Going Against The Flow

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

Sachhyam Dahal[1]

Supervisor(s): Jan Gemert[1]**, Sander Gielisse**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Sachhyam Dahal
Final project course: CSE3000 Research Project
Thesis committee: Jan Gemert, Sander Gielisse, Alex Voulimeneas

Abstract

Optical flow estimation models are currently trained and evaluated on synthetic datasets. However, the generalizability of these models to real-world applications remains unexplored. This study investigates how well two state-of-the-art optical flow estimation models perform on real-world Articulated, Homothetic, and Conformal non-rigid motion. To facilitate evaluation, a manually annotated dataset comprising twenty-four real-world image pairs and sparse vector fields was created. Both models demonstrated performance consistent with synthetic benchmarks on Homothetic and Conformal motion. However, results degraded when evaluating Articulated motion, revealing limitations in real-world applicability for practical applications such as controlled robotics and object tracking.

1 Introduction

1.1 Optical Flow

Optical flow is the perceived motion of brightness patterns in an image. It occurs from the motions of objects, viewers, or light sources [6]. Optical Flow Estimation is the task of finding the optical flow between two images, represented as a vector field. Each element of this vector field corresponds to a pixel translation between frames.

Estimating optical flow is a fundamental task in computer vision, and has applications in fields such as object tracking, autonomous driving, view reconstruction, image segmentation, and surveillance [9]. Optical flow estimators (OFEs) are models designed to estimate this vector field given two images. A variety of OFEs exist, each having distinct model architectures and approaches to solving the optical flow estimation problem. Current state-of-the-art models achieve estimation errors below 1% of the image size, across a variety of datasets.

1.2 The Knowledge Gap

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

Attempts to tackle this issue have been proposed before. For example, The KITTI [5] dataset uses additional sensors alongside their video data, cross-referencing points to create a dense vector field. However, real world sensor information is noisy and can lead to overcorrection. Another proposed solution is to manually annotate dense vector fields for frame pairs. However, this presents multiple issues for both feasibility and accuracy. Manual annotation beyond a small scale has been described as infeasible [11], making the creation of a large-scale real-world dataset manually annotated dataset largely impractical.

1.3 Area of Research

This research looks to investigate the performance of OFEs in real-world scenarios. For this, we focus specifically on non-rigid motion, where an object in motion does not retain its original shape. This kind of motion occurs frequently in the real world [1], but is poorly represented in widely used optical flow datasets such as KITTI or FlyingChairs [3]. Although synthetic datasets such as Sintel [2] or Spring [7] include examples of non-rigid motion, these datasets cover only a small subset of possible cases. Additionally, OFE performance on these datasets may not be generalizable to the real world. To address this knowledge gap, our research aims to answer the following question:
How well do optical flow estimators perform on real-world non-rigid motion?

2 Background Information

2.1 Non-Rigid Motion Taxonomy

A taxonomy on non-rigid motion is presented by [1]. It is composed of Articulated, Quasi-rigid, Homothetic, Isometric, Conformal, Elastic, and Fluid motion, ordered by increasing non-rigidity. The following explanations of each class of non-rigid motion are adapted from [1], and examples have been added for clarity.

Articulated Motion 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 non-rigid. Examples include limb movements in humans and animals, such as elbow or knee flexions.

Quasi-rigid Motion describes when a deforming object can be seen as rigid up to some point in time. For a short time interval, the body behaves as a rigid object, with minimal deformation. However, when viewed along longer time intervals, the motion of the object is non-rigid. Examples include oscillating aeroplane wings, and tree branches bending in wind.

Homothetic Motion occurs when an object 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, and the uniform inflation or deflation of a balloon.

Isometric Motion occurs when the distances between points on the object's surface are preserved. The object undergoes bending or folding without stretching, compressing, or altering internal angles significantly. While the overall shape may change, the local lengths and angles remain consistent across the surface. Examples include folding clothes, rolling paper, and waving flags.

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, and 3D shapes transforming on a 2D projected plane.

Elastic Motion describes an arbitrary deformation of an object, which returns to its original form once forces are removed. The only constraint to this motion is a form of continuity, where objects do not change their topology due to deformation. Examples include a rubber band being stretched, and a spring being compressed.

Fluid Motion describes the motion of objects with no defined shape, and no resistance to external forces. In this type of motion, angles and distances between two points can change arbitrarily. Examples include smoke drifting in air, and liquid being poured onto a surface.

2.2 Optical Flow Estimators

Estimators Optical Flow Estimators (OFEs) are models that can be used to estimate the translation vector field between two images. OFEs can be categorized into two main classes: knowledge-based and learning-based approaches. Knowledge-based models operate via mathematical modelling, under the assumption that the brightness of a pixel stays consistent between frames [4]. However, their performance is limited due to significant computational demand [9]. Learning-based models instead use deep learning architectures to abstract away from handcrafted features. These architectures outperform classical approaches in both accuracy and runtime [9], and are the current state-of-the-art model paradigm for optical flow estimation.

Limitations All current OFE performance is evaluated using synthetic datasets such as Sintel [2], KITTI [5],or FlyingChairs [3]. Synthetic datasets provide a dense translation vector field, where all pixels are annotated with a corresponding translation. This is used to train models and validate performance. However, practical applications of optical flow estimation occur in real-world environments, yet datasets reflecting these conditions with dense ground truth vectors are notably scarce. None of the commonly used optical flow estimation datasets provide densely annotated real-world scenes. This limitation occurs due to the difficulty of manual annotation. Creating dense ground truth vector fields requires pixel-level labelling across multiple images, a process that has been described as ”unmanageable and error prone” beyond a small scale [11]. As a result, performance benchmarks on synthetic data offer limited insight into how models generalize to real-world scenarios.

3 Methodology

3.1 Model Selection

AOptical Flow Estimator (OFE) architectures take various approaches to solving the estimation problem. To make the evaluation more robust, we have chosen to use two OFE architectures; RAFT [10] and DPFlow [8]. These were chosen due to their consistent model performance on Sintel [2] and Spring [7] datasets, which include scenes of non-rigid motion. Additionally, both RAFT and DPFlow reported state-of-the-art performance on the Sintel dataset at release.

3.2 Data Collection

Due to the absence of a real-world validation dataset, a bespoke dataset was created for evaluation. Gathering a complete dataset comprised of all classes in the non-rigid motion taxonomy was infeasible due to time constraints. Therefore, the dataset consists of only Articulated, Homothetic, and Conformal motion examples. This subset was chosen as it encompasses a large range of deformation types and varying degrees of rigidity, as well as its relevance to practical applications in real-world optical flow estimation.

The dataset consists of multiple video scenes, organized by motion class. Each motion class has a corresponding folder. Each folder contains two scenes. From each scene, 4 frame pairs were extracted and annotated, resulting in a total of 24 image pairs and vector fields. This dataset will be made available alongside the project.

3.3 Data Annotation

Creating an Annotation Tool

The ground truth values necessary for evaluation had to to be annotated manually. For this, a custom annotation tool was developed. This tool enables users to import and display a pair of images (.jpg, .png, or .jpeg) or a video (.mp4, .avi, or .mov). Users can navigate through video frames using single-frame skip buttons, or via a scrollbar. They can also specify a frame offset to simulate larger displacements or time periods between frames. Clicking a pixel in the left image marks it with a randomly coloured cross. Clicking a pixel on the right image while one is marked creates a pair, and both crosses change colour to indicate a successful match. A maximum number of allowed pairs can be specified to control annotation density. An undo function is provided to remove the most recently matched pair, both visually and from the stored list. Users can also clear all existing mappings at once using a button. Finally, the tool allows for the export of annotated frame pairs and their corresponding pixel mappings in the format of the KITTI2015 dataset. The images and vector field are saved to the *image_2* and *flow_occ* folders, respectively. The code to the annotation tool will be made available alongside the project.

Determining Annotation Count

Creating a densely annotated vector field for all image pairs in the dataset was infeasible. Instead, each image was sparsely annotated, with a limited number of annotations on key objects within each frame. However, model performance on sparsely annotated data is poorly documented. Specifically, how the number of annotations influences accuracy. The annotation process had to remain feasible within the constraints of time and resources. Finally, any analysis needed to consider the potential for human error. For this, we needed to find a balance of model performance and annotation density. Table 1 shows the results of brief experimentation using annotation counts on one image pair per taxonomy class.

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

From the results, displayed in Table 1, we determined that 40 annotations are optimal for sparse manual annotation of the dataset. This number of annotations remains feasible across all images found in the dataset. An increase to 50 annotations did not produce significant change to Endpoint Error values. Thus, remaining at 40 is sufficient, resulting in a total of 960 annotated pixels across the entire dataset.

4 Optical Flow Model Architectures

The RAFT [10] and DPFlow [8] model architectures were chosen for evaluation based their performance on the Sintel [2] and Spring [7] datasets. These datasets are widely used and primarily contain non-rigid motion. Thus, we assume these architectures are best suited for accurately estimating examples of unseen non-rigid motion.

RAFT

The RAFT model architecture follows a 3 stage pipeline. First, features are encoded using a convolutional network. This encoder network is applied to both images, and maps the inputs to a feature pyramid using multiple residual blocks, each scaled relative to image size. Additionally, an identical context network architecture is used to extract features from only the first image. Next, RAFT constructs a 4D all-pairs correlation volume by taking the dot product of each extracted feature vector across both images. This set of volumes gives information about both large and small displacements. Lookups are performed on all levels of the feature pyramid. The resulting correlation pyramid enables RAFT to retrieve multi-scale similarity information during estimation. The model iteratively refines an estimated flow through a recurrent update operator, based on a convolutional Gated Recurrent Unit (GRU). Each iteration, updates are made to the current flow, progressively improving accuracy. After a fixed number of iterations, the final low-resolution flow field is upsampled to the original image solution.

DPFlow

DPFlow makes use of a ”recurrent encoder-decoder” architecture. The first stage involves using a recurrent dual-pyramid encoder, utilizing both image and feature pyramids. This allows DPFlow to share multi-scale information across different resolutions, while using the image pyramid to retain the input information at many levels. They also allow for shallower levels to access information from deeper levels. This process is applied to both images in the input pair.

The DPFlow architecture replaces standard attention mechanisms with a convolutional Cross-Gated Unit (CGU) as its core component. While attention is commonly used for extracting matching features, they are computationally expensive and generalize poorly to varying input sizes. A CGU instead allows for efficient and scalable processing using local convolutions. The model is trained using Mixture of Laplace loss, similar to that found in SEA-RAFT [12], but adapted for training on multiple scales. The models are trained using 3 scales and 4 iterations per scale.

5 Experimental Setup and Results

5.1 Experimental Setup

Model Checkpoint Selection

Training an optical flow estimation model was infeasible due to time and resource constraints, so pre-trained model checkpoints were used during testing. For both RAFT [10] and DPFlow [8], we used a model checkpoint trained on the Sintel dataset. This was the only available checkpoint where the training set contained examples of non-rigid motion.

Evaluation Metrics

The annotation tool outputs data in the KITTI2015 [5] format. For evaluation, we replace the *image_2* and *flow_occ* folders in a local installation of the KITTI2015 dataset with a combined folder containing all 24 image pairs and their annotations. We evaluate our results using the standard evaluation metrics of the KITTI2015 dataset: Endpoint Error (EE) and Fl-all score. The EE measures the Euclidean distance between a models' predicted optical flow vector and our manually annotated data points. A lower EE indicates higher accuracy and better model performance in estimating optical flow. The Fl-all score is specific to the KITTI2015 dataset, and measures the percentage of optical flow pixels that are considered outliers. A lower Fl-all score indictes more accurate flow estimation, as there are fewer outlier pixels. A pixel is marked as an outlier if the EE is ≥ 3 pixels, and the magnitude of error is $\geq 5\%$ of the ground truth displacement.

Evaluation Criteria

We must also consider how we classify a successful estimation. We consider an ideal estimation score for model evaluation to be an EE score of ≤ 0.5 % of the image resolution, or 3 pixels. When accounting for human error, we established an upper bound for an ideal EE value to be 5 pixels. An EE of ≤ 1 % of the image resolution, or 7 pixels, will be considered satisfactory performance, and still count as an accurate estimation. All EE values > 7 pixels will be considered inaccurate. We do not set a bound on Fl-all score, as it is specific to performance on the KITTI2015 dataset. However, conclusions will still be drawn from the Fl-all value.

5.2 Results

Articulated motion

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 2 shows the performance of both models per Articulated motion scene. All EE measurements fall within the range of 3 to 6 pixels. Fl-all scores show a high rate of inaccuracy, with more than 30% of all estimations being marked as errors. Across both scenes, DPFlow outperforms RAFT by a significant margin. In the *Fish* scene, DPFlow achieves an EE of 0.72 pixels lower than RAFT, and a 6.25% lower Fl-all score. Similarly, in the *Horses* scene, DPflow has an EE of 0.98 pixels less than RAFT, and a 3.15% lower Fl-all score. Articulated motion, represented by abrupt changes in colour, are well illustrated in Figures A2, A4, and A5. However, some flow maps show a gradient change, uncharacteristic of rigid, jointed Articulated movement. Examples include Figures A4 and A7. The models demonstrate that they can recognize the outlines of non-rigid objects, and can estimate non-rigid motion within such objects.

Homothetic motion

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 3 shows the performance of both models per Homothetic motion scene. All EE values fall below 3 pixels. Fl-all scores remain under 3% for all but one evaluation, which marginally exceeds it. RAFT and DPFlow have varying performances for this class of motion. In the *Flowers* scene, we notice a negligible 0.04 EE difference between both models, and an identical Fl-all score. In contrast, the *Buns* scene highlights a notable performance gap. RAFT records an EE that is 1.15 greater than that of DPFlow. Additionally, RAFT exhibits a higher Fl-all score, increasing from DPFlow's 2.5% to 3.13%. Homothetic transformations result in a gradient across the entire screen, best illustrated in Figures B2, B4, and B7. The models demonstrate consistent performance in estimating optical flow under Homothetic transformations involving pan and zoom dynamics. This holds true for both magnification and reduction across the entire field of view.

Conformal motion

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

Table 4 shows the performance of both models per Conformal motion scene. All EE scores remain under 2 pixels. Fl-all results vary between models, but both remain under 10%. DPFlow outperforms RAFT in all instances. For the *Cloth* scene, there is a negligible 0.07 EE difference between the models. However, RAFT's Fl-all score of 5% is significantly higher than DPFlow's 2.5%. For the *Rubix* scene, there is a much more notable performance gap. There is a 0.66 pixel increase from DPFlow to RAFT. Similarly, the Fl-all scores are significantly higher, as RAFT misclassified 12% of all points. DPFlow only misclassified 2% of all points. Conformal motion is visualized by various gradual gradient changes within a single object, representing the deformations. These is best illustrated in Figures C1, C2, and C7. The models perform consistently well when evaluating optical flow for Conformal transformations, such as oblique rotation or smooth, topological deformations that preserve local angles.

General Performance

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

When evaluating the complete dataset, both models achieve EE values below 3 pixels Fl-all scores range between 12% and 16%. Table 5 shows the average EE and Fl-all scores for all three classes of the motion taxonomy, as well as the combined results for the full dataset. Across all types of motion, DPFlow outperforms RAFT in both metrics. On average, RAFT has an EE that is 0.61 pixels and an Fl-all score 3.73% higher than DPFlow. The largest performance gap is seen in Articulated motion, where both EE 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 EE and Fl-all scores.

6 Responsible Research

Data Collection & Annotation All videos included in this dataset were gathered by the author, and not derived from any external, third-party, or online sources. There is no contractual obligation to the confidentiality of this data. All collected data will be publicised alongside this paper. No images within the dataset contain private information on any individual or property. No confidential or personal information was collected or stored.

Motion & Model Selection Taxonomy classes were selected based on real-world abundance. Models trained on Sintel may provide a positive performance bias of optical flow models on real-world settings. However, it is likely that models trained on relevant datasets will be used in real-world settings. Thus, taking the most suitable training checkpoint is reasonable.

Results & Implications This research provides insight into the real-world performance of optical flow models trained exclusively on synthetic data. The conclusions drawn allows current and future researchers to more accurately assess the generalizability of their models beyond synthetic environments. This can reveal potential limitations to current models that are not prevalent in controlled environments, which to be addressed for generalization into real-world settings. Such evaluations are particularly important in safety-critical applications, such as autonomous driving or robotics, where having inaccurate motion estimation could lead to potentially harmful or dangerous outcomes. In this regard, having a limited training set for real-world scenarios can aid with assessing performance in these settings.

7 Discussion

7.1 Summary

Evaluating the full dataset, both RAFT [10] and DPFlow [8] achieve Endpoint Error (EE) values below 3 pixels, with Fl-all scores of 12.10% and 15.83% respectively. However, model performance varies significantly across motion types. The models perform worst on Articulated motion, where EE values range between 3.23 and 5.29 pixels. Similarly, Fl-all scores are consistently high, between 27.5% and 38.75% misclassified points. In contrast, the models perform comparably on Homothetic and Conformal motion. For Homothetic motion, all EE values fall below 3 pixels, All Fl-all scores remain below 3% in all but once instance, where it reaches 3.13%. Finally, the models have varying results for Conformal motion. All EE values fall below 2 pixels, but Fl-all scores vary between 1.88 % and 11.86%.

7.2 Interpretations

When assessed against the evaluation criteria defined in section 5.1, both models demonstrate acceptable performance. Both have an average EE below 3 pixels, which is the boundary for perfect annotation without human error. RAFT achieves an EE of 2.62 pixels, with an Fl-all score of 15.83%. DPFlow achieves an EE of 2.01 pixels, and an Fl-all score of 12.10%. For both models, it is important to recognize the differences that each data subset had on these value totals.

Articulated

The models perform adequately on Articulated motion. Both total EE measurements exceed the 3 pixel bound, but fall within the 5 pixel bound for ideal estimation including a human error offset. However, all Fl-all scores are consistently high. Roughly a third of the data was marked as inaccurate, with an upwards of 62 (38.75% of 160) misclassified points when evaluating the *Horses* scene. While both models produce higher EE values on the *Fish* scene, they achieve lower Fl-all scores compared to their performance on the *Horses* scene. Both models show EE and Fl-all values that are significantly higher than synthetic benchmarks on many datasets. Visual inspection of the estimated flow fields indicate that the models sometimes estimate gradual motion gradients. This is inconsistent with the rigid, jointed motion patterns of Articulated motion.

Homothetic

The models perform exceptionally well on Homothetic motion. Both average EE measurements fall within the 3 pixel threshold for ideal estimation, and Fl-all scores are consistently low, with a maximum of 5 pixels (3.13% of 160 annotations) being misclassified in a single scene. These EE values are consistent with Sintel-clean benchmark results. RAFT has an EE of 1.65 and an Fl-all score of 2.81. The EE values are comparable to the benchmark average EE of 1.61. Similarly, DPFlow achieves an EE of 1.05, which is virtually identical as its synthetic benchmark of 1.046. Visually, the models appear to recognize and allocate appropriate gradients throughout the full image for instances of Homothetic motion.

Conformal

Both models perform reliably on Conformal motion. All EE values fall well below the 3 pixel threshold for ideal estimation. However, Fl-scores vary per model, ranging anywhere between 3 (1.88% of 160) and 19 (11.86% of 160) misclassified pixels for the same scene. The average EE values for these models is comparable to that of Homothetic Motion. RAFT achieves an EE 0.16 pixels lower than its Sintel-clean benchmark, while DPFlow scores only 0.04 pixels higher than its benchmark. These deviations are negligible, indicating that both models perform at a level consistent with their performance on synthetic datasets. However, the Fl-all scores vary significantly per model. In the *Cloth* scene, RAFT misclassifies twice as many points as DPFlow. This disparity is even more prevalent in the *Rubix* scene, where RAFT produces nearly 6.5 times more misclassifications. The models appear to capture both abrupt and gradual texture deformations on the surface of the object, without generalizing motion to the outline of the object, Thus recognizing instances of Conformal motion..

7.3 Implications

These results have several implications for the optical flow estimation of non-rigid motion in real-world scenarios. Both RAFT and DPFlow achieve our ideal estimation, with average EE values falling below 3 pixels for both models. However, the accuracy of these models is largely dependent on the type of motion being estimated. When estimating Articulated motion, both models had abnormally high EE and Fl-all scores, indicating poor accuracy and consistency. For applications of optical flow estimation in the real world, we must consider these limitations. The poor performance in estimating Articulated motion highlights a limitation of these models in estimating complex, jointed deformations. Though the values for EE still fall within our classification of satisfactory, the high Fl-all scores suggest estimation inconsistency. This indicates that these models should be adapted or modified before use in practical applications such as object tracking or autonomous robotics. In contrast, the high consistency of performance across Homothetic and Conformal motion suggest that both models are well suited for scenes exhibiting smooth deformations across varying scales and contexts. In these settings, EE and Fl-all remain close to synthetic benchmarks. This indicates reliable performance in settings where such motion is present, such as autonomous driving, controlled robotics, or image and video processing applications. For any of these applications, the data suggests DPFlow to be the better choice, consistently outperforming RAFT in both EE and Fl-all values.

7.4 Limitations

While the results presented seem promising, several limitations need to be acknowledged. Firstly, our analysis covered only three out of seven classes within the non-rigid motion taxonomy. As a result, not all behaviours of non-rigid motion are represented in the dataset. Additionally, we focused exclusively on non-rigid motion. Thus, no conclusions can be drawn from this data regarding optical flow performance on rigid, quasi-rigid, isometric, elastic or fluid motion.

Furthermore, the dataset used in this study was limited in scope. Combinations of different motion types were not considered. The number of annotated frames was relatively small, and manual annotation introduced potential for human error. Though this was corrected for, conclusions are only accurate to a certain extent. Moreover, we only evaluated two optical flow estimation model architectures – RAFT& DPFlow . While these models represent different approaches and were both state-of-the-art at the time of their release, a broader evaluation across a larger set of models would provide more generalizable results.

Additionally, our analysis utilized only two metrics. Though EE and Fl-all are widely used and standardized metrics, they may not fully capture model behaviour under diverse conditions. Finally, the visual conditions in our dataset were clean, with minimal presence of real-world artefacts. Image effects such as motion blur, occlusion, noise, or distortion were not present. This clarity could contribute to the performance of the models, and does not represent the full range of challenges encountered in real-world applications. Thus, though the results are promising, they can not be assumed to generalize across all current and future real-world datasets, motion types, or artefacts.

8 Conclusions and Future Work

This study aimed to evaluate the performance of Optical Flow Estimators (OPEs) on real-world non-rigid motion, addressing the knowledge gap presented by a limited non-rigid motion examples in existing data and a lack of real-world benchmarks. For this, we created a dataset consisting of real-world Articulated, Homothetic, and Conformal motion. This dataset was composed of 24 image pairs, each with 40 manually annotated ground truth vectors. For evaluation, we used the standard metrics of the KITTI2015 [5] benchmark; the Endpoint Error (EE) and the percentage of misclassified pixels (Fl-all). Based on our image resolution, we defined a successful estimation as having an EE of ≤ 5 pixels. However, we considered estimations with errors of ≤ 7 pixels as accurate, being $\leq 1\%$ of the total image resolution. We then evaluated two state-of-the-art OPE architectures, RAFT [10] and DPFlow [8], against these criteria. These architectures were chosen for their consistent performance on Sintel [2] and Spring [7] datasets, which contain examples of non-rigid motion. Evaluation was done using model checkpoints trained on the Sintel dataset.

While both RAFT and DPFlow demonstrated successful estimations under these conditions, the varied range in performance between different classes of motion reveals limitations to generalizability in real-world applications. For Homothetic and Conformal motion, both OPEs performed consistently with synthetic benchmarks. Visual analysis confirms that both models generalize well to smooth, continuous non-rigid motions. However, performance declined when evaluating Articulated motion. The complex, partially rigid behavior presented significant challenges for both models. Despite EE values for Articulated motion remaining within our acceptable threshold, high Fl-all scores indicate a lack of consistency and robustness required for many real-world applications.

These findings highlight the importance of varied datasets for evaluating OPEs. Though the selected models perform well on synthetic datasets, real-world evaluations suggest under-represented motions in synthetic datasets. This reinforces the need for more comprehensive datasets containing real-world motion patterns. Furthermore, the study did not include visual artefacts, which likely contributed to such high performances across motion classes. In real world scenarios, models must perform well against repeating patterns, occlusions, motion blur, and various lighting scenarios, among others. In conclusion, while this study shows that current OPEs are capable of achieving performance comparable to synthetic benchmarks in real-world non-rigid motion, it also reveals limitations in their consistency between motion types. Our study highlighted a significant performance drop in Articulated motion, which is most relevant to real-world applications such as autonomous robotics or object tracking.

Future work may extend this study by expanding the dataset to include more classes of non-rigid motion, or by increasing the number of scenes within existing classes. Additionally, more optical flow estimation architectures could be tested, beyond the two covered in this paper. Furthermore, a more realistic performance indication may be achieved by adding various image effects that would be present in real-world applications.

References

[1] J.K. Aggarwal et al. “Nonrigid Motion Analysis: Articulated and Elastic Motion”. In: *Computer Vision and Image Understanding* 70.2 (1998), pp. 142–156. ISSN: 1077-3142. DOI: https://doi.org/10.1006/cviu.1997.0620. URL: https://www.sciencedirect.com/science/article/pii/S1077314297906202.

[2] D. J. Butler et al. “A naturalistic open source movie for optical flow evaluation”. In: *European Conf. on Computer Vision (ECCV)*. Ed. by A. Fitzgibbon et al. (Eds.) Part IV, LNCS 7577. Springer-Verlag, Oct. 2012, pp. 611–625.

[3] A. Dosovitskiy et al. “FlowNet: Learning Optical Flow with Convolutional Networks”. In: *IEEE International Conference on Computer Vision (ICCV)*. 2015. URL: http://lmb.informatik.uni-freiburg.de/Publications/2015/DFIB15.

[4] Denis Fortun, Patrick Bouthemy, and Charles Kervrann. “Optical flow modeling and computation: A survey”. In: *Computer Vision and Image Understanding* 134 (2015). Image Understanding for Real-world Distributed Video Networks, pp. 1–21. ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2015.02.008. URL: https://www.sciencedirect.com/science/article/pii/S1077314215000429.

[5] Andreas Geiger et al. “Vision meets Robotics: The KITTI Dataset”. In: *International Journal of Robotics Research (IJRR)* (2013).

[6] Berthold Horn and Brian Schunck. “Determining Optical Flow”. In: *Artificial Intelligence* 17 (Aug. 1981), pp. 185–203. DOI: 10.1016/0004-3702(81)90024-2.

[7] Lukas Mehl et al. *Spring: A High-Resolution High-Detail Dataset and Benchmark for Scene Flow, Optical Flow and Stereo*. 2023. arXiv: 2303.01943 [cs.CV]. URL: https://arxiv.org/abs/2303.01943.

[8] Henrique Morimitsu et al. *DPFlow: Adaptive Optical Flow Estimation with a Dual-Pyramid Framework*. 2025. arXiv: 2503.14880 [cs.CV]. URL: https://arxiv.org/abs/2503.14880.

[9] Syed Tafseer Haider Shah and Xiang Xuezhi. “Traditional and modern strategies for optical flow: an investigation”. In: *SN Applied Sciences* 3.3 (2021), p. 289.

[10] Zachary Teed and Jia Deng. “RAFT: Recurrent All-Pairs Field Transforms for Optical Flow”. In: *CoRR* abs/2003.12039 (2020). arXiv: 2003 . 12039. URL: https://arxiv.org/abs/2003.12039.

[11] Adrian Wälchli and Paolo Favaro. *Optical Flow Dataset Synthesis from Unpaired Images*. 2021. arXiv: 2104.02615 [cs.CV]. URL: https://arxiv.org/abs/2104.02615.

[12] Yihan Wang, Lahav Lipson, and Jia Deng. *SEA-RAFT: Simple, Efficient, Accurate RAFT for Optical Flow*. 2024. arXiv: 2405.14793 [cs.CV]. URL: https://arxiv.org/abs/2405.14793.

The appendices contain the predicted optical flow mappings from DPFlow **only**. Performance was sufficiently similar (<2 pixels) in EE measurements that conclusions for both models can be drawn from these visualizations.

A Articulated Motion Flow Images

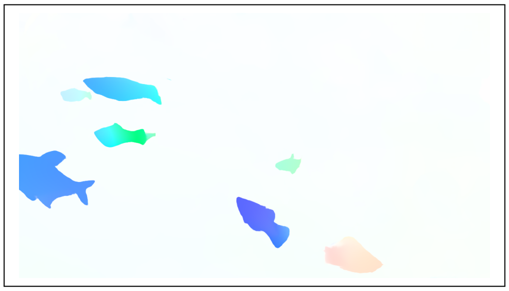

Figure A1: fish-1 Flow Map (DPFlow)

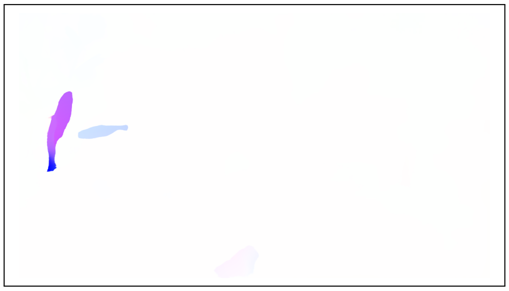

Figure A2: fish-2 Flow Map (DPFlow)

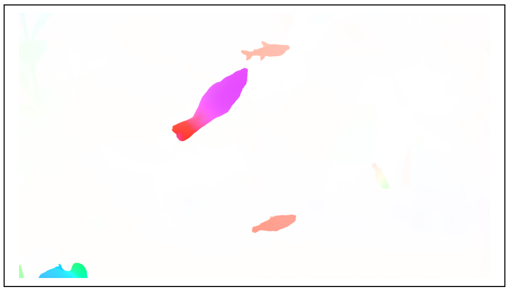

Figure A3: fish-3 Flow Map (DPFlow)

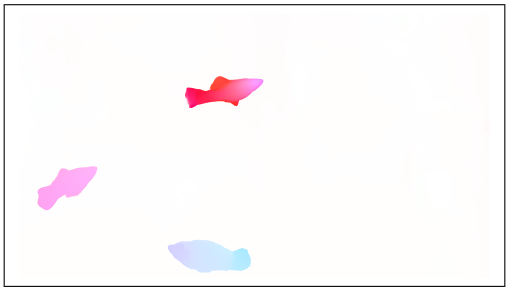

Figure A4: fish-4 Flow Map (DPFlow)

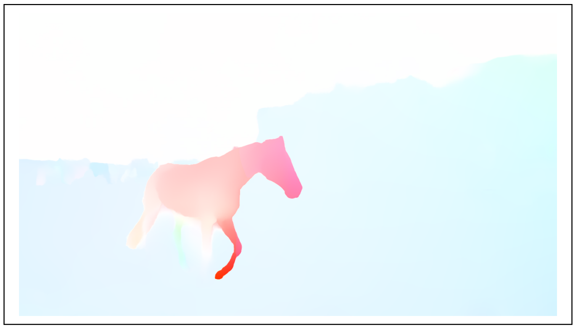

Figure A5: horses-1 Flow Map (DPFlow)

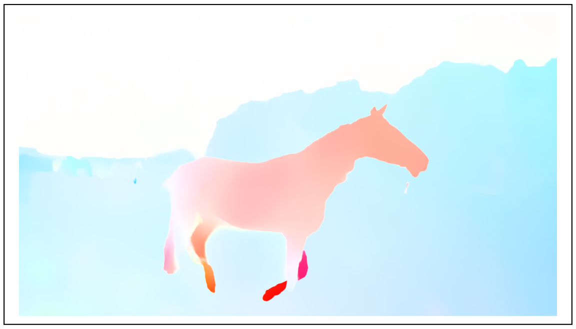

Figure A6: horses-2 Flow Map (DPFlow)

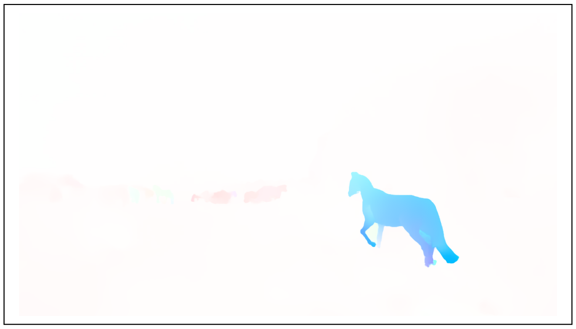

Figure A7: horses-3 Flow Map (DPFlow)

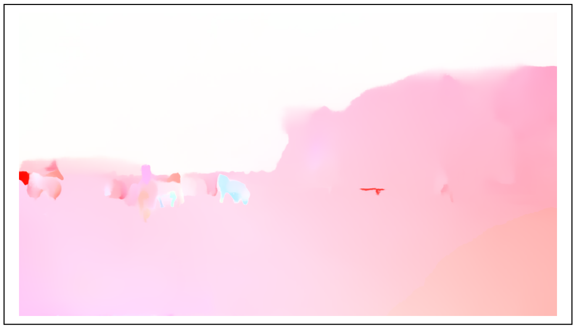

Figure A8: horses-4 Flow Map (DPFlow)

B Homothetic Motion Flow Images

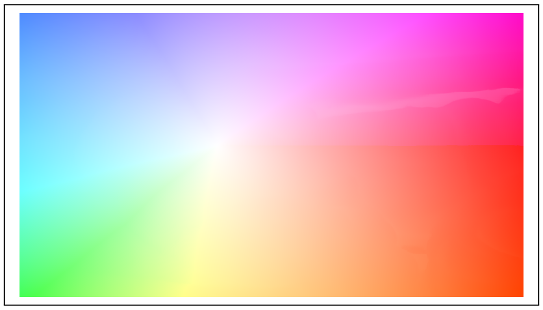

Figure B1: buns-1 Flow Map (DPFlow)

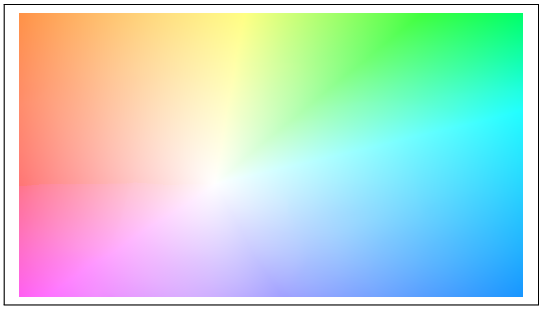

Figure B2: buns-2 Flow Map (DPFlow)

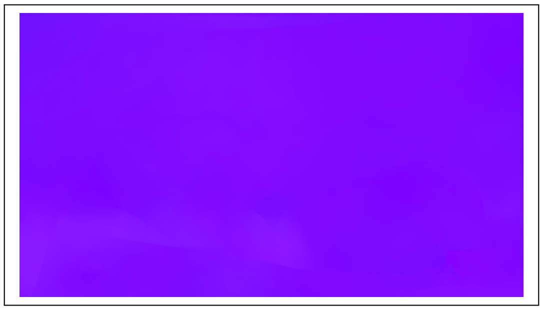

Figure B3: buns-3 Flow Map (DPFlow)

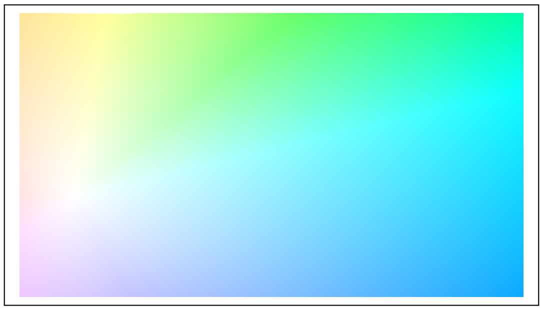

Figure B4: buns-4 Flow Map (DPFlow)

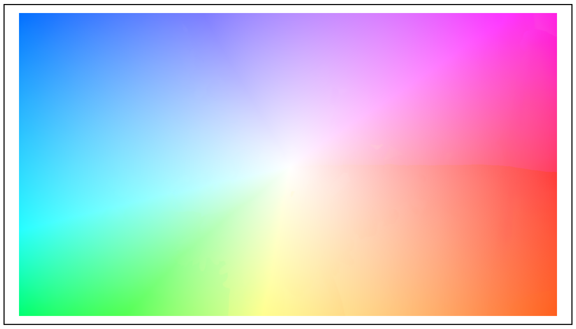

Figure B5: flowers-1 Flow Map (DPFlow)

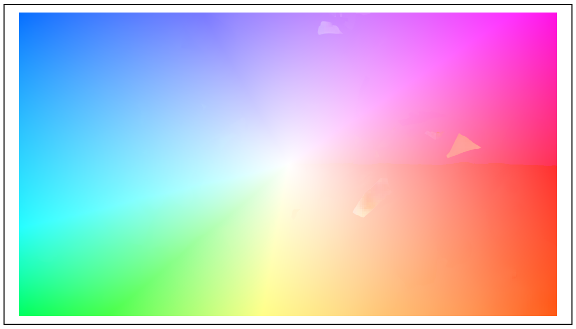

Figure B6: flowers-2 Flow Map (DPFlow)

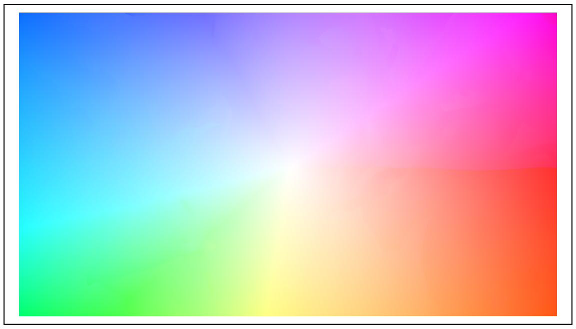

Figure B7: flowers-3 Flow Map (DPFlow)

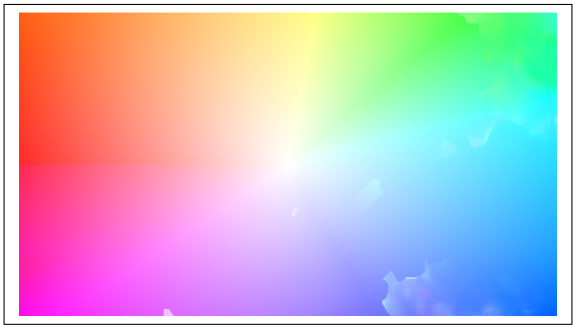

Figure B8: flowers-4 Flow Image

C Conformal Motion Flow Images

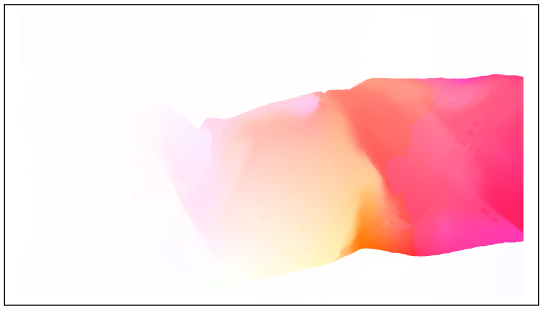

Figure C1: cloth-1 Flow Image

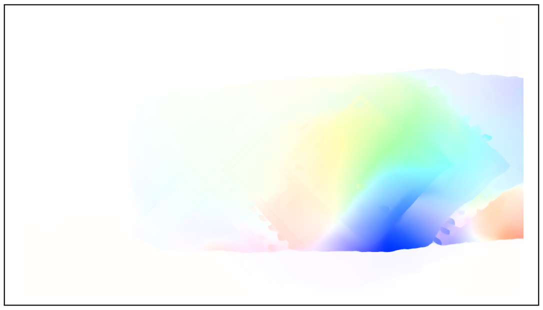

Figure C2: cloth-2 Flow Image

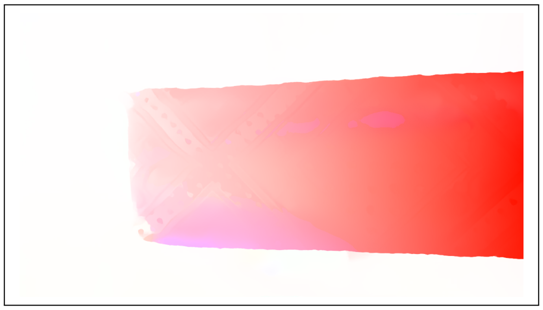

Figure C3: cloth-3 Flow Image

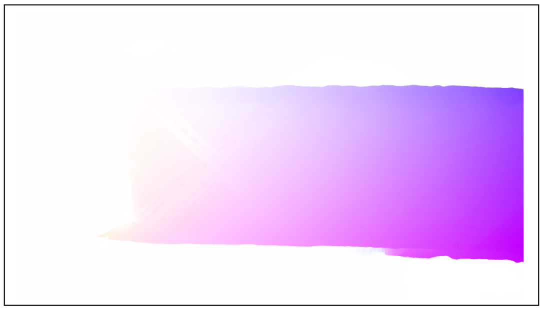

Figure C4: cloth-4 Flow Image

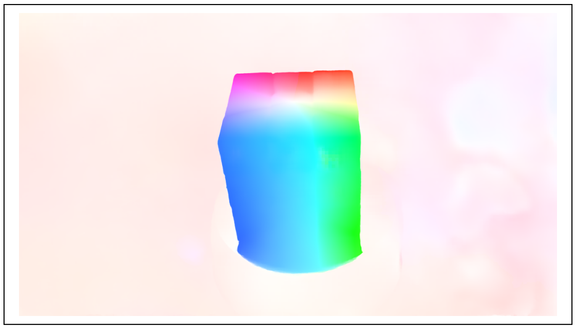

Figure C5: rubix-1 Flow Image

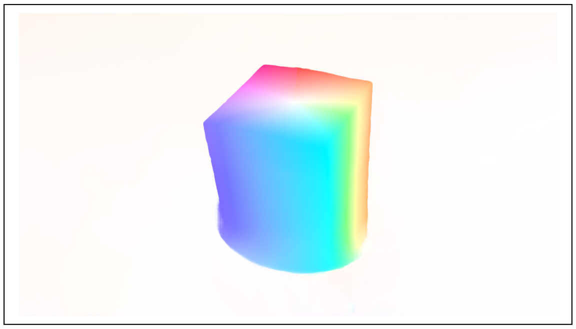

Figure C6: rubix-2 Flow Image

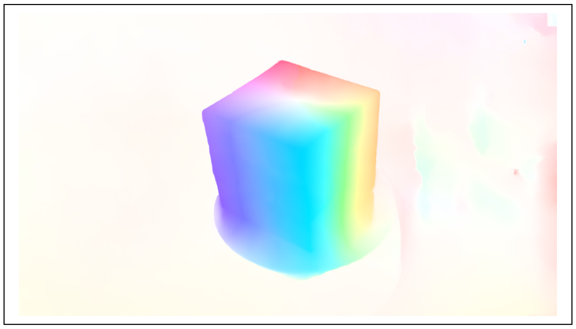

Figure C7: rubix-3 Flow Image

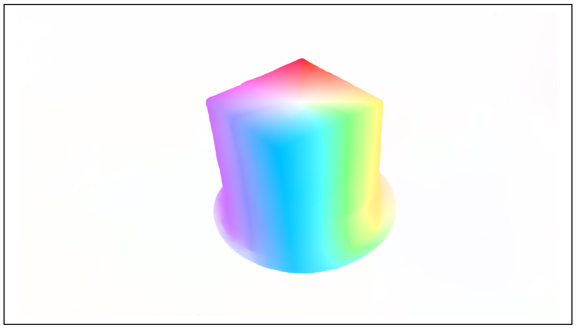

Figure C8: rubix-4 Flow Image