LayeredFlow Supplementary Material

Hongyu Wen, Erich Liang, and Jia Deng

Department of Computer Science, Princeton University
{hongyu.wen, erliang, jiadeng}@princeton.edu

1 Details about LayeredFlow Benchmark Data Collection

1.1 Data Statistics

Detailed statistics are shown in Tab. 1.

1.2 Ground Truth Annotation

We used the Python bindings of AprilTag [3] to detect of AprilTag [14] in raw images.

Upon detection, each AprilTag marker is uniquely identified by its ID, the central point of the tag, and the locations of its four corners. To reduce potential errors arising from distortion, especially because some markers are attached to curved surfaces, we limit the generation of ground-truth correspondences to only the four corners of each tag. Consequently, each tag offers four pairs of correspondences.

1.3 Camera Calibration

The image acquisition process involves two cameras mounted on a tripod. Prior to each image capture session, the cameras are calibrated. For each camera calibration session, we take at least 40 pairs of simultaneous photos with the camera pair. We utilize OpenCV [4] library to interface with each camera.

For each camera's image of the chessboard, we identify the 2D key points — the inner corners of the chessboard. By measuring the size of each square in the chessboard, we are able to determine the position of each corner in world coordinates. The relationship between 2D and 3D key points in homogeneous coordinates is represented by the equation:

$$\mathbf{P}_{2D} = \mathbf{K}[\mathbf{R} \mid \mathbf{t}]\mathbf{P}_{3D} \tag{1}$$

Here, $\mathbf{P}_{2D}$ and $\mathbf{P}_{3D}$ denote the 2D and 3D key points, respectively. $\mathbf{K}$ is the camera's intrinsic matrix, and $[\mathbf{R} \mid \mathbf{t}]$ is the camera's extrinsic matrix. This allows us to solve the camera's intrinsic matrices $\mathbf{K}$ and distortion coefficients.

Subsequently, the two cameras undergo stereo calibration. We assume that the origin of world coordinates is located at the center of the left camera, and we calculate $\mathbf{R}_{cam}$ and $\mathbf{t}_{cam}$, which represent the rotation and translation from

	All	Material			Layer		
		Transparent	Reflective	Diffuse	1	2	3
Optical Flow	152627	120737	21459	10431	136799	13988	1840
Stereo	147607	117262	20479	9866	132991	13048	1568

Table 1: Number of stereo pairs and optical flow pairs in our benchmark, categorized by material property and layer index.

the left camera to the right camera. This is done by jointly calibrating $\mathbf{P}_{3D}$, $\mathbf{P}_{2D_l}$, and $\mathbf{P}_{2D_r}$, where $\mathbf{P}_{2D_l}$ and $\mathbf{P}_{2D_r}$ are the 2D key points on the left and right images, respectively. Utilizing all this information, we proceed to rectify the images, ensuring the corresponding points in the two images lie along the same epipolar lines.

2 Details about Synthetic Data Generation

2.1 Scenes and Assets

Our synthetic dataset was created using 30 diverse scenes, enhanced with 100 non-Lambertian assets and 50 random HDR environment textures. All scenes and assets were acquired from BlendSwap [1] under the Creative Commons license. Note that some of the assets are adopted from the other scenes. We acknowledge creators of all assets and scenes, shown in Tab. 2. All HDR images are acquired from HDRi Haven [2] under the Creative Commons Zero license.

2.2 Ground Truth Generation Details

To generate ground truth for optical flow, a typical approach involves using the vector pass in the Blender Cycles engine [8]. However, Cycles does not inherently support the generation of multi-layer ground truth. To address this limitation, we add several new passes to the engine, enabling it to record information each time a ray strikes a surface during the ray tracing process. Specifically, we modified the Cycles source code to capture data for multiple layer masks, 3D positions (useful for depth and disparity calculations), and motion (for optical flow calculation) each time a ray from air strikes an object surface. This modification allows for the generation of multi-layer ground truth that is perfectly aligned with human perception and preserves the effects of light refraction.

3 Training Details

All models are implemented in PyTorch [15]. The fine-tuned version of RAFT is trained on eight RTX 3090 GPUs with a batch size of 20, directly following the training procedure and data augmentation in RAFT [21]. The learning rate is set to 1e-5.

Type	Category	Link	Creator	Type	Category	Link	Creator
Scene	Kitchen	link	TheCGNinja	Scene	Living Room	link	Wig42
Scene	Kitchen	link	cenobi	Scene	Living Room	link	Mikel007
Scene	Kitchen	link	Warcos	Scene	Living Room	link	blenderjunky
Scene	Kitchen	link	unangelo	Scene	Living Room	link	ermmus
Scene	Kitchen	link	oldtimer	Scene	Living Room	link	oldtimer
Scene	Kitchen	link	blenderjunky	Scene	Bedroom	link	SlykDrako
Scene	Kitchen	link	MarcoD	Scene	Bedroom	link	irokrhus
Scene	Kitchen	link	MimingApe	Scene	Bedroom	link	oldtimer
Scene	Kitchen	link	oldtimer	Scene	Bedroom	link	Yulia
Scene	Kitchen	link	appisolato	Scene	Bedroom	link	Mikel007
Scene	Office	link	ThePefDispenser	Assets	N/A	link	ruwo
Scene	Office	link	LRosario	Assets	N/A	link	Davilion
Scene	Office	link	DragonautX	Assets	N/A	link	MZiemys
Scene	Office	link	fjcar	Assets	N/A	link	MZiemys
Scene	Office	link	Elysia	Assets	N/A	link	Davilion
Scene	Bathroom	link	bobal57	Assets	N/A	link	Zorian
Scene	Bathroom	link	irokrhus	Assets	N/A	link	vicentecarro
Scene	Bathroom	link	wfg5001	Assets	N/A	link	piergi
Scene	Bathroom	link	nacimus	Assets	N/A	link	Bastable
Scene	Bathroom	link	Ndakasha	Assets	N/A	link	arttechsouth

Table 2: Blender assets and scenes.

For multi-layer RAFT is trained on four RTX 3090 GPUs with a batch size of 4. The learning rate is set to 1e-4. When the training images contain $m \geq 1$ layers of true optical flow and the model generates $n > m$ optical flow prediction layers, the final prediction layer is duplicated $n - m + 1$ times to align the dimensions. Specifically, for training with the Sintel [5], which provides a single layer of optical flow ground truth, this duplication occurs n times.

For (S+L) training policy, any image in Sintel dataset will appear 100 times to match the size of our synthetic dataset. The reported results from the checkpoint that has the best performance on validation set of our benchmark.

4 Additional Experiments

4.1 First Layer Optical Flow

We provide additional evaluation results for our single layer experiments. For this set of experiments, we evaluate each method to predict first layer optical flow on pairs of images that have been downsampled by a factor of 8—this is as opposed to our results in the main paper, where we evaluate each method on images downsampled by a factor of 4. Overall, our finetuned RAFT method still outperforms other existing optical flow methods, including the baseline RAFT method. Results are shown in Tab. 3.

Method	All				Transparent				Reflective				Diffuse			
	EPE↓	1px↓	3px↓	5px↓	EPE↓	1px↓	3px↓	5px↓	EPE↓	1px↓	3px↓	5px↓	EPE↓	1px↓	3px↓	5px↓
FlowNet-C [9]	9.71	89.07	61.51	43.93	11.08	89.23	62.43	45.05	6.38	88.25	58.36	40.03	8.53	89.08	54.13	35.35
FlowNet2 [11]	10.07	77.56	54.22	42.13	11.46	78.20	56.15	44.38	6.70	75.39	46.69	33.33	7.66	72.44	43.21	29.41
PWC-Net [19]	9.49	74.93	50.47	39.05	10.90	76.47	52.59	41.43	5.99	69.84	42.99	29.82	6.91	61.85	34.77	25.30
GMA [12]	9.77	72.46	46.93	36.97	12.01	75.48	50.07	40.24	**4.48**	60.85	35.42	24.26	2.26	54.20	25.56	17.98
SKFlow [20]	9.86	72.02	47.44	36.88	12.00	74.90	50.84	40.14	4.78	60.89	35.21	24.40	3.23	54.90	23.23	17.18
CRAFT [18]	10.36	72.34	47.54	37.00	12.65	74.75	50.96	40.47	4.65	64.12	35.10	23.06	3.30	53.08	23.95	18.89
GMFlow [22]	9.09	81.99	51.79	37.75	10.93	83.01	53.87	40.06	5.20	80.02	44.73	28.64	5.01	66.91	35.13	24.79
GMFlow+ [23]	9.46	82.71	53.14	39.70	11.31	83.21	54.91	42.10	6.04	81.61	46.57	29.95	5.71	75.97	41.43	27.85
FlowFormer [10]	10.20	73.59	48.97	38.56	12.51	76.91	52.56	42.27	5.00	61.03	36.18	24.76	**2.17**	52.89	22.90	14.12
RAFT [21]	9.38	71.98	46.46	36.15	11.31	74.65	49.69	39.34	5.57	61.53	35.73	24.57	2.62	56.72	**19.44**	**14.05**
RAFT-ft. (S)	9.74	74.56	49.14	38.60	11.64	77.10	51.94	41.55	5.74	65.16	39.28	27.47	4.13	57.63	28.32	20.06
RAFT-ft. (L)	**7.12**	**69.17**	**40.88**	**29.49**	**8.26**	**71.74**	**43.44**	**31.79**	5.24	59.72	**31.23**	**20.60**	3.03	**51.77**	24.61	15.84
RAFT-ft. (S+L)	7.93	69.20	42.04	32.51	9.23	71.88	44.76	35.05	6.16	**58.68**	32.42	22.55	2.65	54.13	22.02	18.27

Table 3: Representative optical flow methods evaluated on first layer subset of our benchmark using EPE and bad-τ metrics. Images are down-sampled by 8. Best scores are in **bold**. Underlined numbers denote RAFT fine-tuned on our synthetic data outperforming the original version.

4.2 First-Layer Stereo Matching

As mentioned in main paper, our benchmark also provides stereo matching ground-truth. We evaluate effectiveness of existing representative stereo matching methods with public implementation and pre-trained weights on LayeredFlow's first layer points. Stereo pairs with significant y-axis discrepancies are excluded, achieving an average residual y-disparity of 0.36 on images downsampled by a factor of 4 to 540×960. Results are shown in Tab. 4.

As stereo matching methods tend to be sensitive to the scale of images, we also provide results on images that have been downsampled by a factor of 8, shown in Tab. 5. Overall, existing methods generally struggle to achieve good EPE and bad-τ metrics, particularly for transparent and reflective materials. This highlights the challenge of first-layer stereo matching in non-Lambertian settings.

Method	All				Transparent				Reflective				Diffuse			
	EPE↓	1px↓	3px↓	5px↓	EPE↓	1px↓	3px↓	5px↓	EPE↓	1px↓	3px↓	5px↓	EPE↓	1px↓	3px↓	5px↓
PSMNet [6]	74.43	92.16	83.32	77.78	82.92	95.57	89.68	84.78	45.41	82.62	62.38	53.70	17.82	59.91	37.45	31.62
HSMNet [24]	57.38	93.91	87.43	83.03	64.48	98.70	95.19	91.71	33.74	81.14	63.72	54.86	7.40	45.96	23.08	18.49
LEAStereo [7]	54.96	89.58	79.20	74.36	62.19	95.66	87.51	82.96	29.56	70.47	50.54	43.96	9.66	41.80	24.86	21.45
CFNet [16]	40.23	90.45	83.09	77.79	45.90	95.40	90.27	85.61	20.07	75.16	59.53	51.75	5.68	50.27	30.85	22.49
PCWNet [17]	41.59	92.00	84.23	79.58	47.44	97.55	92.81	88.37	20.90	75.80	55.93	50.01	5.68	42.77	22.53	18.92
RAFTStereo [13]	32.50	85.27	75.80	71.25	37.36	92.99	85.04	80.22	13.94	60.15	42.45	38.61	8.55	28.01	21.83	20.01
DLNR [25]	**30.69**	**82.47**	**71.92**	**67.24**	**36.10**	**90.48**	**81.64**	**76.57**	**9.96**	**56.44**	**37.02**	**33.87**	**4.49**	**23.12**	**14.49**	**11.58**

Table 4: Representative stereo matching methods evaluated on first layer subset of our benchmark using EPE and bad-τ metrics. Best scores are in **bold**.

Method	All				Transparent				Reflective				Diffuse			
	EPE↓	1px↓	3px↓	5px↓	EPE↓	1px↓	3px↓	5px↓	EPE↓	1px↓	3px↓	5px↓	EPE↓	1px↓	3px↓	5px↓
PSMNet [6]	**14.73**	84.77	**63.11**	**51.82**	**16.44**	**88.89**	**69.15**	**57.37**	8.11	72.52	41.74	31.42	4.77	44.55	19.38	15.16
HSMNet [24]	18.70	89.83	76.93	66.72	20.83	94.79	84.36	74.55	11.25	75.62	51.69	37.92	3.04	39.43	18.45	14.75
LEAStereo [7]	15.51	82.92	64.61	54.40	17.77	89.44	71.92	61.46	6.71	61.42	38.07	27.04	2.81	28.83	14.64	13.59
CFNet [16]	16.38	84.81	70.06	60.20	18.39	91.07	76.72	66.67	9.26	65.46	47.98	37.85	1.98	27.18	15.38	10.75
PCWNet [17]	17.56	89.44	76.35	66.52	19.86	96.20	84.27	74.42	9.14	69.14	49.40	38.17	2.21	24.57	14.38	11.13
RAFTStereo [13]	16.72	84.14	70.34	59.71	19.05	91.48	78.62	67.21	7.75	59.88	40.47	31.44	2.82	23.45	13.03	12.99
DLNR [25]	15.91	**82.06**	68.79	60.69	18.70	91.15	78.28	70.24	**4.80**	**49.88**	**33.58**	**24.25**	**0.92**	**16.50**	**7.21**	**3.10**

Table 5: Representative stereo matching methods evaluated on first layer subset of our benchmark using EPE and bad-τ metrics. Images are down-sampled by 8. Best scores are in **bold**.

References

1. Blendswap. https://www.blendswap.com 2
2. Hdri haven. https://hdri-haven.com 2
3. Lib-apriltag. https://github.com/duckietown/lib-dt-apriltags 1
4. Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000) 1
5. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: A. Fitzgibbon et al. (Eds.) (ed.) European Conf. on Computer Vision (ECCV). pp. 611–625. Part IV, LNCS 7577, Springer-Verlag (Oct 2012) 3
6. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5410–5418 (2018) 4, 5
7. Cheng, X., Zhong, Y., Harandi, M., Dai, Y., Chang, X., Li, H., Drummond, T., Ge, Z.: Hierarchical neural architecture search for deep stereo matching. Advances in Neural Information Processing Systems **33**, 22158–22169 (2020) 4, 5
8. Community, B.O.: Blender - a 3D modelling and rendering package. Blender Foundation, Stichting Blender Foundation, Amsterdam (2018), http://www.blender.org 2
9. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2758–2766 (2015) 4
10. Huang, Z., Shi, X., Zhang, C., Wang, Q., Cheung, K.C., Qin, H., Dai, J., Li, H.: Flowformer: A transformer architecture for optical flow. In: European Conference on Computer Vision. pp. 668–685. Springer (2022) 4
11. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2462–2470 (2017) 4
12. Jiang, S., Campbell, D., Lu, Y., Li, H., Hartley, R.: Learning to estimate hidden motions with global motion aggregation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9772–9781 (2021) 4
13. Lipson, L., Teed, Z., Deng, J.: Raft-stereo: Multilevel recurrent field transforms for stereo matching. In: 2021 International Conference on 3D Vision (3DV). pp. 218–227. IEEE (2021) 4, 5
14. Olson, E.: AprilTag: A robust and flexible visual fiducial system. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). pp. 3400–3407. IEEE (May 2011) 1

15. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf 2
16. Shen, Z., Dai, Y., Rao, Z.: Cfnet: Cascade and fused cost volume for robust stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13906–13915 (2021) 4, 5
17. Shen, Z., Dai, Y., Song, X., Rao, Z., Zhou, D., Zhang, L.: Pcw-net: Pyramid combination and warping cost volume for stereo matching. In: European conference on computer vision. pp. 280–297. Springer (2022) 4, 5
18. Sui, X., Li, S., Geng, X., Wu, Y., Xu, X., Liu, Y., Goh, R., Zhu, H.: Craft: Cross-attentional flow transformer for robust optical flow. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 17602–17611 (2022) 4
19. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8934–8943 (2018) 4
20. Sun, S., Chen, Y., Zhu, Y., Guo, G., Li, G.: Skflow: Learning optical flow with super kernels. Advances in Neural Information Processing Systems **35**, 11313–11326 (2022) 4
21. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European Conference on Computer Vision (ECCV) (2020) 2, 4
22. Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Tao, D.: Gmflow: Learning optical flow via global matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8121–8130 (2022) 4
23. Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Yu, F., Tao, D., Geiger, A.: Unifying flow, stereo and depth estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 4
24. Yang, G., Manela, J., Happold, M., Ramanan, D.: Hierarchical deep stereo matching on high-resolution images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5515–5524 (2019) 4, 5
25. Zhao, H., Zhou, H., Zhang, Y., Chen, J., Yang, Y., Zhao, Y.: High-frequency stereo matching network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1327–1336 (2023) 4, 5