Learning Large Motion Estimation from Intermediate Representations with a High-Resolution Optical Flow Dataset Featuring Long-Range Dynamic Motion

Supplementary Material

Our supplementary material provides additional information about our proposed method and includes detailed discussions on the following contents that were not extensively covered in the main paper:

- Implementation details (Sec. A).
- Comparison with existing datasets and benchmarks for optical flow (Sec. B).
- Comparison with local cost volume (Sec. C)
- Additional analysis of Sec. 6 of the main paper (Sec. D).
- Design of distance loss (Sec. E)
- Additional qualitative results (Sec. F).
- Supplementary details of RelayFlow-4K (Sec. G).
- Hyper-parameter analysis (Sec. H).
- Limitation and future work (Sec. I).

A. Implementation details

Our implementation builds upon RAFT [17] and GMA [7], adopting their main configurations to ensure a consistent network structure, except for the downscaling factor. Previous methods [7, 16, 17] commonly employ a 1/8 downscaling for correlation computations, resulting in complex inference when estimating flow in 4K images. Consequently, we modified the downscaling factor from 1/8 to 1/16. With 1/8 downscaling and a batch size of 1, memory consumption surpasses 60 GB, rendering it infeasible for standard GPUs. In contrast, 1/16 downscaling enables 4K resolution inference with less than 20 GB of memory, effectively overcoming this limitation. We set 100K training iterations per stage and crop images to 1920×1088 for training. We train our model with the batch size 1. The inital learning rate is $2e^{-4}$ and, and it is halved at the beginning of each stage before resuming training. Following prior works [7, 17], we use $\beta = 0.8$ and $\gamma = 0.85$ for $\mathcal{L}_{dis}$, and a distance loss threshold ζ_D of 368 pixels with $\alpha = 0.3$. We set the stage K from 0 to 4, and adjusted the refinement iteration M using 12 during training and 24 during evaluation. Other components, such as the optimizer and related settings, also follow the training configuration of prior works.

B. Comparison with existing datasets and benchmarks for optical flow

We compare recent optical flow datasets in Tab. B, providing an overview of the available images, ground truths, and additional data characteristics such as stereo disparity, intermediate flow, and distance maps

Table A. Generalization performance comparison with local cost volume.

Train Data	Scale	Method	Sintel (train)		KITTI-15 (train)		RelayFlow (test)		
			Clean	Final	F1-epe	F1-all	All	Match	Unmatch
C+T	1/8	RAFT	**1.43**	2.71	**5.04**	**17.4**	OOM	OOM	OOM
		RAFT-Local	1.46	**2.68**	**5.04**	17.5	34.93	27.03	210.64
	1/16	RAFT	1.86	3.16	7.52	26.5	**16.99**	**8.32**	**210.00**

C. Comparison with local cost volume

As mentioned in Tab. 3 of the main paper, RAFT encountered OOM at 1/8 resolution on the RelayFlow test set, due to the high-resolution images. To mitigate this, we employed a local cost volume [17] formulation to reduce memory usage and enable evaluation. As shown in Tab. A, the result with the local cost volume avoided OOM but performed poorly on large motions. It even performs worse than 1/16 RAFT model on the RelayFlow test set.

D. More about comparison with approaches using accumulation and the intermediate frames

We continue the description of the comparison with the accumulation methods from Sec. 6. We demonstrate more about our experiment settings and utilized datasets.

D.1. HS-Sintel dataset

As mentioned in previous work [18], for HS-Sintel [6], GT flow is only provided at 1008 FPS and is not available for other frame rates. Therefore, following the standard protocol [18], we utilize the 24 FPS GT flows from MPI Sintel [2] as ground truth labels to assess the predictions derived from 1008 FPS video sequences of HS-Sintel.

We present a comparison with other methods on HS-Sintel in the 1st column of Tab. 4 of the main paper. Existing methods [9, 18] employ a multi-frame approach, utilizing intermediate frames not only during training but also at inference time. Notably, our method achieves the best performance in non-occlusion regions and the second-highest overall performance. This indicates that our proposed methods are effective not only on our dataset, RelayFlow-4K, but also on others, particularly excelling in matching regions as discussed in the main paper.

D.2. CVO dataset

We also conduct experiments on the CVO dataset [18], which has a relatively small resolution of 512×512. As mentioned in Sec. 3.2 of the main paper, the CVO dataset provides optical flow annotations for intermediate frames,

Table B. Overview of optical flow datasets. We present available images and ground truths for optical flow (OF). We check whether each dataset provides stereo disparity (ST), intermediate flow (inter. flow), and a match map (match.). † : partially exist.

Dataset	Venue	OF	ST	#images	#gt frames	#pix	inter. flow	match.	scenes	source	ph.realism	motion
RelayFlow-4K (Ours)	-	✓	✓	8428	37900	8.3M	✓	✓	35	CGI	high	realistic
CVO [18]	ICCV'23	✓	✗	83594	262724	0.3M	✗	✓	11942	CGI	low	random
Spring [11]	CVPR'23	✓	✓	5953	23812	2.1M	✗	✓	47	CGI	high	realistic
AutoFlow [15]	CVPR'21	✓	✗	40000	40000	0.3M	✗	✗	n/a	CGI	low	random
HumanOF [13]	IJCV'20	✓	✗	238900	238900	0.4M	✗	(✓)†	18432	CGI	med.	rand./human
VKITTI2 [3]	arXiv'20	✓	✓	21210	84840	0.5M	✗	✓	5	CGI	med.	automative
Vimeo-90K [19]	IJCV'19	✓	✗	89800	89800	0.3M	✗	✗	4278	mixed	low	random
HS Sintel [6]	CVPR'17	✓	✗	4730	4704	1.8M	✓	✓	13	CGI	high	realistic
VIPER [14]	ICCV'17	✓	✗	186285	372570	2.1M	✗	✗	184	CGI	high	automotive
Driving [10]	CVPR'16	✓	✓	4392	17568	0.5M	✗	✓	1	CGI	med.	automotive
FlyingThings3D [10]	CVPR'16	✓	✓	24084	96336	0.5M	✗	✓	2676	CGI	low	random
HD1K [8]	CVPRW'16	✓	✗	1074	1074	2.8M	✗	✓	63	real	high	automotive
Monkaa [10]	CVPR'16	✓	✓	8640	34560	0.5M	✗	✓	8	CGI	low	random
FlyingChairs [4]	ICCV'15	✓	✗	22872	22872	0.2M	✗	✗	n/a	CGI	low	random
KITTI 2015 [12]	CVPR'15	✓	✓	400	400	0.5M	✗	✓	n/a	real	high	automotive
KITTI 2012 [5]	CVPR'12	✓	✓	389	389	0.5M	✗	✓	n/a	real	high	automotive
MPI Sintel [2]	ECCV'12	✓	✓	1593	1593	0.4M	✗	✓	35	CGI	high	realistic
Middlebury-OF [1]	IJCV'11	✓	✗	16	16	0.2M	✗	✗	16	HT/CGI	med.	small

similar to our work and can therefore be used to train our framework. We follow the training settings of AccFlow, the method accompanying the CVO dataset, to implement our proposed flow estimation strategy while incorporating supplementary techniques, such as matching cost distillation and incremental time-step learning.

We present our results with others on the CVO dataset in the 2nd and 3rd columns of Tab. 4 of the main paper, where our method achieves the second-best performance overall. Notably, the CVO dataset features random object movements, and our approach demonstrates higher flow estimation accuracy by utilizing only the reference and target frames during inference, without relying on intermediate frames. When comparing the base model (*e.g.*, GMA and RAFT) with *Relay* (Ours), our approach demonstrates significant improvements in non-occluded regions. This consistent trend across all datasets highlights the generalization ability of our method. However, it remains relatively weak in occluded regions. A key advantage of our approach is that it uses intermediate frames only during training. In other words, it employs the same test setup as the baseline, offering a significant benefit over other methods.

D.3. Computational costs

We further compare our strategy with other optical flow estimation methods in terms of runtime and memory usage during the inference phase, using the RAFT model [17] as a baseline. For AccFlow [18], we applied the setting from the original paper, using 5 intermediate frames for a total of 7 frames. In contrast, both our method and the baseline use only 2 frames: the reference and target frames.

Table. 5 of the main paper presents the results of runtime and memory usage measured across various image sizes. We measure the computational cost using an NVIDIA A6000 GPU, and for memory usage, we record values after a sufficient warm-up process rather than the peak memory. Our proposed method requires only the reference and target frames during the inference phase, identical to the baseline, and does not require any additional modules, resulting in the same runtime and memory usage as the baseline. On the other hand, AccFlow requires intermediate frames during the inference process, just as it does during its training phase. Consequently, this results in significantly higher runtime requirements during inference. Furthermore, AccFlow relies on additional modules to accumulate optical flow between adjacent frames, which increases the demand for memory during inference. As a result, when performing flow estimation between high-resolution 4K images, out-of-memory (OOM) issues due to the peak memory usage. While the approach of accumulating intermediate frames during the inference phase achieves strong performance, its practicality decreases as image resolution increases due to a sharp rise in inference time and memory usage. In contrast, our method proves to be well-suited for the current trend toward high-resolution data.

E. Design of distance loss

Our motivation for down-weighting the loss based on the distance from matched regions in occluded areas stems from that these regions lack sufficient visual cues, making predictions noisy and leading to high, unstable gradients. Prioritizing regions with reliable correspondences helps the model learn stable patterns, while unmatched areas are refined via neighboring propagation. As shown in Tab. C, alternatives like removing occluded regions or up-weighting them by distance lead to reduced performance.

Table C. Ablation of design of distance loss

Loss Ablation Study	All		Match		Unmatch	
	EPE	1px	EPE	1px	EPE	1px
RAFT	18.54	25.90	8.31	23.24	223.82	84.91
+ Removing Occlusion	18.23	17.90	8.43	14.95	236.01	83.73
+ Matched-region Inverse Distance Loss	18.40	20.40	9.47	18.17	217.02	**76.24**
+ Matched-region Distance Loss (**Ours**)	**15.36**	**17.71**	**6.89**	**14.79**	**203.73**	82.58

Table D. Hyper-parameter analysis of α in Eq. (8) with RAFT [17].

α	All		Match		Unmatch	
	EPE	1px	EPE	1px	EPE	1px
0.1	14.73	15.20	6.04	12.34	207.8	78.91
0.3	11.69	15.52	4.08	12.64	181.01	79.58
0.5	12.99	15.67	5.09	12.79	188.54	79.54
0.7	12.92	16.06	5.18	13.21	185.01	79.36
0.9	13.02	15.56	5.22	12.69	186.46	79.39

F. Additional qualitative results

We show more qualitative comparisons for optical flow to showcase the effectiveness of our method in Fig. A. Our method qualitatively improves the performance of existing models, regardless of the base model used. Notably, it estimates flows for large displacements previously unattainable and accurately captures object boundaries, enabling fine optical flow estimation.

G. Details of RelayFlow-4K dataset

We visualize more samples from our RelayFlow-4K dataset in Fig. B, C, D, and E. We present the current image and its subsequent frame, along with annotations that include the optical flow and depth map. For training efficiency, we also provide additional data such as a match map and a distance map. Although not explicitly shown in the sample figures, each frame is accompanied by an additional frame linked through a stereo setting (right frame), and optical flow between the right frames is also provided.

We provide comprehensive scene configurations and the corresponding baselines used for each scene through configuration files included with the dataset. These files include all the necessary details required for further training and analysis. Additionally, the annotations in our dataset, such as optical flow, depth, and distance maps, are refined and standardized into a consistent discrete range prior to distribution. For the match map, unlike traditional occlusion maps, we precisely identify occlusion regions while improving the pixel matching process through an additional masking step. This involves masking pixels with significant RGB differences that exceed a predefined threshold, ensuring accurate and reliable construction of the match map for pixel-level correspondence.

H. Hyper-parameter analysis

As shown in Tab. D, we conduct a hyper-parameter analysis of the weight α for the matching cost distillation loss in the main paper. To this end, we perform a hyper-parameter analysis by applying only the distance and matching cost distillation losses without using the incremental time-step learning strategy. Specifically, we train at stage 0 and then directly skip to stage 4, applying the matching cost distillation loss and distance loss.

We reveal the impact of the trade-off between EPE and 1px across all match and unmatch points. Across all categories, the lowest EPE is achieved when $\alpha = 0.3$. Although the best 1px accuracy is observed at $\alpha = 0.1$, this setting resulted in significantly higher EPE values. From the perspective of 1px accuracy, the next best option is $\alpha = 0.3$. Considering the trade-off between EPE and 1px accuracy, $\alpha = 0.3$ was determined to be the most suitable value for our experiments and the application of matching cost distillation loss and was thus selected as the optimal setting. Apart from these detailed performance metrics, the proposed matching cost distillation demonstrates robustness, showing minimal performance variation with different loss weights, α.

I. Limitation and future work

As shown in Tab. 2 of the main paper, in regions unmatched with small displacements (s0-40), a slight performance decline compared to the original model is observed. We believe that this performance drop may stem from our proposed method of prioritizing matched regions and large displacements, particularly in the KD loss. However, this trade-off is considered marginal compared torall performance gains. Addressing this learning bias represents a limitation of our work and an avenue for future research.

References

[1] Simon Baker, Daniel Scharstein, James P Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International journal of computer vision*, 92:1–31, 2011. 2

[2] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 611–625. Springer, 2012. 1, 2

[3] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 2

[4] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Pro-*

ceedings of the IEEE international conference on computer vision, pages 2758–2766, 2015. 2

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 2

[6] Joel Janai, Fatma Guney, Jonas Wulff, Michael J Black, and Andreas Geiger. Slow flow: Exploiting high-speed cameras for accurate and diverse optical flow reference data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3597–3607, 2017. 1, 2

[7] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9772–9781, 2021. 1

[8] Daniel Kondermann, Rahul Nair, Katrin Honauer, Karsten Krispin, Jonas Andrulis, Alexander Brock, Burkhard Gussefeld, Mohsen Rahimimoghaddam, Sabine Hofmann, Claus Brenner, et al. The hci benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 19–28, 2016. 2

[9] SukHwan Lim, John G Apostolopoulos, and AE Gamal. Optical flow estimation using temporally oversampled video. *IEEE Transactions on Image Processing*, 14(8):1074–1087, 2005. 1

[10] 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. 2

[11] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4981–4991, 2023. 2

[12] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015. 2

[13] Anurag Ranjan, David T Hoffmann, Dimitrios Tzionas, Siyu Tang, Javier Romero, and Michael J Black. Learning multi-human optical flow. *International Journal of Computer Vision*, 128:873–890, 2020. 2

[14] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *Proceedings of the IEEE international conference on computer vision*, pages 2213–2222, 2017. 2

[15] Deqing Sun, Daniel Vlasic, Charles Herrmann, Varun Jampani, Michael Krainin, Huiwen Chang, Ramin Zabih, William T Freeman, and Ce Liu. Autoflow: Learning a better training set for optical flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10093–10102, 2021. 2

[16] Shangkun Sun, Yuanqi Chen, Yu Zhu, Guodong Guo, and Ge Li. Skflow: Learning optical flow with super kernels. *Advances in Neural Information Processing Systems*, 35: 11313–11326, 2022. 1

[17] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 1, 2, 3

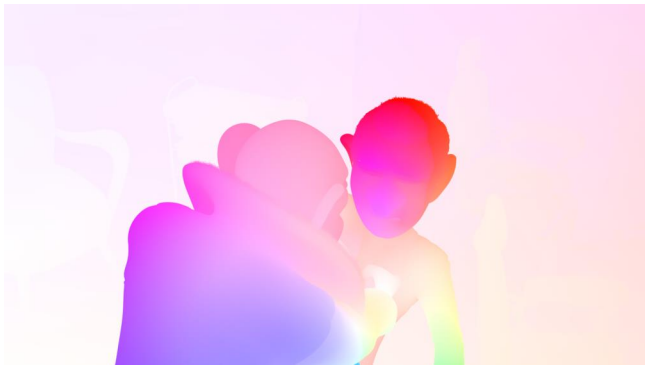
[18] Guangyang Wu, Xiaohong Liu, Kunming Luo, Xi Liu, Qingqing Zheng, Shuaicheng Liu, Xinyang Jiang, Guangtao Zhai, and Wenyi Wang. Accflow: Backward accumulation for long-range optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12119–12128, 2023. 1, 2

[19] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision (IJCV)*, 127 (8):1106–1125, 2019. 2

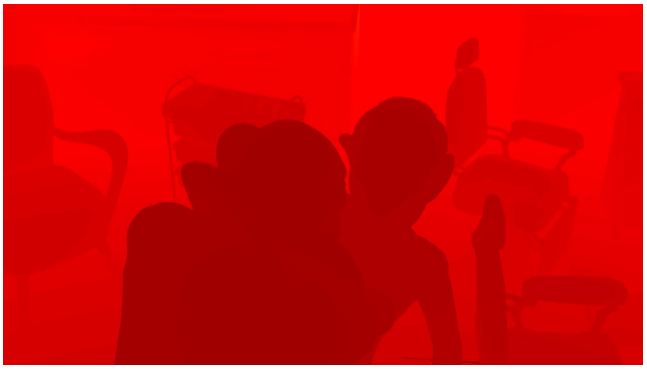
Figure A. Qualitative comparison on the test set of RelayFlow-4K. We present three examples with varying motion ranges. 'Stat.' represents the distribution of optical flow in each sample.

(a) Present image

(b) Next image

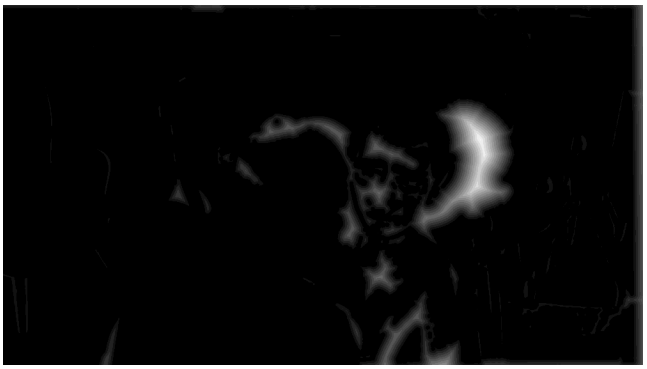


(c) Flow



(d) Depth map

(e) Match map



(f) Distance map

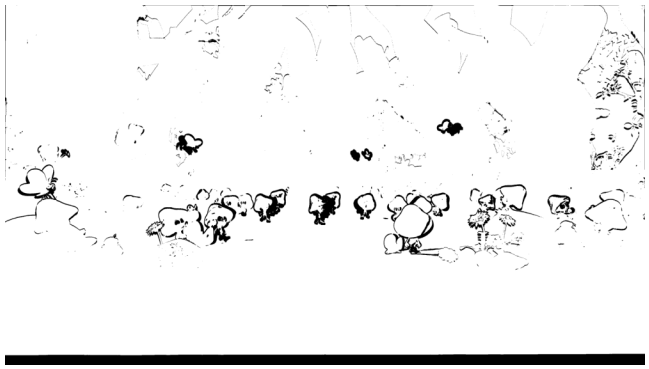
Figure B. Sample data of RelayFlow-4K. All sample annotations and maps here is aligned with the present image.

(a) Present image

(b) Next image

(c) Flow

(d) Depth map



(e) Match map

(f) Distance map

Figure C. Sample data of RelayFlow-4K. All sample annotations and maps here is aligned with the present image.

(a) Present image

(b) Next image

(c) Flow

(d) Depth map

(e) Match map

(f) Distance map

Figure D. Sample data of RelayFlow-4K. All sample annotations and maps here is aligned with the present image.

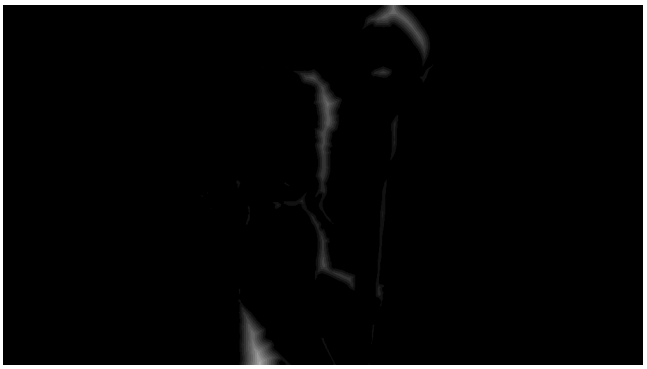
(a) Present image

(b) Next image

(c) Flow

(d) Depth map

(e) Match map



(f) Distance map

Figure E. Sample data of RelayFlow-4K. All sample annotations and maps here is aligned with the present image.