Bridging the Gap: A Real-World Dataset and Evaluation of Optical Flow Models in Large Displacement Scenarios

Marijn Timmerije
Supervisors: Jan van Gemert, Sander Gielisse
EEMCS, Delft Technical University
CSE3000 Research Project: Computer Science and Engineering

Abstract

Optical flow models excel on synthetic benchmarks but can struggle with real-world scenarios involving large displacements, which are critical for applications like autonomous navigation and augmented reality. To address this, we introduce a novel real-world dataset and evaluation framework, using a specialized annotation tool to capture ground truth optical flow in scenarios with fast movements and close-range objects. Our approach minimizes confounders, providing clear insights into model performance with large displacements. Findings show recent models outperform the previous state-of-the-art, RAFT, across all tested scenarios. Both the annotation tool and dataset are available to support further research.

1. Introduction

Optical flow estimation is a fundamental computer vision task that quantifies the apparent motion of objects between consecutive video frames, typically visualized as a vector field where each arrow represents the direction and magnitude of pixel displacement (Fig 1). Another common visualization uses colour-coding (Fig 2), where the direction of movement is encoded as the hue of a pixel and the magnitude as its intensity.

Figure 1. Arrow visualization of flow between 2 overlaid frames.

This technique is critical for applications like autonomous driving [4], robotics [19], object/scene tracking[12], surveillance [2], and video compression

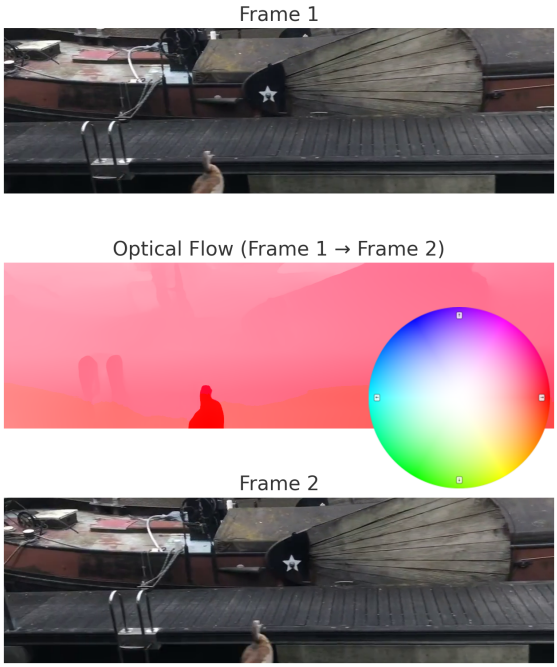

Figure 2. Optical flow visualization with color direction wheel. Predominant red indicates rightward movement or leftward camera motion. The dark red suggests an object is moving faster.

[16]. It involves estimating the motion of pixels between consecutive frames in a video, enabling the analysis of dynamic scenes and the extraction of structural and motion-related information.

Despite significant advancements in optical flow estimation techniques, such as the RAFT (Recurrent All Pairs Field Transform) [17] model, many optical flow models are generally evaluated and benchmarked on synthetic data sets [1], such as Flying Chairs [7] and MPI Sintel [5]. Their performance on real-world data remains underexplored. Although these data sets provide a controlled en-

vironment for benchmarking, they may not fully capture the complexity and unpredictability of real-world scenarios. This can lead to poor performance when these models are deployed in practical applications that include occlusions, lighting changes, low-textured surfaces, repeating patterns, fast movements, and objects moving close to the camera. The last two examples both lead to large displacements between frames and are the focus of this paper.

This research aims to address the following primary question: *How do existing optical flow models perform in real-world scenarios with large displacements of 20 pixels or more?* To answer this question, we will first attempt to answer the following sub-questions:

1. *How do existing models compare in terms of accuracy and robustness, measured by endpoint error (EPE) and percentage of outliers (Fl-all), when evaluated on real-world data with large displacements?*
2. *What are the challenges in annotating real-world videos, especially in scenarios with large displacements?*

To address these questions, we developed an annotation tool for generating optical flow ground truth in real-world videos. Unlike benchmarks like KITTI, which combine multiple challenges without configurable confounders, we are using this tool to create a focused dataset that isolates large displacements while systematically minimizing other variables (e.g., lighting changes, occlusions, non-rigid transformations). This enables targeted evaluation of model capabilities in specific scenarios.

2. Previous Work

Optical flow estimation has evolved significantly from its early mathematical foundations to modern deep learning-based approaches. Old models used to have the brightness constancy assumption, that states that a pixel's brightness remains constant between frames. Modern learning-based approaches do not do this anymore.

The datasets with which models are benchmarked have also seen significant improvements over time to become more complete and challenging.

2.1. Models

Early methods like those by Lucas and Kanade [13] and Horn and Schunck [9] established the mathematical foundations for flow estimation. These classical approaches, while fundamental, were limited by their reliance on local smoothness assumptions.

The introduction of deep learning revolutionized optical flow estimation. FlowNet [8] was among the first to demonstrate that neural networks could be incorporated to better estimate optical flow. This was followed by improvements in FlowNet 2.0 [11] which introduced improvements to the architecture and better training procedures.

RAFT [17] represented a significant advancement by introducing a recurrent architecture that iteratively refines flow estimates, achieving state-of-the-art results on standard benchmarks. Many new developments since then have built on the foundation laid by RAFT.

To address the specific challenge of large displacements, several approaches have been proposed. GMFlow [18] and GMFlowNet [20] utilize global matching mechanisms that better handle large motions between frames by considering the entire feature space. FlowFormer [10] and its improved version FlowFormer++ [15] use transformer architectures to model dependencies, showing strength in scenarios with significant object movement. Another model that seems to be very promising in the area of large displacements is MemFlow [6], which currently sits on top of the Sintel and KITTI-15 leaderboards.

2.2. Benchmark Datasets

Early benchmarks like Middlebury [3] focused on small, controlled movements, but the demands of the field quickly outpaced their simplicity.

The release of Sintel [5] changed this, by providing synthetic sequences with complex motion, large displacements, occlusions and other challenging scenarios. It came in two versions: clean (pristine images) and final (with motion blur and shadows), the latter provided a much more realistic dataset while still having a perfect ground truth.

Real-world datasets like KITTI-15 [14] contain a lot of situations in its scenes that display large displacements, however this dataset is not human-annotated as its ground truth is a sparse annotation based on LiDAR readings.

Other benchmarks also exist that benchmark other difficult scenarios, as explored in the review by Alfarano et al. [1]. Currently these are still used as the main benchmarks for evaluating optical flow models and both Sintel and KITTI have leaderboards where researchers can submit the performance of their models.

3. Methods

To perform the benchmarking, we needed to create a dataset that we could specifically test for various real-world phenomena. To do so, we needed to gather footage, annotate it, and export these annotations in a format that models such as RAFT could work with.

3.1. Tool Development

To facilitate precise and practical optical flow annotation, we developed a custom Python-based tool that uses OpenCV for image processing and PyQt for the graphical interface. Figure 3 shows a screenshot of the main interface where a video is loaded. Its features are highlighted in table 1.

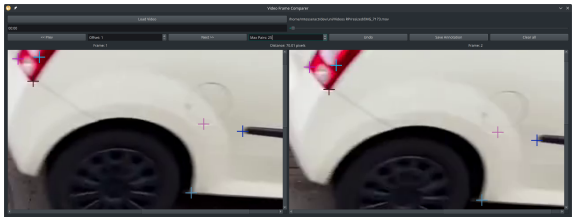

Figure 3. The annotation tool interface showing the dual-frame view with zoom functionality. Corresponding points between frames are marked with colored crosses.

Feature	Description
Video Formats	Supports .mp4, .avi and others for broad compatibility
Dual-frame View	Synchronized side-by-side playback
Precision Tools	Pixel selection with zoom and cross markers
Frame Selection	Non-consecutive frame processing option by skipping frames
KITTI Output	Commonly-supported sparse annotation format

Table 1. Annotation Tool Features

3.2. Data Collection

The dataset comprises five optical flow scenarios (Table 4), each designed to cause large visual displacements while minimizing confounders like lighting changes, non-rigid motion, repetitive patterns and occlusions. Frame selection prioritized visual speed (within the image's space) over absolute speeds, with displacements validated through annotation statistics (Section 2).

We preprocessed all suitable footage using FFmpeg to match KITTI-15's 1242×375 resolution by cropping the original captures and removing audio tracks. This standardization prevented resolution-based artifacts when evaluating models pre-trained on KITTI-15 data. We then organized the processed footage into our five target categories for annotation.

3.3. Annotation

Using our custom tool, we annotated the captured footage by selecting frame pairs with clear large displacements across all five scenarios. For each pair, we manually identified up to 10 corresponding points while avoiding problematic regions like motion-blurred areas, low-texture surfaces, and occlusions. Through iterative verification, we ensured point correspondence accuracy before exporting the final annotations in KITTI-15 format. This process yielded 50 annotated frame pairs per scenario, collectively representing 250 validated samples that matched our displacement targets, as later shown in Table 4.

3.4. Chosen Metrics

To evaluate optical flow models in large displacement scenarios, we selected metrics that capture both accuracy and robustness relevant within the evaluation of large displacements.

Endpoint Error (EPE): The Euclidean distance (in pixels) between predicted and ground truth flow vectors. This directly measures motion estimation accuracy, with large displacements magnifying errors that are critical for stress-testing models. Over a dataset we can define it as follows:

$$\text{EPE}_i = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(u_i - u_{\text{gt},i})^2 + (v_i - v_{\text{gt},i})^2} \quad (1)$$

where:

- N is the total number of pixels in the dataset.
- (u_i, v_i) is the estimated flow vector for pixel i.
- $(u_{\text{gt},i}, v_{\text{gt},i})$ is the ground truth flow vector for pixel i.

Percentage of Outliers (Fl-all): The percentage of flow vectors classified as outliers, where errors exceed either 3 pixels or 5% of the flow magnitude. It follows the standard implementation set in the KITTI benchmark and is defined as follows:

$$\text{Fl-all} = \frac{100\%}{N}\sum_{i=1}^{N}\mathbb{I}\left(\text{EPE}_i > 3.0 \text{ px} \wedge \frac{\text{EPE}_i}{\|\mathbf{f}_i\|} > 0.05\right) \quad (2)$$

where:

- N is the total number of pixels in the dataset.
- EPE_i is the endpoint error for pixel i
- $\|\mathbf{f}_i\|$ is the magnitude for pixel i
- $\mathbb{I}(\cdot)$ is the indicator function.

Accuracy Thresholds (Acc 1px/3px/5px): The percentage of flow vectors with errors relative to ground truth below these thresholds. These can reveal precision at tolerances relevant to specific applications.

Displacement Range Binned EPE: These are EPE metrics segmented by displacement magnitude. These will be used to see which models perform best within which displacement ranges. The selected ranges are *(s0-10px)*, *(s10-40px)* and *(s40+px)* as these are commonly included in benchmarks.

3.5. Evaluation

All models were evaluated using their official KITTI-15 pre-trained weights to ensure fair comparison and leverage

real-world feature learning from this established dataset.

We adapted the standard KITTI-15 evaluation scripts to accommodate our dataset while adding relevant metrics for displacement-range analysis. The modified scripts computed endpoint error across specific ranges (s0-10px, s10-40px, and s40+px), Fl-all outlier rates, and accuracy thresholds at 1px, 3px, and 5px precision levels.

3.6. Ethical Considerations

The collection and annotation of real-world video data raises some important ethical concerns.

Privacy: All videos were recorded in public spaces or with the explicit consent of the participants (when relevant). Faces and license plates, if recognizable, were anonymized or removed during pre-processing to prevent identification.

Transparency: The dataset, validation scripts, and source code of the annotation tool have been publicly released to ensure reproducibility and to facilitate scrutiny of the ground truth quality and generated statistics. They can be found in section 6.

LLM Use: Since an LLM has been used during the development of the annotation tool, we have listed the prompts that were used. It was mainly used to generate boilerplate code for the GUI. It has also been used to help adjust evaluation scripts to extract relevant statistics.

3.7. Model Selection

Given our focus on large displacement scenarios, we prioritize models that have demonstrated strong performance in this specific challenge. Our evaluation includes:

- **RAFT** [17]: The foundational model that serves as our baseline for comparison. It uses iterative refinement of the flow field using recurrent neural networks to achieve high performance. At the time of its release, it achieved state-of-the-art performance.
- **FlowFormer++** [15]: Transformer-based model with improved long-range dependency modeling.
- **GMFlow** [18]: Incorporates global matching to better handle large motions, showing particular strength in scenarios with significant object movement.
- **GMFlowNet** [20]: This model combines global matching with iterative updates, similar to RAFT.
- **MemFlow** [6]: MemFlow introduces a memory module to store historical motion states, which helps predict future motion.
- **MemFlow-T** [6]: This is a variant of MemFlow that replaces its feature encoder with a more robust vision transformer. Based on its placement in the leaderboards, it performs even better than MemFlow itself on multiple metrics.

Our selection criteria primarily focus on performance improvements on large displacement scenarios compared to RAFT, architectural features specifically designed for handling large displacements, and strong performance on benchmark sequences known for challenging motion patterns. By comparing these models against the original RAFT implementation's performance on our dataset, we aim to quantify the improvements these specific models have made in handling large displacements.

4. Results and Discussion

Table 3 shows the results of the evaluation of the models on the "Combined" dataset. The underlined values show which model performed best on a given metric. In tables 5 and 6 are results of the benchmarks on the smaller datasets. These do not include the displacement-based EPE as these ranges have too few annotations (as little as 0.4%) for reasonable statistical analysis.

4.1. Interpretation

MemFlow-T emerged as the top performer in 8 out of 10 metrics in the "Combined" dataset. It has the lowest EPE of 2.451px and the lowest Fl-all outlier rate of 5.1%. The basic version of MemFlow performed quite well on most metrics, but it performed the second worst on EPE, with an error of 5.953px, which is interesting, as on other metrics it almost always performed second best, right behind its vision transformer-augmented variant.

On the other hand, RAFT, which used to be the state-of-the-art during its initial release, did not manage to outperform any of the other models in any scenario, suggesting that major advancements have been made in the handling of large displacements during the years since its release.

Scenario-Specific Insights: On tables 5 and 6, the performance on the individual datasets has also been recorded. There it can be seen that there's considerably more variation in performance.

In many of the scenarios, MemFlow-T still outperforms the other models, but on the "Moving Camera" dataset, MemFlow outperforms it on every metric. This is interesting as it suggests that the vision transformer MemFlow-T is augmented with,might cause a reduction in accuracy under specific circumstances.

Conversely, in the "Fake Low FPS" dataset, where motion blur was minimized but displacements were artificially large, MemFlow-T's robust performance (EPE: 3.123 px) contrasted sharply with MemFlow's high error (EPE: 20.870 px), implying that there are certain situations where MemFlow's default feature encoder fails to produce correct results, while MemFlow-T's vision transformer does handle them correctly. While the exact cause (e.g., architectural sensitivity to frame-skipping) remains unclear and falls outside this study's scope, the results show a major divergence

in model behavior under these circumstances.

4.2. Discussion

The evaluation of optical flow models on our real-world dataset reveals critical insights into their performance in large displacement scenarios. MemFlow(-T) and FlowFormer++ consistently outperformed other models, including the established baseline RAFT. This shows that in the years following the creation of RAFT, considerable improvements have been made to optical flow estimation of large displacements.

A notable finding is the discrepancy between model performance and annotation reliability. Motion blur and occlusions introduced significant ambiguity in ground truth labeling, as seen in Figure 4, where blurred regions made precise correspondence difficult. This raises the possibility that models like MemFlow-T may, in some cases, surpass human annotation accuracy where motion blur or occlusions obscure ground truth.

4.2.1 Challenges

Annotating the ground truth for scenarios that contained large displacements presented significant recording and annotation challenges for humans.

Motion Blur: One of the primary challenges is motion blur, which occurs when objects move rapidly relative to the camera's exposure time. This phenomenon creates smeared edges and causes textures to lose sharpness, making precise pixel correspondence difficult to establish. Motion blur particularly affects fast-moving objects or scenes with rapid camera motion, potentially rendering them nearly impossible to annotate accurately. The visual ambiguity caused by motion blur forces annotators to use their personal judgment about the true position of moving objects, potentially introducing inconsistencies in the ground truth data. This can be clearly observed in Figure 4, where the blurred regions demonstrate how motion blur obscures boundaries of objects and reduces the distinctiveness of visual features needed to accurately annotate them.

Figure 4. An example of motion blur on a vehicle in the dataset. There are few if any points that can be annotated with high confidence.

In a few frame sets, a technique was used to mitigate this by selecting non-consecutive frames from videos of moderately moving objects. This approach artificially amplifies displacements while preserving image clarity, enabling more accurate annotation. This method can be justified for several reasons. Low framerate cameras/video streams or short shutter speeds naturally produce similar frame intervals, making it an approximation of such real-world scenarios. Additionally, by removing image blur, we can evaluate model performance on large motions without the confounding effects of visual artifacts.

Occlusion and Disappearance: Another major difficulty came from occlusions, especially self-occlusion caused by perspective shifts. Objects that move quickly often move out of the frame or features become (temporarily) hidden, creating several challenges for optical flow estimation. This includes the disappearance of correspondence points between frames and the appearance of new objects without clear origin points. Additionally, fast-moving objects frequently experience self-occlusion where parts of the object obscure other parts due to their rapid motion relative to the camera.

Another problem with occlusions in this dataset is that they are a challenging topic on their own that also poses a significant challenge for optical flow models. Because of this, care needs to be taken that occlusions don't become a bigger challenge than large displacements in this dataset.

4.3. Limitations

The current dataset, comprising 250 frames across five different scenarios, was limited by constraints in time and resources, as well as data loss. Expanding the dataset to include over 100 frames per scenario would significantly enhance statistical confidence. This expansion is particularly crucial for both the lower end of the spectrum, where sub-10px displacements currently represent less than 2.5% of all annotations, and the higher end, where only 3.9% of annotations exceed 100px in the "Moving Camera" dataset, these sparsities can be seen in Table 2. The limited proportion of sub-10px annotations, skews the s0-10 EPE for example, effectively turning this range into an error bin and leading to unexpected results within this displacement range. To address this issue, it would be beneficial to intentionally incorporate more low displacement annotations into the dataset. This adjustment would make the s0-10 EPE statistic more reflective of the models' actual performance.

4.4. Future Research

Although our data set reveals critical insights about large-displacement optical flow, several avenues remain unexplored.

Dataset	>10px (%)	10-20px (%)	20-50px (%)	50-100px (%)	>100px (%)
Fast Objects	99.6	0.0	34.2	26.3	39.1
Moving Camera	99.8	5.0	47.7	43.3	3.9
Panning	94.8	8.2	34.8	47.0	4.8
Small Objects	95.2	3.4	24.8	42.1	25.0
Fake Low Fps	98.2	3.8	24.2	29.7	40.4
Combined	97.5	4.1	33.1	37.7	22.6

Table 2. Dataset Annotation Statistics

Model	EPE (px)	Fl-all (%)	Acc 1px (%)	Acc 3px (%)	Acc 5px (%)	>10px EPE (px)
RAFT [17]	6.669	11.325	47.536	83.630	88.921	6.650
FlowFormer++ [15]	2.708	6.581	49.080	87.937	93.363	2.562
GMFlow [18]	2.791	8.077	41.558	84.637	92.152	2.674
GMFlowNet [20]	4.944	9.487	49.420	85.704	90.508	4.956
MemFlow [6]	5.953	5.953	53.868	90.297	94.650	5.864
MemFlow-T[6]	2.451	5.118	53.090	90.430	95.023	2.300
	>10px Fl-all (%)	**>20px EPE (px)**	**>20px Fl-all (%)**	**s0-10 EPE (px)**	**s10-40 EPE (px)**	**s40+ EPE (px)**
RAFT [17]	12.264	6.822	12.547	7.858	3.347	7.983
FlowFormer++ [15]	7.030	2.550	7.126	13.108	4.211	2.768
GMFlow [18]	8.511	2.675	8.633	14.092	4.302	2.716
GMFlowNet [20]	10.368	5.017	10.853	5.695	3.823	5.691
MemFlow [6]	5.822	5.954	5.851	14.816	3.559	6.875
MemFlow-T [6]	4.838	2.311	5.233	15.029	3.171	2.489

Table 3. Core Performance on "Combined" Dataset

4.4.1 Cross-Dataset Generalization

The datasets "Fast Objects," "Moving Camera", and "Panning" exhibit characteristics similar to those found in the KITTI dataset. Therefore, it would be insightful to evaluate their performance when models are pre-trained on other datasets. This approach could provide valuable insights into different models' adaptability and robustness across different data sources.

4.4.2 More Dynamic Scenes

Due to resource limitations and also in part to reduce confounders, it was not feasible to generate exceptionally dynamic scenes. As a result, a lot of the test data contains similar linear motion, where either almost all features in an image are moving the same direction, with possibly a static background. For similar reasons, it has also resulted in a lot of videos being re-used to generate multiple frame annotations.

5. Conclusion

This study addressed the critical gap in evaluating optical flow models for real-world scenarios with large displacements by introducing a novel dataset and annotation framework. Our approach was designed to isolate and analyze large displacements while systematically minimizing confounders such as lighting variations, occlusions, non-rigid motion and giving a clearer insight into model performance on specific conditions.

The results demonstrate that recent models, particularly MemFlow(-T) and FlowFormer++, significantly outperform the previously established state-of-the-art RAFT across all tested scenarios. Their success highlights the effectiveness of memory modules and transformers in predicting large displacements flows, providing guidance for future model development.

6. Code and Data

Our annotation tool was implemented in Python using OpenCV and PyQt. The complete source code is available

on GitHub[1]. Part of the code was written with the help of an LLM. The prompts used are included as well. The FFmpeg script used for resizing is also included in the repository.

The dataset that was created is also available on Github [2] This repository also contains the FFmpeg script, as well as the validation scripts used to extract performance metrics from the models discussed in the paper so that others can replicate the benchmarking.

References

[1] Andrea Alfarano, Luca Maiano, Lorenzo Papa, and Irene Amerini. Estimating optical flow: A comprehensive review of the state of the art. *Computer Vision and Image Understanding*, 249:104160, 2024. 1, 2

[2] Joshan Athanesious, Vasuhi Srinivasan, Vaidehi Vijayakumar, Shiny Christobel, and Sibi Chakkaravarthy Sethuraman. Detecting abnormal events in traffic video surveillance using superorientation optical flow feature. *IET Image Processing*, 14(9):1881–1891, 2020. 1

[3] Simon Baker, Daniel Scharstein, J. 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):1–31, 2011. 2

[4] Aseem Behl, Omid Hosseini Jafari, Siva Karthik Mustikovela, Hassan Abu Alhaija, Carsten Rother, and Andreas Geiger. Bounding boxes, segmentations and object coordinates: How important is recognition for 3d scene flow estimation in autonomous driving scenarios? In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2593–2602, 2017. 1

[5] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conf. on Computer Vision (ECCV)*, pages 611–625. Springer-Verlag, 2012. 1, 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*, 2024. 2, 4, 6

[7] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. v.d. Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 1

[8] Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip Häusser, Caner Hazırbaş, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning Optical Flow with Convolutional Networks, 2015. arXiv:1504.06852 [cs]. 2

[9] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1):185–203, 1981. 2

[10] 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, 2022. arXiv:2203.16194 [cs]. 2

[11] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks, 2016. arXiv:1612.01925 [cs]. 2

[12] Kiran Kale, Sushant Pawar, and Pravin Dhulekar. Moving object tracking using optical flow and motion vector estimation. In *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, pages 1–6, 2015. 1

[13] Bruce Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision (ijcai). 1981. 2

[14] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[15] Xiaoyu Shi, Zhaoyang Huang, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1599–1610, 2023. 2, 4, 6

[16] Chuanbo Tang, Xihua Sheng, Zhuoyuan Li, Haotian Zhang, Li Li, and Dong Liu. Offline and online optical flow enhancement for deep video compression, 2023. 1

[17] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 4, 6

[18] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, 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. 2, 4, 6

[19] Zhiyong Zhang, Huaizu Jiang, and Hanumant Singh. Neuflow: Real-time, high-accuracy optical flow estimation on robots using edge devices, 2024. 1

[20] Shiyu Zhao, Long Zhao, Zhixing Zhang, Enyu Zhou, and Dimitris Metaxas. Global Matching with Overlapping Attention for Optical Flow Estimation, 2022. arXiv:2203.11335 [cs]. 2, 4, 6

[1] https://github.com/IrisPetre99/RP3000

[2] https://github.com/mtesseracttech/LargeDisplacementAnnotations

Dataset	Description	Key Characteristics
Fast Objects	Recordings of moving vehicles captured from close range, emphasizing large pixel displacements.	• Dominant displacements: 20-100px (65.3%) • Most >100px cases (39.1%) • Fast foreground objects with static background • Challenges: Motion blur, reflections
Moving Camera	Footage from a moving tram showing relative motion.	• 99.8% displacements >10px • 47.7% mid-range (20-50px) • Few >100px (3.9%) • Consistent, linear motion patterns
Panning	Controlled camera panning across outdoor scenes.	• 94.8% displacements >10px • 47.0% 50-100px motions • Few >100px (4.8%) • Consistent, linear motion patterns
Small Objects	Close-range footage of primarily pieces of teaware being moved around or being panned over.	• 95.2% displacements >10px • 25.0% >100px cases • Challenges: Low texture due to light
Fake Low FPS	Artificial displacements via frame skipping.	• 98.2% displacements >10px • 40.4% >100px motions • Minimal motion blur • Skips between 5 and 10 frames.
Combined	Aggregate of all datasets.	• 97.5% displacements >10px • Most Balanced distribution • 250 annotated frames

Table 4. Current Dataset Composition and Displacement Characteristics

Dataset	Model	EPE (px)	Fl-all (%)	Acc 1px (%)	Acc 3px (%)	Acc 5px (%)
Fast Objects	RAFT	11.788	23.600	29.071	62.369	73.488
	FF++	4.259	9.600	34.967	74.720	85.099
	GMF	5.506	14.800	26.991	69.011	81.243
	GMFN	8.227	19.600	32.395	68.034	77.411
	MF	4.865	9.601	36.141	77.960	88.399
	MF-T	4.770	9.221	37.029	78.061	87.134
Moving Camera	RAFT	1.484	7.022	55.516	91.292	95.295
	FF++	1.449	8.475	55.165	89.813	95.870
	GMF	1.472	5.811	51.845	92.443	97.252
	GMFN	1.409	4.358	61.979	93.901	97.351
	MF	1.058	2.719	67.203	95.385	98.408
	MF-T	1.289	3.804	66.378	94.975	97.523
Panning	RAFT	6.672	7.739	66.937	91.556	91.956
	FF++	1.165	2.037	70.181	96.556	98.022
	GMF	1.769	3.870	59.034	93.489	95.644
	GMFN	8.628	8.961	64.345	90.022	91.044
	MF	0.890	1.044	70.937	98.756	99.378
	MF-T	0.891	0.622	72.880	99.178	99.778
Small Objects	RAFT	2.236	6.221	39.956	87.567	94.467
	FF++	2.081	5.530	40.394	91.152	95.733
	GMF	2.144	6.912	35.569	87.621	95.333
	GMFN	2.039	6.912	40.870	88.617	93.867
	MF	2.083	7.333	46.182	91.285	95.533
	MF-T	2.181	7.652	40.176	89.566	94.767
Fake Low Fps	RAFT	11.164	10.558	46.198	85.367	89.400
	FF++	4.585	7.371	44.691	87.444	92.089
	GMF	3.063	8.367	34.354	80.622	91.289
	GMFN	4.417	6.375	47.513	87.944	92.867
	MF	20.870	9.067	48.876	88.100	91.533
	MF-T	3.123	4.289	48.987	90.370	95.911

Table 5. General Performance Metrics for Separate Datasets

Dataset	Model	>10px EPE (px)	>10px Fl-all (%)	>20px EPE (px)	>20px Fl-all (%)
Fast Objects	RAFT	11.634	25.668	11.634	25.668
	FF++	4.136	10.367	4.136	10.367
	GMF	5.377	15.745	5.377	15.745
	GMFN	8.085	20.884	8.085	20.884
	MF	4.712	9.490	4.712	9.490
	MF-T	4.618	9.132	4.618	9.132
Moving Camera	RAFT	1.483	7.955	1.482	7.955
	FF++	1.448	8.649	1.440	8.394
	GMF	1.470	6.150	1.452	6.165
	GMFN	1.408	4.322	1.415	4.036
	MF	1.057	2.719	1.051	2.719
	MF-T	1.288	3.804	1.283	3.566
Panning	RAFT	6.927	8.796	7.622	9.710
	FF++	1.192	2.130	1.256	2.664
	GMF	1.816	4.630	1.918	5.051
	GMFN	8.966	10.162	9.737	11.427
	MF	0.904	1.088	0.920	1.187
	MF-T	0.909	0.648	0.937	1.048
Small Objects	RAFT	2.065	8.261	1.927	8.226
	FF++	1.871	7.394	1.746	7.341
	GMF	2.155	8.794	2.013	8.770
	GMFN	2.088	9.806	1.974	11.434
	MF	1.894	6.856	1.744	6.791
	MF-T	2.007	7.472	1.887	9.053
Fake Low Fps	RAFT	11.152	10.500	11.536	10.714
	FF++	4.108	6.411	4.030	6.315
	GMF	2.516	7.078	2.508	6.973
	GMFN	4.391	6.661	4.366	6.480
	MF	20.553	8.767	20.953	8.583
	MF-T	2.625	2.967	2.665	2.891

Table 6. Large Displacement Performance Metrics for Separate Datasets