

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

Hoonhee Cho*, Yuhwan Jeong*, and Kuk-Jin Yoon

Visual Intelligence Lab., KAIST, Korea

{gnsngsngml, jeongyh98, kjyoon}@kaist.ac.kr

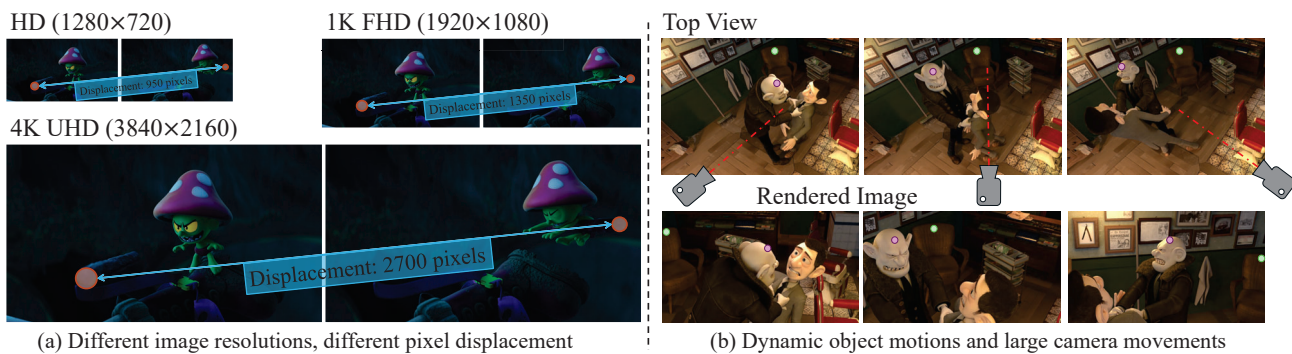


Figure 1. **Motivation for Long-Range Optical Flow Research.** (a) As image resolution increases, the pixel displacement within the same scene grows proportionally. With the increasing availability of high-resolution data, scenarios with large pixel displacements have become more common, posing challenges for accurate optical flow estimation. (b) In addition to object motion, dynamic camera movements, particularly rotational motion, can lead to significant pixel displacements between frames. This issue arises even in consecutive frames when the motion is highly dynamic relative to the frame rate, making optical flow estimation more challenging.

Abstract

With advancements in sensor and display technologies, high-resolution imagery is becoming increasingly prevalent in diverse applications. As a result, optical flow estimation needs to adapt to larger image resolutions, where even moderate movements lead to substantial pixel displacements, making long-range motion estimation more critical than ever. However, existing datasets primarily focus on short-range flow in low-resolution settings, limiting the generalization of models to high-resolution scenarios with large displacements. Additionally, there is a lack of suitable datasets for evaluating model capacity in long-range motion estimation, further hindering progress in this area. To address this, we introduce *RelayFlow-4K*, high-resolution 4K optical flow dataset designed to capture diverse motion patterns, including long-range intermediate frame flows. While such datasets provide valuable training resources, long-range estimation remains challenging due to increased matching ambiguity. Simply incorporating these datasets does not inherently improve performance. To this end, we propose a novel training frame-

work that integrates matching cost distillation and incremental time-step learning to refine cost volume estimation and stabilize training. Additionally, we leverage the distance map, which measures the distance from unmatched regions to their nearest matched pixels, improving occlusion handling. Our approach significantly enhances long-range optical flow estimation in high-resolution settings. Our datasets and code are available at <https://github.com/Chohoonhee/RelayFlow-4K>.

1. Introduction

Optical flow represents a dense motion field that maps pixel correspondences between consecutive frames. While recent advancements [7, 15, 72] have significantly improved performance, challenges remain, particularly in high-resolution and long-range motion scenarios. As sensor and display technologies advance, high-resolution images are increasingly common, amplifying pixel displacements even for moderate movements and making long-range optical flow estimation critical. For example, when capturing a dynamically moving object with a handheld ego camera, substantial pixel shifts occur due to camera and ob-

*Denotes equal contribution.

Table 1. Overview of optical flow datasets with available images, ground truths for optical flow (OF), stereo disparity (ST), intermediate flow (inter. flow), and a match map (match.). † : partially exist. More comparisons are provided in the *Supple.*

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 [70]	ICCV'23	✓	✗	83594	262724	0.3M	✗	✓	11942	CGI	low	random
Spring [46]	CVPR'23	✓	✓	5953	23812	2.1M	✗	✓	47	CGI	high	realistic
HS Sintel [24]	CVPR'17	✓	✗	4730	4704	1.8M	✓	✓	13	CGI	high	realistic
FlyingThings3D [45]	CVPR'16	✓	✓	24084	96336	0.5M	✗	✓	2676	CGI	low	random
HD1K [32]	CVPRW'16	✓	✗	1074	1074	2.8M	✗	✓	63	real	high	automotive
FlyingChairs [8]	ICCV'15	✓	✗	22872	22872	0.2M	✗	✗	n/a	CGI	low	random
KITTI 2015 [48]	CVPR'15	✓	✓	400	400	0.5M	✗	✓	n/a	real	high	automotive
MPI Sintel [2]	ECCV'12	✓	✓	1593	1593	0.4M	✗	✓	35	CGI	high	realistic

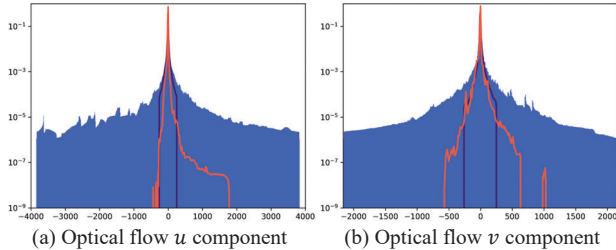


Figure 2. Comparison of flow statistics between RelayFlow-4K (blue), CVO [70] (purple), and Spring [46] (orange).

ject movements, necessitating long-range motion estimation. Similarly, in video frame interpolation [5, 9, 13, 16, 27, 30, 59], accurate long-range optical flow [57] is essential to synthesize smooth intermediate frames and prevent ghosting artifacts, especially in high-resolution content. Specifically, as shown in Fig. 1 (a), increasing image resolution leads to larger pixel distances even when capturing the same scene. Additionally, Fig. 1 (b) illustrates that significant camera motion, particularly rotation, induces long-range optical flow. In such cases, the absence of intermediate frames or motion exceeding the frame rate often requires inferring long-range flow from only the two available frames during inference. A key challenge in high-resolution long-range optical flow research is the lack of suitable datasets for training and evaluation. While other low-level vision tasks [35, 57] have long incorporated 4K and higher-resolution data, most existing optical flow studies, as summarized in Tab. 1, remain confined to resolutions below 1K. Furthermore, as shown in Fig. 2, conventional datasets primarily contain flow displacements within 500 pixels, whereas the shift toward higher-resolution imagery increasingly demands accurate flow estimation beyond this range [35, 57]. This underscores the need for datasets that accommodate the growing requirements of high-resolution long-range optical flow research.

To address this, we introduce RelayFlow-4K, a new high-resolution optical flow dataset providing a wide range of pixel displacement and dynamic motion. In addition, it offers stereo images with disparity data for stereo matching. RelayFlow-4K dataset also provides a match map and a distance map, which measures the distance from unmatched regions to their nearest matched pixels. This feature pro-

vides stable training by effectively handling occlusions.

Even with high-resolution and long-range datasets, long-range flow estimation remains challenging. Recent high-performing approaches [26, 61, 64] typically generate a cost volume and refine it iteratively, which plays a crucial role in determining the initial flow direction and overall performance. A major challenge is the increased number of matching candidates compared to local-range flow, causing greater ambiguity in identifying correct matches. Therefore, simply incorporating long-range datasets into training does not inherently lead to performance improvements. Additionally, as resolution increases, the cost volume grows, leading to excessive memory use. To mitigate this, more feature downsampling is often required, which reduces the ability to capture fine details. Effective guidance during training is therefore essential for stable learning and more accurate flow direction discovery. Therefore, we propose an effective approach using knowledge distillation [11, 55] to utilize information from intermediate frames only during the training process. By merging the cost volumes generated from the intermediate frames, we generate an aggregated matching cost that is relatively clear and accurate. Moreover, we adapt the curriculum learning strategy [1, 58] for entire learning process, named incremental time-step learning. With this learning strategy, the model is gradually trained, beginning with frames exhibiting small displacement distributions and progressing to frames with larger displacements, ultimately resulting in robustness across all distributions. Additionally, our incremental time-step learning approach complements our proposed cost volume distillation scheme exceptionally well, enhancing its effectiveness. Finally, we leverage the distance map generated from the match map, which measures the distance from unmatched regions to their nearest matched pixels, enabling more effective training by improving occlusion handling.

2. Related works

Optical flow. Numerous approaches [8, 17–19, 34, 51, 62, 73, 79, 86] have been introduced to enhance the accuracy of optical flow estimation using convolution layers. However, these approaches struggled with coarse resolution errors and missed small, fast-moving objects. To address these,

RAFT [64] introduced a novel solution with a 4D correlation volume and iterative updates. Building on RAFT, new elements such as global motion aggregation [26], multi-scale network [21, 23], super kernels [63], and on-going researches [7, 15, 22, 29, 43, 50, 61, 81, 83] have been introduced, advancing state-of-the-art performance in optical flow estimation. In addition, there are several unsupervised optical flow estimation methods [3, 14, 20, 28, 33, 39–42, 44, 47, 52, 53, 60, 67, 68, 75, 76, 82, 84, 85] to break down the fidelity of the ground truth data. Beyond two-frame estimation, multi-frame methods leverage temporal context [54, 56, 65]. [6] uses a memory buffer for motion features, and [56] propagates auxiliary cues across frames. However, both approaches overlook long-range motion.

Optical flow for large displacements. One approach to addressing long-range motion involves leveraging an accumulation strategy derived from high frame rate optical flow [37]. In line with this approach, SlowFlow [24] developed a synthetic dataset with an extremely high frame rate, enabling the estimation of optical flow across frames. The recent works attempt to solve large displacement connection by utilizing intermediate frames with per-pixel tracking [4, 10, 31], recursively backward accumulation [70], and a memory bank [6]. However, depending on the situation, intermediate frames may need to be generated, or in cases such as timelapse videos or rapidly moving motion exceeds the frame rate, intermediate frames with smaller pixel movements may be absent. Our work focuses on such challenging scenarios to overcome these difficulties.

Knowledge distillation (KD) [11, 55, 66] is widely used in low-level vision to reduce computational costs while maintaining performance. KD has notably improved super-resolution [71, 78] and depth estimations [66, 74]. In flow estimation, KD helps create compact models and addresses occlusions and ambiguities by filling in unknown regions [12, 25, 40, 60]. However, its potential to transfer knowledge from simpler to more complex tasks remains underexplored, highlighting an area for further research.

Curriculum learning structures training by moving from simpler to more complex, enhancing model robustness and efficiency. Bengio *et al.*, [1] showed that curriculum strategy improves generalization by allowing models to build foundational knowledge first. Recent works [36, 49, 58, 69, 77] extend this framework to deep learning applications across domains like object detection and language processing, often integrating KD to guide complex tasks through foundational knowledge from simpler pre-trained models.

3. RelayFlow-4K Dataset

3.1. Dataset creation

We created RelayFlow-4K using the open-source software Blender. This dataset includes 8,428 frames at 4K reso-



Figure 3. Optical flow with different skip levels. RelayFlow-4K provides flows for adjacent frames (0 skips) and cross-frame flows up to 4 skips, which are uncommon in standard datasets.

lution (3840×2160) from four animations, Charge, Agent 327, Spring, and Sprite Fright. In the case of the Spring scene, some images partially overlap with the existing Spring dataset [46]. However, instead of following the conventional data generation method derived directly from the original animation timeline, we applied temporal stretching and compression, extracting flows and data from temporally distant timestamps. Moreover, there is a significant difference in data resolution between our dataset and theirs (1K Spring vs. 4K ours.) To increase data diversity, we converted a single camera to stereo with 15 cm, 30 cm, or 45 cm baselines, providing ground truth for depth and optical flow over time, suitable for scene flow training and evaluation. Consequently, we obtained data with large displacements in a manner distinct from the existing dataset.

3.2. Dataset components

The demand for intermediate frames, which refer to the frames between the start (t) and end ($t + n$) frames in a sequence, has consistently been needed for reasons such as optical tracking [4, 31] and accumulated approaches [70]. Wu *et al.* [70] provides CVO dataset which provides optical flow annotations across frames; however, due to its small image resolution, it lacks sufficient displacement magnitude between frames when compared to conventional optical flow datasets. In contrast, RelayFlow-4K offers diverse motion, including large displacements, at 4K resolution, and provides both intermediate frames and corresponding flow annotations. More specifically, we provide optical flow annotations between frames separated by larger intervals, referred to as cross-frame flows. RelayFlow-4K’s optical flow annotations include diverse flows between frames up to five frames apart from a reference image, with skips ranging from 1 to 4 frames (see Fig. 3.) This setup offers a diverse set of flow annotations between the reference image and multiple frames at varying intervals. All flow annotations consist of both forward and backward directions.

In addition, RelayFlow-4K provide various masks to

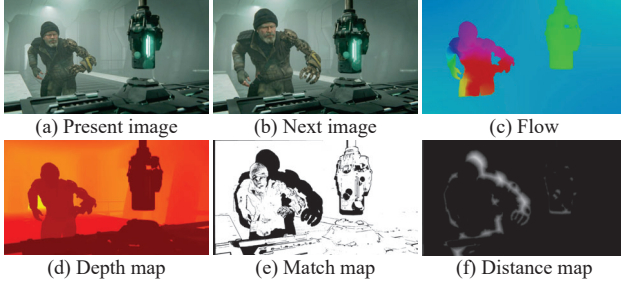


Figure 4. Sample data of RelayFlow-4K. All sample annotations and maps here is aligned with the present image.

support flow research. A match map, derived by checking whether forward warping followed by backward warping returns a pixel to its original position, highlights regions consistent across flows. This enables occlusion analysis and facilitates flow model tuning based on reliably matched, non-occluded areas. It can serve a similar role to the occlusion mask in typical optical flow datasets; however, it provides more advanced information by detecting finer changes in image pixels. We also include a distance map, which calculates the Euclidean distance (unit: pixel) from occluded regions to non-occluded areas, offering further insights into occlusion dynamics. Figure 4 is a sample for our dataset and more details are described in the supplementary details. Overall, RelayFlow-4K provides left and right 4K resolution images captured from a stereo camera setup, along with corresponding depth maps. Therefore, similar to the existing Spring dataset [46], it can be used for scene flow evaluation. Moreover, since we provide both intermediate frames and flow annotations, it can be utilized as data for high-resolution and long-range optical tracking [4, 31].

3.3. Dataset statistic

We illustrate the distribution of RelayFlow-4K with a histogram of flow magnitudes in Fig. 2, comparing it with recent datasets, CVO [70] and Spring [46]. For a fair comparison, we use only the forward flow from the left camera (if stereo is available) in each dataset. Our dataset consistently captures a wider range of larger movements, including sufficient data for negative displacements in the u component.

4. Methods

Problem setup. Given two images, the per-pixel displacement from one image I_i to another I_j is referred to as the optical flow $\mathcal{V}_{i \rightarrow j} = (\mathcal{V}^1, \mathcal{V}^2)$. This flow links the point (u, v) of I_i to the point $(u + \mathcal{V}^1(u), v + \mathcal{V}^2(v))$ of I_j . The long-range optical flow problem we aim to solve involves cases where the values of $\mathcal{V}^1(u)$ and $\mathcal{V}^1(v)$ range from small to large values (e.g., 1000 pixels). Unlike previous approaches [38, 70] that rely on intermediate frames, we address long-range scenarios where only the two images, source and reference, are available.

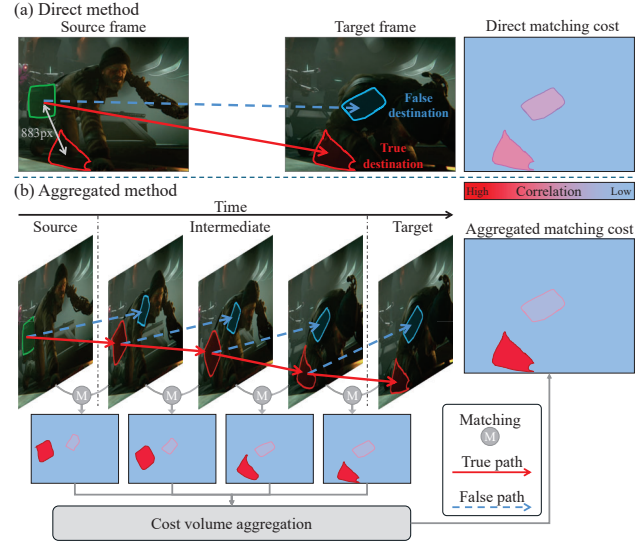


Figure 5. **Motivation of matching cost distillation.** (a): The direct computation of matching cost between the two frames. (b): The accurate long-range matching cost calculation using intermediate frames with relative short displacements only during training.

4.1. Matching cost distillation

As shown in Fig. 5 (a), when similar pixels exist in the target frame relative to the source frame with large displacements, direct methods often struggle to establish a clear matching cost. In contrast, as shown in Fig. 5 (b), the cost volume from intermediate frames contains locally matching pixels, making accurate cost volume estimation easier. Previous methods [6, 70] have addressed long-range flow by accumulating intermediate flows, but they are limited by inference time, memory usage, and practicality, as they require intermediate frames during inference. To overcome these issues, we utilize intermediate frames only during training, relying solely on the first and last frames at inference. More specifically, we propose cost volume distillation, which leverages intermediate frames to improve matching accuracy. The cost volume derived from adjacent frames is more precise than that between frames with large displacements, as local matching is more reliable. To distill this knowledge into the long-range cost volume, we first generate an aggregated cost volume, which serves as a foundation for transferring knowledge to the long-range cost volume, improving accuracy in estimating large displacements.

Given a set of N images, $\mathcal{I} = \{I_1, I_2, \dots, I_N\}$, we first generate the intermediate cost volume set $\mathcal{C} = \{C_{1,2}, C_{2,3}, \dots, C_{N-1,N}\}$. The cost volume [26, 64] is computed as the correlation between features generated by the feature encoder f_θ , as follows:

$$C_{m,n}^{ijkh} = \langle f_\theta(I_m)^{ij}, f_\theta(I_n)^{kh} \rangle \in \mathbb{R}^{(H \times W) \times (H \times W)}, \quad (1)$$

where H, W is the spatial dimension. This represents the correlation between the features generated by the feature

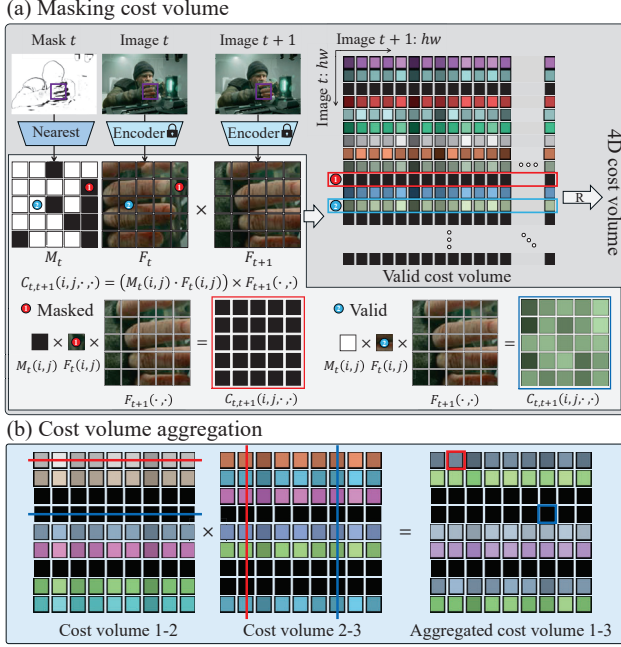


Figure 6. Cost volume aggregation using the match map. (a) shows how valid cost volumes are created with a match map (mask). (b) is the process of aggregation.

encoder f_θ for two images I_m and I_n . The cost volume represents the spatial correlation between two features, so ideally, we can generate the aggregated cost volume, $\hat{C}_{1,N}$, between the two end frames as follows:

$$\hat{C}_{1,N} = \prod_{t=1}^{N-1} C_{t,t+1} \in \mathbb{R}^{(H \times W) \times (H \times W)}. \quad (2)$$

This aggregated cost volume represents a clearer matching correlation, but it has issues such as occlusion. While there may be no occlusion between frames 1 and N , if an occlusion exists between frames 1, 2, \dots , and N , the correlation for those pixels could be relatively inaccurate.

To handle this, we propose a technique that uses a match map to mask during the aggregation of the cost volume. As shown in Fig. 6, we use a match map to apply a mask to the features, assigning a lower correlation score to unmatched regions when calculating the aggregated cost volume. This masked cost volume can be calculated as follows:

$$\tilde{C}_{t,t+1}^{ijkh} = \langle M_{t,t+1}^{ij} \cdot f_\theta(I_t)^{ij}, f_\theta(I_{t+1})^{kh} \rangle, \quad (3)$$

where $t \in \{1, \dots, N-1\}$. The aggregated cost volume considering the matching region is as follows:

$$\hat{C}_{1,N} = \prod_{t=1}^{N-1} \tilde{C}_{t,t+1}. \quad (4)$$

Then, we apply the Kullback–Leibler divergence so that the directly computed cost volume learns the distribution

Algorithm 1 Incremental time-step learning at Stage K

Batch input: Optical Flow Network \mathcal{F} ,

Consecutive image set $\mathcal{S} = \{I_t\}_{t=1}^N$,

Match map set $\mathcal{M} = \{M_{t,t+1} \mid t \in [1, N-1]\}$,

Distance map/optical flow between the ends $\mathcal{D}_{1,N}/\mathcal{V}_{1,N}^{gt}$.

- 1: Let stage K , then $2 \leq N \leq K+1$
- 2: Copy $\mathcal{F} \leftarrow \mathcal{F}$ for distillation
- 3: **for** each epoch **do**
- 4: **for** each batch **do**
- 5: Obtain $C_{1,N}, \mathcal{V}_{1,N}$ using \mathcal{F}, I_1, I_N
- 6: Obtain $\hat{C}_{1,N}$ using $\hat{\mathcal{F}}, \mathcal{S}, \mathcal{M}$ with Eq. (4)
- 7: Calculate \mathcal{L}_{KD} using $C_{1,N}, \hat{C}_{1,N}$ with Eq. (5)
- 8: Calculate \mathcal{L}_{dis} using $\mathcal{D}_{1,N}, \mathcal{V}_{1,N}$ with Eq. (6)
- 9: Update \mathcal{F} to minimize \mathcal{L}_{total} with Eq. (8)
- 10: **end for**
- 11: **if** convergence **then**
- 12: Proceed to the next stage, $K+1$
- 13: **end if**
- 14: **end for**

of the aggregated cost volume. To prevent comparisons in masked regions, we exclude those areas from the computation. Additionally, for KD computation, we apply softmax to each cost volume before performing the calculations, and the equation is as follows:

$$\mathcal{L}_{KD} = \text{KL} \left(\text{Softmax}(\hat{C}_{1,N}) \parallel \text{Softmax}(C_{1,N}) \right) \quad (5)$$

where $\hat{C}_{1,N}^{ijkh} \neq 0$

4.2. Incremental time-step learning

Long-range optical flow is challenging not only because of non-local matching but also because the supervision loss has a large scale, which can make learning the motion of all distributions unstable. Therefore, we propose a strategy where the motion of all distributions is learned sequentially rather than learning them all at once. As mentioned in Sec. 3.2, RelayFlow-4K provides frames with various skips and corresponding optical flows. Therefore, the larger the skip in the frames, the more it corresponds to long-range optical flow, which means more challenging sets.

Therefore, we train on smaller skip scenarios and gradually progress to flow estimation learning in larger skip situations. We define training with 0-skip data up to K -skip data as Stage K . In Stage 0, training proceeds as in standard flow estimation, only using 0 skip frames with adjacent flow GT. In subsequent stages, as larger skip cases are introduced into the training data, intermediate frames naturally emerge from Stage 1 onward. Beyond simply initiating each stage’s training from the output of the previous stage, we also employ the matching cost distillation described in Sec. 4.1. This approach provides additional priors for learning large displacements at each stage. To prevent

Table 2. Quantitative results on the RelayFlow-4K test set, showing total EPE, EPE by flow magnitude, and 1px/5px outlier rates. **-Relay* indicates our method. Best values are in **bold**, second-best are underlined. Lower is better.

Method	All						Match						Unmatch					
	EPE	s0-40	s40-160	s160+	1px	5px	EPE	s0-40	s40-160	s160+	1px	5px	EPE	s0-40	s40-160	s160+	1px	5px
SKFlow [63]	29.48	4.84	77.14	585.89	53.14	35.10	19.56	4.73	75.76	491.75	51.95	33.40	250.01	9.25	85.75	743.45	79.57	72.81
CRAFT [61]	22.58	1.82	29.25	542.50	41.12	13.80	12.77	1.53	25.29	434.02	38.94	11.12	240.64	13.57	53.88	724.07	89.44	73.45
RAFT [64]	18.54	1.37	12.16	466.83	25.90	8.15	9.31	0.99	8.16	342.76	23.24	5.65	223.82	16.45	37.08	674.50	84.91	63.60
RAFT-Relay	9.81	1.39	6.17	230.32	12.96	5.07	<u>3.29</u>	0.69	<u>2.99</u>	107.50	10.20	2.97	154.83	29.72	25.97	435.89	<u>74.27</u>	<u>51.89</u>
GMA [26]	22.08	1.18	21.25	557.18	21.70	8.32	12.00	0.82	17.81	444.62	18.89	5.80	246.43	15.70	42.64	745.60	84.20	64.56
GMA-Relay	7.96	1.45	<u>8.28</u>	<u>173.57</u>	<u>16.27</u>	5.07	2.23	<u>0.74</u>	1.99	55.62	<u>13.74</u>	<u>3.04</u>	<u>135.18</u>	29.63	<u>28.77</u>	<u>371.01</u>	72.50	50.08
GMFlowNet [80]	25.24	3.10	45.34	559.94	50.08	25.91	15.32	2.91	41.30	454.90	48.57	23.73	245.95	<u>10.59</u>	70.47	736.79	83.71	74.46
GMFlowNet-Relay	13.24	<u>1.23</u>	9.94	324.93	19.37	7.83	5.51	7.36	6.48	193.66	16.82	5.55	184.96	20.96	31.48	544.66	76.23	58.56
GMFlow [72]	15.44	1.99	15.43	358.95	39.93	10.69	7.53	1.46	11.49	240.93	37.83	8.25	191.29	23.22	39.99	556.51	86.70	64.94
GMFlow-Relay	<u>8.92</u>	2.28	14.30	170.26	55.27	10.10	3.96	1.80	10.20	<u>77.10</u>	53.84	7.84	119.17	21.73	39.80	326.19	87.11	60.39

instability, the model used to generate the aggregated cost volume for \mathcal{L}_{KD} is a copy of the model trained in the previous stage. When each stage starts, the initial learning rate is halved, and this approach is repeated up to Stage 4. Algorithm 1 shows the detailed incremental time-step learning at stage K . \mathcal{L}_{dis} will be discussed in the following section.

4.3. Matched-region distance loss

In supervised optical flow estimation, gradients are applied uniformly across matched and unmatched regions. Unmatched areas, often caused by motion, become more prevalent in dynamic scenes and significantly impact long-range flow loss. Simply excluding them from loss calculation improves performance in matched regions but degrades accuracy in unmatched areas, necessitating a more adaptive optimization approach. Motivated by the observation that optical flow networks infer the flow of unmatched regions using nearby matched regions [2], we propose a novel loss function, **matched-region distance loss** (\mathcal{L}_{dis}). Specifically, the reliability of a flow prediction at an unmatched point (u, v) is determined by its Euclidean distance to the nearest matched region. By weighting the loss based on this distance, our approach enables effective learning from both matched and unmatched areas, improving flow estimation in large-displacement scenarios. The matched-region distance loss is formulated as follows:

$$\mathcal{L}_{dis}(u, v) = \|\mathcal{V}^{gt} - \mathcal{V}\|_{(u,v)}^1 \left(1 - \beta \frac{D(u, v)}{\zeta_D}\right) \quad (6)$$

$$D(u, v) = \max(\|(u, v) - (x, y)\|^2, \zeta_D) \quad (7)$$

where (x, y) is a position of the nearest non-occluded region, \mathcal{V}^{gt} is the ground truth flow, β is a hyper-parameter, and ζ_D is a threshold for the maximum value. A weight of 1 is applied to pixels in matched regions, equivalent to the original supervised loss. If the pixel is located in an unmatched area, instead of assigning a loss of 0, a reduced loss based on distance is applied. If the distance exceeds this threshold, only $1 - \beta$ of the loss is assigned. Since optical flow networks [26, 61, 64] typically apply loss iteratively, we also define a distance loss, $\mathcal{L}_{dis,i}$, at each iteration

i . Finally, combined with the KD loss, the total loss in our training strategy is calculated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{dis} + \alpha \cdot \mathcal{L}_{KD} = \sum_{i=1}^M \gamma^{M-i} \mathcal{L}_{dis,i} + \alpha \cdot \mathcal{L}_{KD}. \quad (8)$$

5. Experiments

5.1. Experimental settings

We train and evaluate all methods on RelayFlow-4K. We construct the test set by separating 8 scenes from the entire dataset. For evaluation, the test data includes flow between adjacent frames (0-skip) as well as flow between frames that are further apart (from 1-skip to 4-skip), providing a comprehensive assessment. During inference, the resolution of both the images and flow ground truth is 4K (3840×2160).

We compared several leading methods, including RAFT [64], GMA [26], GMFlow [72], SKFlow [63], GMFlowNet [80] and CRAFT [61]. We evaluated the performance of our method by applying it to multiple models, adopting RAFT, GMA, GMFlowNet, and GMFlow.

5.2. Experimental results

Table 2 presents the performance evaluation on RelayFlow-4K, not only for all regions but also for matched and unmatched regions. We assessed performance using the commonly used optical flow metric, End-Point Error (EPE), categorized by flow magnitude as follows: magnitudes up to 40px (s0-40), regions of medium-size displacements with magnitudes of 40-160px (s40-160) and regions of even larger displacements exceeding 160px (s160+). Additionally, we evaluated 1-pixel and 5-pixel outliers.

Although other methods represent recent leading approaches, they face challenges on RelayFlow-4K due to the dataset’s diverse range of flow displacements, encompassing both small and large motions. In contrast, our method, *Relay*, significantly improves performance across all regions for every baselines, reducing errors by over 40% in long-range scenarios, s160+. The observed performance improvements primarily come from improvements in matching regions. For instance, with RAFT, the EPE decreases by 6.02 and the 1px error rate by 13.04%.

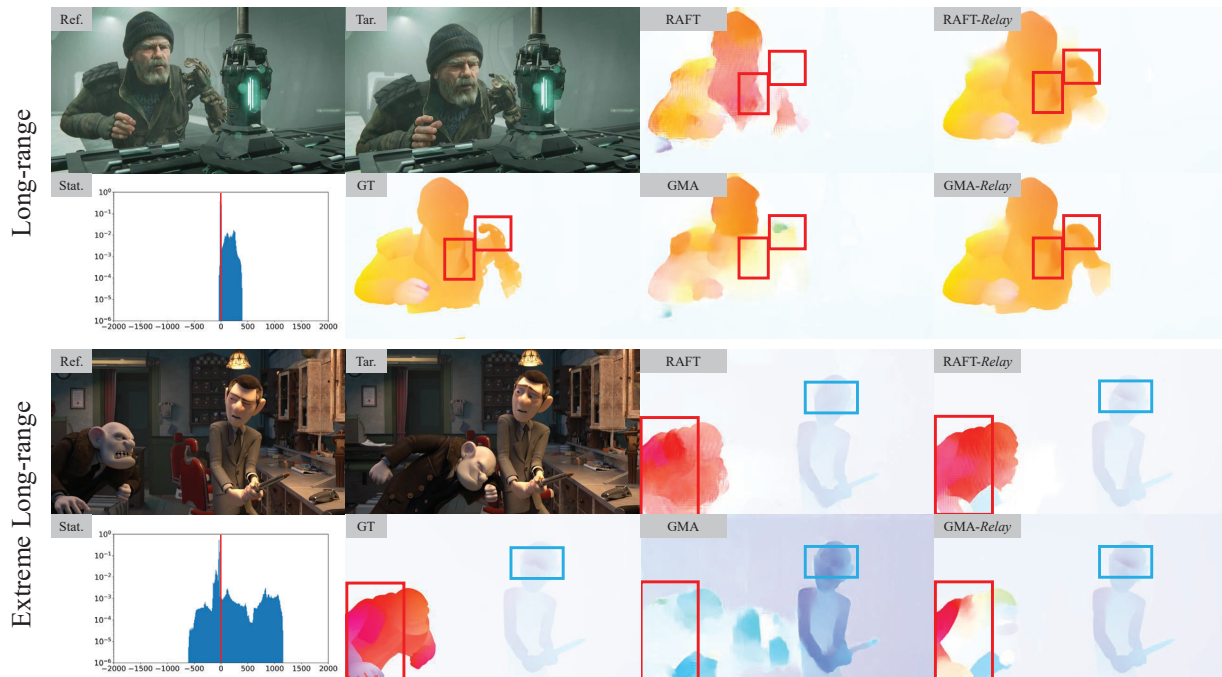


Figure 7. Qualitative comparison on the test set of RelayFlow-4K. ‘Stat.’ represents the distribution of optical flow in each sample.

Table 3. Generalization performance finetuning evaluation using our dataset and our method. “OOM” refers to the out-of-memory issue on an NVIDIA A6000 (48 GB). Ours (*Relay*) uses the proposed module only when R is input, not C or T.

Train Data	Scale	Method	Sintel (train)		KITTI-15 (train)		RelayFlow (test)		
			Clean	Final	F1-epc	F1-all	All	Match	Unmatch
C+T	1/8	RAFT	1.43	2.71	5.04	17.4	OOM	OOM	OOM
C+T+R	1/8	RAFT	1.41	2.72	4.93	17.18	OOM	OOM	OOM
		RAFT-Relay	1.26	2.65	4.77	17.11	OOM	OOM	OOM
C+T	1/8	GMA	1.30	2.74	4.69	17.1	OOM	OOM	OOM
C+T+R	1/8	GMA	1.26	2.54	4.53	15.94	OOM	OOM	OOM
		GMA-Relay	1.23	2.45	4.06	15.19	OOM	OOM	OOM
C+T	1/16	RAFT	1.86	3.16	7.52	26.52	16.99	8.32	210.00
C+T+R	1/16	RAFT	1.81	3.13	7.06	25.96	19.39	10.54	216.32
		RAFT-Relay	1.67	2.86	5.60	21.77	15.64	7.39	199.04
C+T	1/16	GMA	1.86	3.12	8.13	27.75	17.20	8.44	212.01
C+T+R	1/16	GMA	1.90	3.14	7.75	27.55	17.31	8.41	215.34
		GMA-Relay	1.74	3.00	6.30	24.21	14.75	6.59	196.11

As shown in Fig 7, we present two examples with different motion ranges, which are not observed in other datasets, demonstrating that our method consistently predicts accurate flow regardless of object movement. Our approach precisely captures the main object’s long-range motion, outperforming existing methods. While improving accuracy for one range often compromises the other, our method achieves high accuracy for both long- and short-range motion in high-resolution images, as shown with the red box (large displacement) and blue box (short-range) examples.

5.3. Generalization performance

To show the generalization performance about our proposed method and our dataset, we do test using the conventional optical datasets and ours together. Table 3 shows the performance on Sintel [2], KITTI [48], and RelayFlow-4K using

the baseline methods (RAFT [64] and GMA [26] trained on FlyingChairs [8] and FlyingThings3D [45] with down-sampling scales of 1/8 and 1/16. While a 1/8 scale allows training on RelayFlow-4K with cropping, inference requires full-size inputs, leading to memory constraints. We finetune the pre-trained models on our train set using the baseline methods, both with and without our proposed strategy. The results indicate that incorporating our dataset while training alone improves performance. Moreover, applying our proposed method further enhances overall performance.

6. Comparison with other approaches

6.1. Experiment settings

For long-range optical flow estimation, several studies leverage intermediate frames through accumulation [24, 70] during the training and inference process. To compare our method with accumulation-based approaches, we conduct experiments on the HS-Sintel [24] and CVO [70] datasets, which provide intermediate frames. We follow the standard protocol [70], incorporating our proposed techniques: matching cost distillation and incremental time-step learning. Note that our method does not require intermediate frames while the inference.

6.2. Experiments results

Table 4 presents a performance comparison with other methods. The 1st column shows results on HS-Sintel, where our method achieves the highest accuracy in non-occluded regions and the second-best overall performance.

Table 4. Experiments on additional benchmark datasets. We evaluate our approach on the HS-Sintel [24] and CVO [70] datasets. Note that Acc [70] and Lim [38] use intermediate frames during inference to accumulate flow, denoted separately as (MF).

Method	HS-Sintel			CVO (Clean)			CVO (Final)		
	ALL	NOC	OCC	ALL	NOC	OCC	ALL	NOC	OCC
RAFT [64]	2.567	1.426	7.717	4.445	1.948	11.73	4.537	2.003	11.70
+ Ours	1.574	0.652	6.907	4.310	1.598	11.24	4.301	1.560	11.43
+Lim ^(MF)	3.657	1.611	12.36	23.34	6.543	32.90	13.02	7.033	33.82
+Acc ^(MF)	1.383	0.930	4.546	2.634	1.155	7.302	2.707	1.249	7.295
GMA [26]	2.520	1.469	7.600	4.638	2.342	11.33	4.633	2.114	11.36
+ Ours	1.615	0.819	6.223	4.189	1.566	10.85	4.512	1.762	11.22
+Lim ^(MF)	3.306	1.381	11.70	11.39	5.833	31.28	11.68	6.130	31.35
+Acc ^(MF)	1.434	0.950	4.770	2.732	1.181	7.438	2.808	1.261	7.495

Table 5. Comparison of computational cost. We compare ours and AccFlow [70]. Unit of time and memory are ms and GB.

Image Size	1280 × 720		2560 × 1440		3840 × 2160	
	Time	Mem.	Time	Mem.	Time	Mem.
RAFT [64]	50.10	1.97	137.71	4.90	310.26	15.47
+ Ours	50.10	1.97	137.71	4.90	310.26	15.47
+Acc ^(MF)	355.02	2.81	1567.30	9.41	OOM	OOM

The 2nd and 3rd columns display results on the CVO dataset, where our approach ranks second overall. Notably, the CVO dataset involves random object movements, and our method achieves higher flow estimation accuracy by utilizing only the reference and target frames without intermediate frames during inference. Comparing the base models (*e.g.*, GMA and RAFT) with *Relay* (Ours) reveals that our method significantly improves performance in non-occluded regions. This consistent trend across all datasets highlights the generalization ability of our method.

6.3. Computational costs

We compare our method with optical flow methods targeting long-range displacement in runtime and memory usage during inference, using RAFT [64] as the baseline. AccFlow [70] uses 7 frames (5 intermediate), while our method and the baseline use only 2 frames. Table 5 shows runtime and memory usage across image sizes on an NVIDIA A6000. Our method matches the baseline, while AccFlow’s intermediate frames and additional modules significantly increase computational cost, causing OOM issues for 4K images. Details are in the Supplementary.

7. Ablation studies and discussions

Analysis of our proposed strategy. To study the effectiveness of the proposed module, we incrementally added each module to the base model for evaluation in Tab. 6. Each module operates independently and contributes to performance improvement on its own. When combined, they further enhance performance beyond individual use. Notably, incremental time-step learning and \mathcal{L}_{KD} yield the best synergy, and using all proposed methods together ultimately achieves the highest performance.

Table 6. Ablation study of the proposed methods with RAFT [64]. ‘Incr.’ denotes the incremental time-step learning.

Method		All		Match		Unmatch		
\mathcal{L}_{dis}	Incr.	\mathcal{L}_{KD}	EPE	1px	EPE	1px	EPE	1px
✓			18.54	25.90	9.31	23.24	223.82	84.91
	✓		15.36	17.71	6.89	14.79	203.73	82.58
		✓	13.16	15.50	5.11	12.56	192.07	80.91
			12.70	16.79	4.96	13.83	184.85	82.53
✓	✓		12.52	14.85	5.04	12.07	178.95	76.48
✓		✓	11.69	15.52	4.08	12.64	181.01	79.58
	✓	✓	11.16	15.07	3.97	12.17	171.06	79.73
✓	✓	✓	9.81	12.96	3.29	10.20	154.83	74.27

Table 7. Per-stage results of Incremental Time-Step Learning

Stage	All			
	EPE	s0-40	s40-160	s160+
0	14.04	1.50	12.98	335.67
1	12.82	1.48	8.71	308.59
2	12.40	1.33	7.48	302.68
3	11.08	1.31	7.14	266.95
4	9.81	1.39	6.17	230.32

Table 8. Effect of masking while cost volume aggregation

Method	All		Match		Unmatch	
	EPE	1px	EPE	1px	EPE	1px
w/o mask	12.10	13.71	4.92	10.91	171.82	75.87
w/ mask (Ours)	9.81	12.96	3.29	10.20	154.83	74.27

Stepwise performance improvement. We hypothesized that as the model progresses from smaller to larger time-step stages, its performance would gradually improve, with gains reflected across all flow distributions. To validate this, we present the performance at the end of each stage in Tab. 7. Stepwise improvements are observed across nearly all metrics, with particularly notable and consistent gains in long-range cases, such as s160+.

Valid mask of the cost volume. As mentioned in Fig. 6 and Eq. (3), we propose a masking approach to perform weighted cost aggregation that accounts for unmatched regions. In Tab. 8, we conducted an ablation study of cost volume masking. We observed that applying masking led to performance improvements across all evaluation metrics. By suppressing the influence of inaccurate pixels during cost volume aggregation, our approach preserves precise positional relationships between distant pixels, resulting in more accurate flow estimation.

8. Conclusion

We address the core challenges in optical flow estimation with large displacements and dynamic motions by introducing a novel training strategy that uses intermediate frames. To support long-range flow tasks, we present the RelayFlow-4K dataset, offering high-resolution flow data with diverse displacements. This work enhances optical flow accuracy in complex scenarios and provides a foundation for future work in areas such as video frame interpolation, object tracking, and action recognition.

9. Acknowledgments.

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF2022R1A2B5B03002636), and by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2024-00457882, AI Research Hub Project).

References

- [1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 2, 3
- [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. 2, 6, 7
- [3] Hoonhee Cho, Jae-Young Kang, and Kuk-Jin Yoon. Temporal event stereo via joint learning with stereoscopic flow. In *European Conference on Computer Vision*, pages 294–314. Springer, 2024. 3
- [4] Seokju Cho, Jiahui Huang, Seungryong Kim, and Joon-Young Lee. Flowtrack: Revisiting optical flow for long-range dense tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19268–19277, 2024. 3, 4
- [5] Hanqiu Deng, Zhaoxiang Zhang, Shihao Zou, and Xingyu Li. Bi-directional frame interpolation for unsupervised video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2634–2643, 2023. 2
- [6] Qiaole Dong and Yanwei Fu. Memflow: Optical flow estimation and prediction with memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19068–19078, 2024. 3, 4
- [7] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Rethinking optical flow from geometric matching consistent perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1337–1347, 2023. 1, 3
- [8] 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 *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 2, 7
- [9] Pedro Figueirêdo, Avinash Paliwal, and Nima Khademi Kalantari. Frame interpolation for dynamic scenes with implicit flow encoding. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 218–228, 2023. 2
- [10] Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *European Conference on Computer Vision*, pages 59–75. Springer, 2022. 3
- [11] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 3
- [12] Markus Hofinger, Samuel Rota Bulò, Lorenzo Porzi, Arno Knapitsch, Thomas Pock, and Peter Kontschieder. Improving optical flow on a pyramid level. In *European Conference on Computer Vision*, pages 770–786. Springer, 2020. 3
- [13] Yinlin Hu, Yunsong Li, and Rui Song. Robust interpolation of correspondences for large displacement optical flow. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 481–489, 2017. 2
- [14] Hsin-Ping Huang, Charles Herrmann, Junhwa Hur, Erika Lu, Kyle Sargent, Austin Stone, Ming-Hsuan Yang, and Deqing Sun. Self-supervised autoflow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11412–11421, 2023. 3
- [15] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: A transformer architecture for optical flow. In *European conference on computer vision*, pages 668–685. Springer, 2022. 1, 3
- [16] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *European Conference on Computer Vision*, pages 624–642. Springer, 2022. 2
- [17] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Lite-flownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8981–8989, 2018. 2
- [18] Junhwa Hur and Stefan Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5754–5763, 2019.
- [19] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 2
- [20] Woobin Im, Tae-Kyun Kim, and Sung-Eui Yoon. Unsupervised learning of optical flow with deep feature similarity. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 172–188. Springer, 2020. 3
- [21] Azin Jahedi, Lukas Mehl, Marc Rivinius, and Andrés Bruhn. Multi-scale raft: Combining hierarchical concepts for learning-based optical flow estimation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1236–1240. IEEE, 2022. 3
- [22] Azin Jahedi, Maximilian Luz, Marc Rivinius, and Andrés Bruhn. Ccmr: High resolution optical flow estimation via coarse-to-fine context-guided motion reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6899–6908, 2024. 3
- [23] Azin Jahedi, Maximilian Luz, Marc Rivinius, Lukas Mehl, and Andrés Bruhn. Ms-raft+: High resolution multi-scale

- raft. *International Journal of Computer Vision*, 132(5): 1835–1856, 2024. 3
- [24] 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. 2, 3, 7, 8
- [25] Huaizu Jiang, Deqing Sun, Varun Jampani, Zhaoyang Lv, Erik Learned-Miller, and Jan Kautz. Sense: A shared encoder network for scene-flow estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3195–3204, 2019. 3
- [26] 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. 2, 3, 4, 6, 7, 8
- [27] Xin Jin, Longhai Wu, Jie Chen, Youxin Chen, Jayoon Koo, and Cheul-hee Hahm. A unified pyramid recurrent network for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1578–1587, 2023. 2
- [28] Rico Jonschkowski, Austin Stone, Jonathan T Barron, Ariel Gordon, Kurt Konolige, and Anelia Angelova. What matters in unsupervised optical flow. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 557–572. Springer, 2020. 3
- [29] Hyunyoung Jung, Zhuo Hui, Lei Luo, Haitao Yang, Feng Liu, Sungjoo Yoo, Rakesh Ranjan, and Denis Demandolx. Anyflow: Arbitrary scale optical flow with implicit neural representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5455–5465, 2023. 3
- [30] Tarun Kalluri, Deepak Pathak, Manmohan Chandraker, and Du Tran. Flavr: Flow-agnostic video representations for fast frame interpolation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2071–2082, 2023. 2
- [31] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Co-tracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023. 3, 4
- [32] Daniel Kondermann, Rahul Nair, Katrin Honauer, Karsten Krispin, Jonas Andrulis, Alexander Brock, Burkhard Gusefeld, 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
- [33] Lingtong Kong and Jie Yang. Mdflow: Unsupervised optical flow learning by reliable mutual knowledge distillation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(2):677–688, 2022. 3
- [34] Wei-Sheng Lai, Jia-Bin Huang, and Ming-Hsuan Yang. Semi-supervised learning for optical flow with generative adversarial networks. *Advances in neural information processing systems*, 30, 2017. 2
- [35] Junyong Lee, Myeonghee Lee, Sunghyun Cho, and Seungyong Lee. Reference-based video super-resolution using multi-camera video triplets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17824–17833, 2022. 2
- [36] Siyang Li, Xiangxin Zhu, Qin Huang, Hao Xu, and C-C Jay Kuo. Multiple instance curriculum learning for weakly supervised object detection. *arXiv preprint arXiv:1711.09191*, 2017. 3
- [37] SukHwan Lim and Abbas El Gamal. Optical flow estimation using high frame rate sequences. In *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, pages 925–928. IEEE, 2001. 3
- [38] 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. 4, 8
- [39] Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6489–6498, 2020. 3
- [40] Pengpeng Liu, Irwin King, Michael R Lyu, and Jia Xu. DdfLOW: Learning optical flow with unlabeled data distillation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8770–8777, 2019. 3
- [41] Pengpeng Liu, Michael Lyu, Irwin King, and Jia Xu. Self-low: Self-supervised learning of optical flow. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4571–4580, 2019.
- [42] Pengpeng Liu, Irwin King, Michael R Lyu, and Jia Xu. Flow2stereo: Effective self-supervised learning of optical flow and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6648–6657, 2020. 3
- [43] Ao Luo, Xin Li, Fan Yang, Jianguo Liu, Haoqiang Fan, and Shuaicheng Liu. Flowdiffuser: Advancing optical flow estimation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19167–19176, 2024. 3
- [44] Kunming Luo, Chuan Wang, Shuaicheng Liu, Haoqiang Fan, Jue Wang, and Jian Sun. Upflow: Upsampling pyramid for unsupervised optical flow learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1045–1054, 2021. 3
- [45] 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, 7
- [46] Lukas Mehl, Jenny Schmalfluss, 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, 3, 4
- [47] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 3
- [48] 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, 7
- [49] Nicola Milano and Stefano Nolfi. Automated curriculum learning for embodied agents a neuroevolutionary approach. *Scientific reports*, 11(1):8985, 2021. 3
- [50] Henrique Morimitsu, Xiaobin Zhu, Roberto M Cesar, Xiangyang Ji, and Xu-Cheng Yin. Rapidflow: Recurrent adaptable pyramids with iterative decoding for efficient optical flow estimation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2946–2952. IEEE, 2024. 3
- [51] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017. 2
- [52] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12240–12249, 2019. 3
- [53] Zhe Ren, Junchi Yan, Bingbing Ni, Bin Liu, Xiaokang Yang, and Hongyuan Zha. Unsupervised deep learning for optical flow estimation. In *Proceedings of the AAAI conference on artificial intelligence*, 2017. 3
- [54] Zhile Ren, Orazio Gallo, Deqing Sun, Ming-Hsuan Yang, Erik B Sudderth, and Jan Kautz. A fusion approach for multi-frame optical flow estimation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2077–2086. IEEE, 2019. 3
- [55] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 2, 3
- [56] Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12469–12480, 2023. 3
- [57] Hyeonjun Sim, Jihyong Oh, and Munchurl Kim. Xvfi: extreme video frame interpolation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14489–14498, 2021. 2
- [58] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum self-paced learning for cross-domain object detection. *Computer Vision and Image Understanding*, 204:103166, 2021. 2, 3
- [59] Mandyam V Srinivasan. An image-interpolation technique for the computation of optic flow and egomotion. *Biological cybernetics*, 71(5):401–415, 1994. 2
- [60] Austin Stone, Daniel Maurer, Alper Ayvaci, Anelia Angelova, and Rico Jonschkowski. Smurf: Self-teaching multi-frame unsupervised raft with full-image warping. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 3887–3896, 2021. 3
- [61] Xiuchao Sui, Shaohua Li, Xue Geng, Yan Wu, Xinxing Xu, Yong Liu, Rick Goh, and Hongyuan Zhu. Craft: Cross-attentional flow transformer for robust optical flow. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 17602–17611, 2022. 2, 3, 6
- [62] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 2
- [63] 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. 3, 6
- [64] 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. 2, 3, 4, 6, 7, 8
- [65] Bo Wang, Yifan Zhang, Jian Li, Yang Yu, Zhenping Sun, Li Liu, and Dewen Hu. Splatflow: Learning multi-frame optical flow via splatting. *International Journal of Computer Vision*, pages 1–23, 2024. 3
- [66] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3048–3068, 2021. 3
- [67] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu. Occlusion aware unsupervised learning of optical flow. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4884–4893, 2018. 3
- [68] Yang Wang, Peng Wang, Zhenheng Yang, Chenxu Luo, Yi Yang, and Wei Xu. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8071–8081, 2019. 3
- [69] Jerry Wei, Arief Suriawinata, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Louis Vaickus, Charles Brown, Michael Baker, Mustafa Nasir-Moin, Naofumi Tomita, et al. Learn like a pathologist: curriculum learning by annotator agreement for histopathology image classification. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2473–2483, 2021. 3
- [70] 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. 2, 3, 4, 7, 8

- [71] Chengxing Xie, Xiaoming Zhang, Linze Li, Haiteng Meng, Tianlin Zhang, Tianrui Li, and Xiaole Zhao. Large kernel distillation network for efficient single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1283–1292, 2023. 3
- [72] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezaatofghi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130, 2022. 1, 6
- [73] Jia Xu, René Ranftl, and Vladlen Koltun. Accurate optical flow via direct cost volume processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1289–1297, 2017. 2
- [74] Jingwen Ye, Yixin Ji, Xinchao Wang, Kairi Ou, Dapeng Tao, and Mingli Song. Student becoming the master: Knowledge amalgamation for joint scene parsing, depth estimation, and more. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2829–2838, 2019. 3
- [75] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1983–1992, 2018. 3
- [76] Jason J Yu, Adam W Harley, and Konstantinos G Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *Computer Vision—ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part III 14*, pages 3–10. Springer, 2016. 3
- [77] Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Multi-task curriculum framework for open-set semi-supervised learning. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 438–454. Springer, 2020. 3
- [78] Yiman Zhang, Hanting Chen, Xinghao Chen, Yiping Deng, Chunjing Xu, and Yunhe Wang. Data-free knowledge distillation for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7852–7861, 2021. 3
- [79] Shengyu Zhao, Yilun Sheng, Yue Dong, Eric I-Chao Chang, and Yan Xu. Maskflownet: Asymmetric feature matching with learnable occlusion mask. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2
- [80] Shiyu Zhao, Long Zhao, Zhixing Zhang, Enyu Zhou, and Dimitris Metaxas. Global matching with overlapping attention for optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17592–17601, 2022. 6
- [81] Zihua Zheng, Ni Nie, Zhi Ling, Pengfei Xiong, Jiangyu Liu, Hao Wang, and Jiankun Li. Dip: Deep inverse patch-match for high-resolution optical flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8925–8934, 2022. 3
- [82] Yiran Zhong, Pan Ji, Jianyuan Wang, Yuchao Dai, and Hongdong Li. Unsupervised deep epipolar flow for stationary or dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12095–12104, 2019. 3
- [83] Shili Zhou, Ruian He, Weimin Tan, and Bo Yan. Samflow: Eliminating any fragmentation in optical flow with segment anything model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7695–7703, 2024. 3
- [84] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 989–997, 2019. 3
- [85] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 36–53, 2018. 3
- [86] Shay Zweig and Lior Wolf. Interponet, a brain inspired neural network for optical flow dense interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4563–4572, 2017. 2