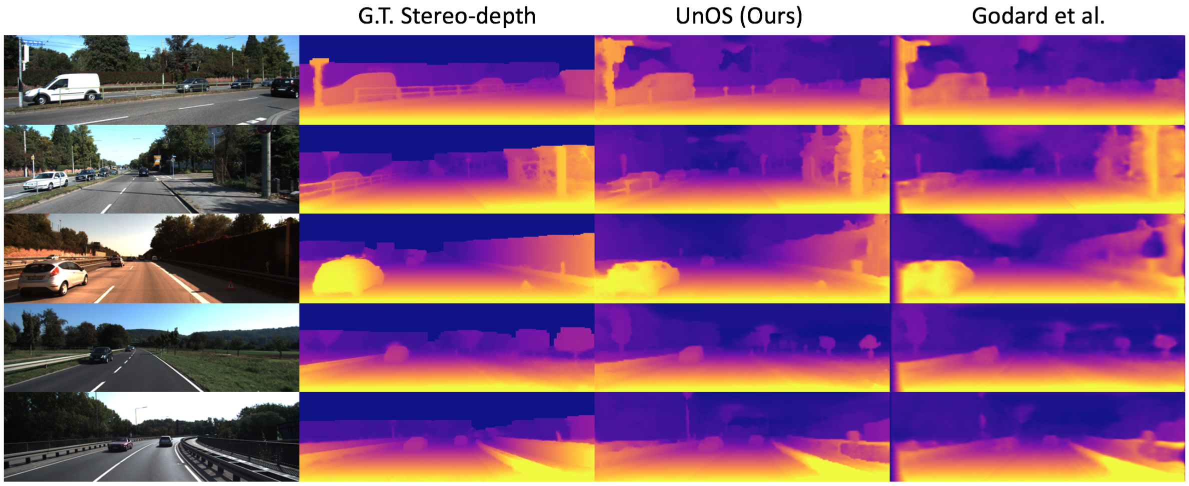

Figure 1. More qualitative results of stereo-depth estimation. From left to right: image, ground truth depth, our depth, and depth from Godard *et al.* [2]

[5] Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, and Ram Nevatia. Every pixel counts: Unsupervised geometry learning with holistic 3d motion understanding. *arXiv preprint arXiv:1806.10556*, 2018.

Method	KITTI 2012			KITTI 2015		
	train noc	train occ	train all	train move	train static	train all
w/o RDVO	1.06	5.72	1.74	6.74	5.28	6.04
Full	1.04	5.18	1.64	5.30	5.39	5.58

Table 2. Ablation study of the optical flow task.

Method	Abs Rel	Sq Rel	RMSE	RMSE log	D1-all
w/o RDVO	0.052	0.610	3.569	0.126	6.442%
Full	0.049	0.515	3.404	0.121	5.943%

Table 3. Ablation study of the stereo depth task.

Method	flow	disparity
UnOS(Stereo-only)	–	15.0%
UnOS(FlowNet-only)	4.18	–
UnOS(Full)	3.16	15.1%

Table 4. Quantitative evaluation on the Monkaa dataset.

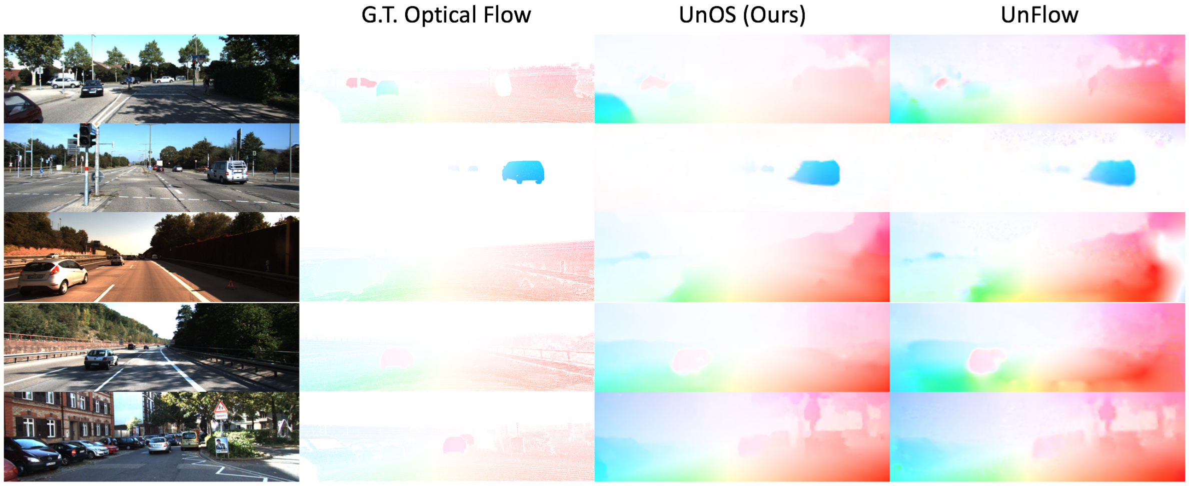

Figure 2. More qualitative results of optical flow estimation. From left to right: image, ground truth optical flow, our optical flow, and optical flow from UnFlow-CSS [4].

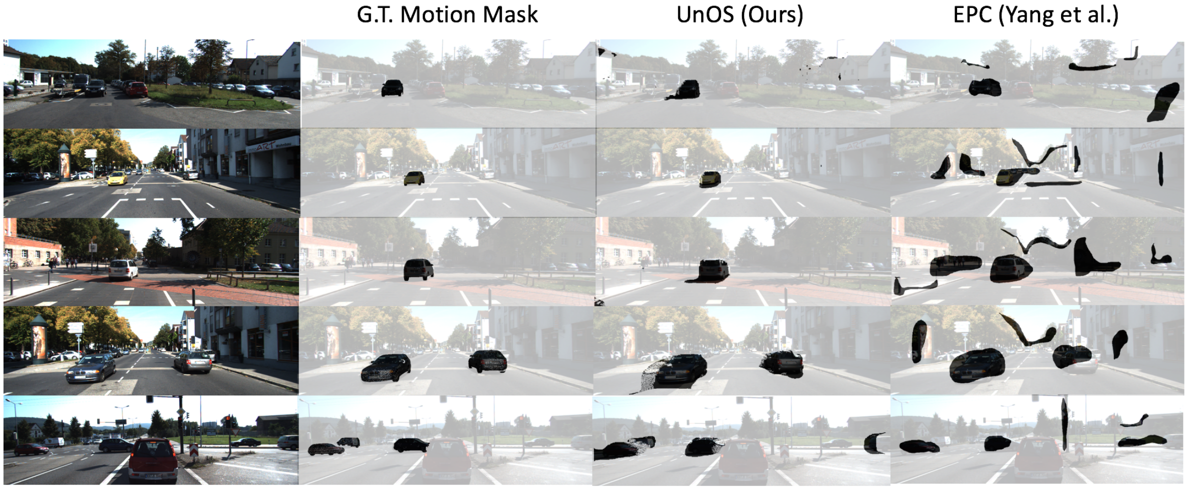

Figure 3. More qualitative results of motion segmentation mask estimation. From left to right: image, ground truth motion mask, our motion mask, and motion mask from EPC [5].

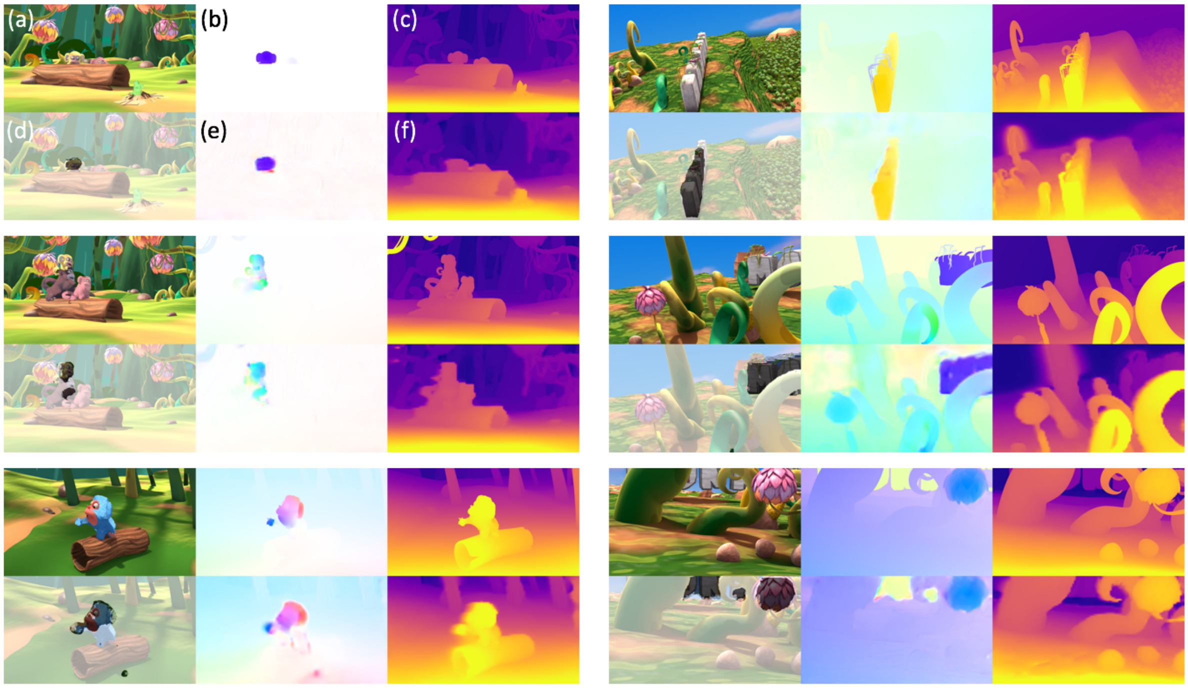

Figure 4. Quantitative examples of our method on the Monkaa dataset. (a) original left image. (b) ground truth optical flow. (c) ground truth disparity. (d) our estimated motion mask overlaid on the left image. (e) our estimated optical flow. (f) our estimated disparity.