Evaluating Optical Flow Estimation Models on Real-World Non-Rigid Motion

1: Introduction & Background Information

Optical Flow is the perceived motion of brightness patterns in an image. It can occur from the motions of objects, viewers, or light sources around the object. Optical Flow Estimation is the task of finding pixel translations between two images. This transformation is represented as a dense vector field, where each element of this field represents the pixel transformation between frames.

Estimating optical flow is a fundamental task in computer vision and has applications in fields such as object or gesture tracking, autonomous driving, view reconstruction, image segmentation, or surveillance.

Currently, OFEs are evaluated using synthetic datasets, which contain dense vector fields only. There are currently no real-world datasets containing dense vector fields. However, all practical applications of optical flow estimation occur in real-world contexts. Thus, there is no benchmark to properly assess OFE performance in real-world scenarios. Judging performance solely via performance on synthetic data may not accurately reflect a model's true effectiveness on real-world data.

We seek to investigate the performance of optical flow models on real-world motion. For this, we focus specifically on non-rigid motion, where an object in motion does not retain its original shape. This occurs frequently in the real world but is poorly represented in widely used optical flow datasets.

A taxonomy of non-rigid motion types is composed of Articulated, Quasi-rigid, Homothetic, Isometric, Conformal, Elastic, and Fluid Motion, ordered by increasing non-rigidity. We focus on Articulated, Homothetic, and Conformal motion.

Articulated Motion is the most constrained form of motion of the taxonomy. It occurs when two rigid objects are connected via one or multiple joints. Despite the rigidity of each individual object, the motion of the joined object is relatively non-rigid. Examples include limb movements, such as elbow or knee flexions.

Homothetic Motion occurs when an object uniformly scales while preserving shape. In this, the distance between points scales uniformly, while angles and proportions remain consistent. This type of motion is best described as an expansion or contraction. Examples include the magnification or reduction of a digital image, or the uniform inflation or deflation of a balloon.

Conformal Motion occurs when an object's distances scale non-uniformly while preserving internal angles. While a global object shape distortion can occur, the object's local geometry remains relatively similar. Examples include stretching cloth with printed designs, or 3D shapes rotating on a 2D projected plane.

2: Research Question

How well do optical flow estimation models perform on real-world non-rigid motion?

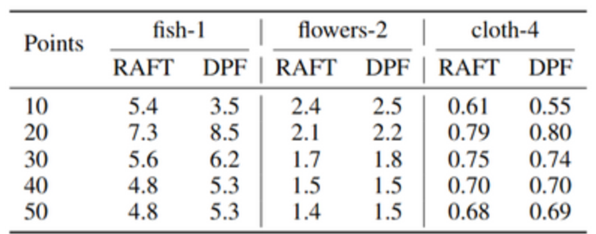
3: Methodology

OFEs vary in use cases. Model architectures are trained on various datasets to ensure optimal model performance. To make the evaluation more robust, we have chosen to use both RAFT and DPFlow architectures. These architectures were chosen due to their consistent model performance on Sintel and Spring datasets, which include scenes containing non-rigid motion.

No annotated dataset of real-world motion exists. Thus, a new custom-made dataset was created from filming various scenarios of non-rigid motion. The dataset consists of 3 folders, one for each of the motion classes. These folders each contain 2 scenes, where a scene is a video containing an example of that type of non-rigid motion. From each scene, 4 pairs of frames were extracted. Thus, we have 24 image pairs, 8 for each type of motion.

Our collected data then needed ground truth values to evaluate our models against. To allow for manual annotation for the collected dataset, a custom annotation tool was developed. It allows users to load, navigate, and annotate video frames or image pairs. It exports annotations into the KITTI2015 format. Annotating every pixel for all twenty-four images is infeasible. Instead, sparsely annotating data was considered. However, model performance on sparsely annotated data is poorly documented. For this, we need to find a balance of model performance and annotation density.

Table 1: Endpoint Errors of listed models evaluated across varying annotated pixel counts for specified dataset subfolders



Points	fish-1		flowers-2		cloth-4	
	RAFT	DPF	RAFT	DPF	RAFT	DPF
10	5.4	3.5	2.4	2.5	0.61	0.55
20	7.3	8.5	2.1	2.2	0.79	0.80
30	5.6	6.2	1.7	1.8	0.75	0.74
40	4.8	5.3	1.5	1.5	0.70	0.70
50	4.8	5.3	1.4	1.5	0.68	0.69

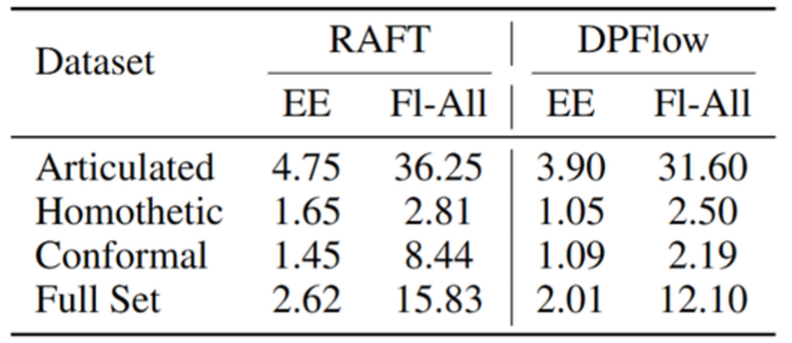
We determined that 40 mappings are optimal for manual annotation of the dataset. This number remains manageable to annotate for all images found in the dataset. Increasing the number of annotations to 50 did not result in any significant improvements in end-point error. With this, we have a total of 960 annotated pixels across the entire dataset, which we deem sufficient to generalize model performance.

Training a model was infeasible due to time and resource constraints, so we instead chose to use pre-trained checkpoints for testing. For both RAFT and DPFlow, we used a model checkpoint trained on the SINTEL dataset. We evaluate the results of our dataset using the standard evaluation metrics of the KITTI dataset: End Point Error (EPE) and Fl-all score.

4: Results

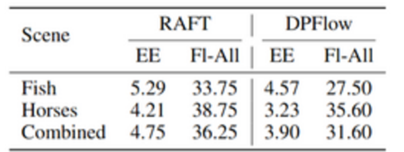
When evaluating the complete dataset, both models achieve End Point Error values below 3 pixels Fl-all scores range between 12%and 16%. The largest performance gap is seen in Articulated motion, where both EPE and Fl-all scores are significantly higher for both models. In contrast, both models show comparable performance in Homothetic and Conformal motion. In all cases, DPFlow maintains a slight advantage in both EPE and Fl-all scores.

Table 5: Endpoint Error and Fl-Scores for the full dataset



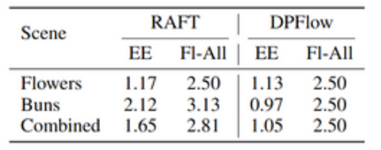
Dataset	RAFT		DPFlow	
	EE	Fl-All	EE	Fl-All
Articulated	4.75	36.25	3.90	31.60
Homothetic	1.65	2.81	1.05	2.50
Conformal	1.45	8.44	1.09	2.19
Full Set	2.62	15.83	2.01	12.10

Table 2: Endpoint Error and Fl-Scores for Articulated Motion



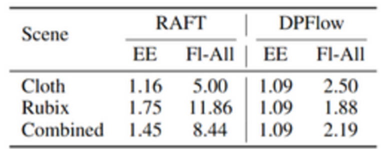
Scene	RAFT		DPFlow	
	EE	Fl-All	EE	Fl-All
Fish	5.29	33.75	4.57	27.50
Horses	4.21	38.75	3.23	35.60
Combined	4.75	36.25	3.90	31.60

Table 3: Endpoint Error and Fl-Scores for Homothetic Motion



Scene	RAFT		DPFlow	
	EE	Fl-All	EE	Fl-All
Flowers	1.17	2.50	1.13	2.50
Buns	2.12	3.13	0.97	2.50
Combined	1.65	2.81	1.05	2.50

Table 4: Endpoint Error and Fl-Scores for Conformal Motion



Scene	RAFT		DPFlow	
	EE	Fl-All	EE	Fl-All
Cloth	1.16	5.00	1.09	2.50
Rubix	1.75	11.86	1.09	1.88
Combined	1.45	8.44	1.09	2.19

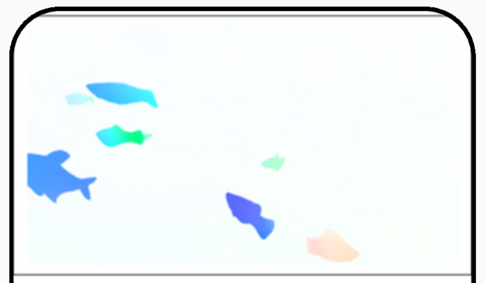


Figure A1: fish-1 Flow Map (DPFlow)

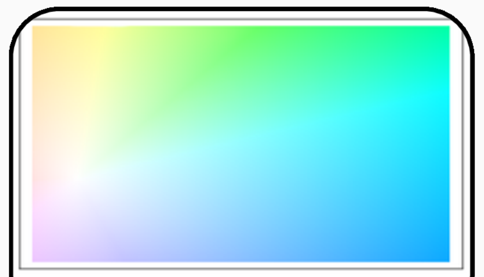


Figure B4: buns-4 Flow Map (DPFlow)

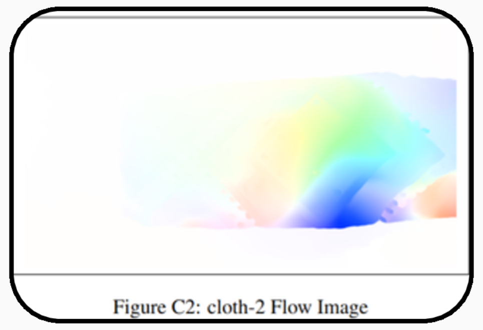


Figure C2: cloth-2 Flow Image

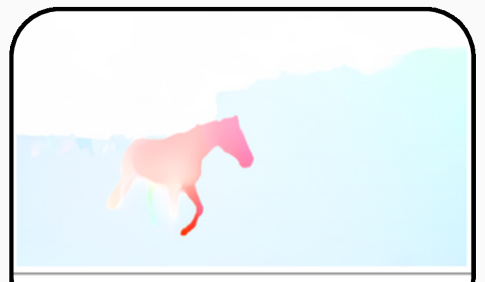


Figure A5: horses-1 Flow Map (DPFlow)

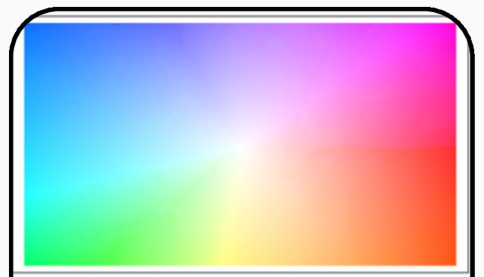


Figure B7: flowers-3 Flow Map (DPFlow)

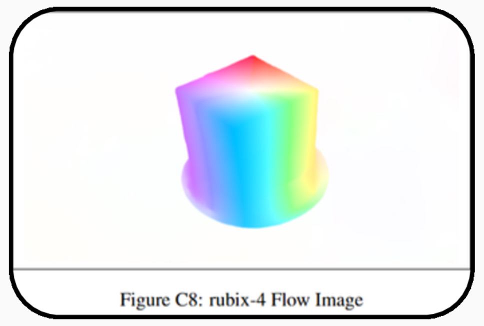


Figure C8: rubix-4 Flow Image

5: Conclusions

While both RAFT & DPFlow demonstrated adequate optical flow estimation under these established conditions, the variations in their performance across motion classes reveals limitations in their generalizability to real-world settings. Both optical flow models performed consistently with synthetic benchmarks for Homothetic and Conformal motion, but performance declined when evaluating Articulated motion. Despite EPE values remaining within the acceptable threshold, high Fl-all scores indicate a lack of consistency and robustness required for many real-world applications.

The findings from this study highlight the importance of more varied datasets for evaluating optical flow models. Though the selected models perform well on synthetic datasets, their real-world performance indicate limitations not captured by synthetic data. Additionally, this highlights the need for more comprehensive datasets consisting of real-world motion patterns.

Author: Sachhyam Dahal. **Supervisor: Sander Gielesse.** **Responsible Professor: Jan van Gemert**