

On the Real-World Generalisability of Optical Flow Models

Petter Reijalt , Sander Gielisse , Rickard Karlsson , and Jan van Gemert 

Computer Vision Lab, TU Delft, Delft, the Netherlands

Abstract. Real-world deployment of vision models to broadly benefit society is arguably a main research objective. In optical flow, however, the difficulty to obtain the ground truth has focused research mainly on synthetic data and domain-specific benchmarks. Here, we investigate the severity of this mismatch. We study how well modern optical flow estimation models generalise to real-world video and question if accuracy on synthetic benchmark proxies actually predicts accuracy on real-world optical flow. To address this, we build a real-world evaluation benchmark and evaluate the real-world generalisability of a broad set of recent optical flow models using standard checkpoints. Our benchmark contains 8,204 frame pairs across TAP-Flow, Slow Flow, and our own dataset FlowFactor. FlowFactor is a manually annotated real-world benchmark of 1,000 HD frame pairs organised into four confounding factors: large displacements, repetitive textures, occlusions, and lighting variation. Each setting mainly varies only one factor, enabling diagnostic, confounder-specific analysis. Using FlowFactor, we reveal that performance on varying lighting and large displacements correlates most strongly with real-world accuracy, and that improvements on large-motion regimes can trade off against robustness in small-motion, stationary scenes. Our experiments show that progress on Sintel, KITTI and Spring only weakly predicts accuracy on real-world data, highlighting the need for a broad real-world optical flow benchmark. Interestingly, scaling up the amount of training data does not necessarily resolve the gap, calling for new innovative research instead of simply scaling data and compute. Code and datasets are available at <https://github.com/Petter6/real-world-optical-flow>.

Keywords: Optical Flow · Real-world Benchmark · Generalisation

1 Introduction

Optical flow, the apparent 2D motion of pixels between consecutive frames, is a foundation for real-world applications ranging from action recognition and video editing to autonomous driving and robotics [3]. Modern optical flow methods [10, 16, 19, 32, 35, 36] learn from annotated data. Due to the difficulty of obtaining annotated per-pixel ground truth optical flow, the training data is created synthetically, such as in FlyingChairs (C) [10] and FlyingThings3D (T) [26]



Fig. 1: How well does optical flow quality on current benchmarks generalise to real world optical flow quality? Optical flow generalisability is, amongst others, measured on synthetic benchmarks [16, 32, 40] such as Sintel [7]. In this paper, we investigate if current-day benchmarks like Sintel can accurately measure generalisability to real-world settings. Both optical flow predictions above were made by the *things* checkpoint of FlowFormerPlusPlus [32]. Whereas the prediction on the synthetic frame is nearly perfect, the real-world example reveals noticeable errors, particularly around the legs and parts of the arms, and a loss of fine detail in the hair. This leads us to ask the question: How well do current-day benchmarks measure generalisability to real-world benchmarks?

and their combination (C+T). For evaluating the generalisation of optical flow it is shown that synthetic pre-training can already yield competitive results [16] on differently distributed datasets such as Sintel [7] and KITTI [28]. A notable gap in evaluating optical flow generalisability is diverse real-world data, as illustrated in Fig. 1. Such real-world data is not used due to hardships of obtaining optical flow ground truth where obtaining pixel-accurate motion for every frame is not trivial, due to occlusions, non-rigid motion, lighting changes *etc.* Manually annotating every pixel is essentially infeasible: even sparse tracking requires multiple annotator hours for only tens of points in a few seconds of video [9]. Dense ground truth for real scenes therefore typically arises either from *synthetic rendering* or from specialised hardware and reconstruction pipelines [4, 18, 28].

In this paper we ask the question whether or not optical flow models generalise to out-of-distribution (OOD) real-world data. This is important because in the real-world, it is not possible to finetune on in-distribution data, because of the difficulty of obtaining real-world optical flow annotations. Thus, OOD generalisation is often benchmarked by training on C+T and evaluating on more realistic datasets [7, 27, 28]. These datasets are, however, narrow in domain, mainly driving cars, or are synthetic in nature. Recent literature [2] suggests the generalisation ability of optical flow models has stagnated when it comes to synthetic corruptions. Here, we thus create our own benchmark, using broad, real-world data to estimate the generalisation gap. We incorporate in our real-world benchmark, three real-world datasets, namely Slow Flow [18], TAP-Vid-DAVIS [9], and FlowFactor. We created the FlowFactor dataset to give a more fine-grained error metric, to improve understanding of optical flow failure modes. Most benchmarks report on broad metrics such as average end-point-error (EPE), where averaging over multiple confounders makes confounder-specific gains hard to measure. This makes model design harder, because it’s unclear what the model can actually

improve. We create a broad real-world dataset to isolate different confounding factors, see Fig. 2. We choose the following confounders, according to classical analyses and more recent overviews: occlusions, large displacements, illumination changes, and complex textures. As they remain enduring challenges for optical flow models [11, 16, 32, 45, 48, 49]. This leads to our FlowFactor dataset. Our main contributions are:

- We introduce *FlowFactor*, a new sparsely annotated real-world optical flow benchmark designed around four isolated challenge factors: occlusions, lighting changes, repetitive textures, and large displacements. This enables factor-specific analysis of failure modes rather than a single aggregate error.
- Using FlowFactor, we provide a detailed analysis of where current models fail and highlight directions for developing optical flow methods that generalise more reliably to real-world applications.
- We create a real-world optical flow evaluation benchmark containing 8,204 samples. We transform the TAP-Vid-DAVIS [9] dataset into an optical flow dataset, and combine this with the Slow Flow [18] and FlowFactor datasets into a benchmark. Using this benchmark we do a comprehensive benchmark of the current state of real-world optical flow accuracy.

2 Related Work

2.1 Optical Flow Datasets and Benchmarks

Early evaluation efforts date back to Barron *et al.* [5], combining both synthetic data from the Yosemite sequence [15] and simple real toy datasets. Later, the Middlebury benchmark emerged as a relatively small, laboratory-style dataset [4]. While crucial for methodology development, these datasets are limited in resolution, diversity, and motion complexity.

The large-scale data requirements of deep learning triggered a wave of synthetic datasets. FlyingChairs [10] renders chairs undergoing planar motions over random natural-image backgrounds to support FlowNet training [10]. Mayer *et al.* [26] later introduced a variety of optical flow datasets, including FlyingThings3D, which simulates everyday objects moving in 3D. These datasets remain the standard for pre-training, often referred to as C+T (FlyingChairs + FlyingThings3D) in recent works. Customisable dataset generators such as Kubric [14] allow users to synthesise many variants of scenes and tasks, including optical flow, by randomising scene content and rendering parameters. Autoflow [34] allows for learning a dataset, based on the accuracy on a target dataset. This dataset is learned by a hyperparameter search using evolutionary algorithms. All of these pre-training datasets are synthetic in nature. This leads to a generalisation gap to the real-world.

Evaluation benchmarks tend to be more realistic in nature. MPI Sintel [7] is based on an open-source animated film and includes challenging effects such as motion blur, long-range motions, and non-rigid deformations. KITTI 2012/2015



Fig. 2: Factorisation of Optical Flow failure modes. (a) In existing optical flow benchmarks, multiple confounders, such as large displacements, occlusions, repetitive textures, and lighting variations, co-occur throughout the scene. This makes it difficult to attribute accuracy drops to specific confounders. (b) We instead isolate these confounders in a controlled real-world setting, enabling a more targeted analysis of known hard individual factors, that affect optical-flow accuracy, namely large displacements, repetitive textures, varying lighting, and occlusions. This dataset can then be used to pinpoint more fine-grained weaknesses of optical flow models.

[13, 28] provide real driving scenes, where optical flow ground truth is obtained semi-automatically by using LIDAR and camera parameters to estimate non-moving objects, and fitting 3D CAD models to moving vehicles. HD1K [21] extends real-world driving benchmarks to a higher resolution and more diverse weather and illumination. More recently, Spring [27] offers a large, high-resolution, computer-generated benchmark for stereo, optical flow, and scene flow, drawn from a photorealistic Blender short film with $4\times$ super-resolved ground truth. Slow Flow [18] is a real-world benchmark that takes a hardware-driven approach by using high-speed cameras. By exploiting the approximate linearity of small motions, the authors reconstruct accurate reference flow fields for a variety of real-world scenes.

Other specialised datasets focus on particular phenomena such as low-light conditions with varying brightness and noise [49] or robustness to a broad range of visual corruptions [31].

While this ecosystem of benchmarks is large and diverse, most of them report aggregate metrics, such as EPE, or only coarse breakdowns such as occluded vs. non-occluded regions, foreground vs. background, or broad displacement bins. More fine-grained statistics, such as rigid vs. non-rigid regions and low vs. high texture, are beginning to appear [27], but they are not measured in isolation: multiple confounding factors change simultaneously, making it difficult to attribute failures to specific causes. Our work is complementary in that FlowFactor is explicitly designed to isolate individual factors: occlusions, lighting changes, repetitive textures and large displacements; in a controlled, real-world setting.

2.2 Sparse Motion Annotations and Point Tracking

Obtaining dense ground truth for arbitrary real videos remains prohibitively expensive, motivating research on sparse annotations. Over the years, various real-world optical flow datasets have used human involvement to different degrees: fitting CAD models to moving objects [28], manually selecting object contours and interpolating flow inside [22, 24, 50], annotating motion boundaries [41], or directly annotating individual pixels [9]. These approaches exploit the human visual system’s sensitivity to motion, depth ordering, and object grouping [33]. Our FlowFactor dataset, the TAP-Flow dataset, and the densely labeled Slow Flow benchmark [18] combined, yield a suite of real-world datasets that cover both sparse and dense labels across diverse conditions. For our FlowFactor dataset we choose sparse manual labeling because it is the most versatile approach and supports annotation of non-rigid motion.

2.3 Robustness, generalisation, and factorised evaluation

Due to the use of synthetic training data, several surveys and benchmarks have begun to question the robustness and generalisation of optical flow models. Specialised benchmarks probe particular phenomena, such as low-light conditions [49] or robustness to synthetic corruptions including noise, blur, and compression [31]. FlowBench [2] explicitly aims to benchmark in-distribution accuracy, reliability, and generalisation via corruption and adversarial attacks, and reports that as accuracy on standard benchmarks has increased, reliability and robustness have not necessarily followed.

Another line of work focuses on generalisation across datasets. Many optical flow models claim strong distribution generalisation by training on less realistic synthetic datasets [17, 26] and evaluating on more realistic validation sets such as Sintel [7], Spring [27], or KITTI [28]. The Robust Vision Challenge [1] and related efforts encourage evaluation on multiple benchmarks to assess cross-dataset robustness reporting on the mean error on all of these datasets combined, though hyperparameter tuning is allowed on those benchmarks beforehand, making it difficult to fully disentangle generalisation from dataset-specific tuning [6].

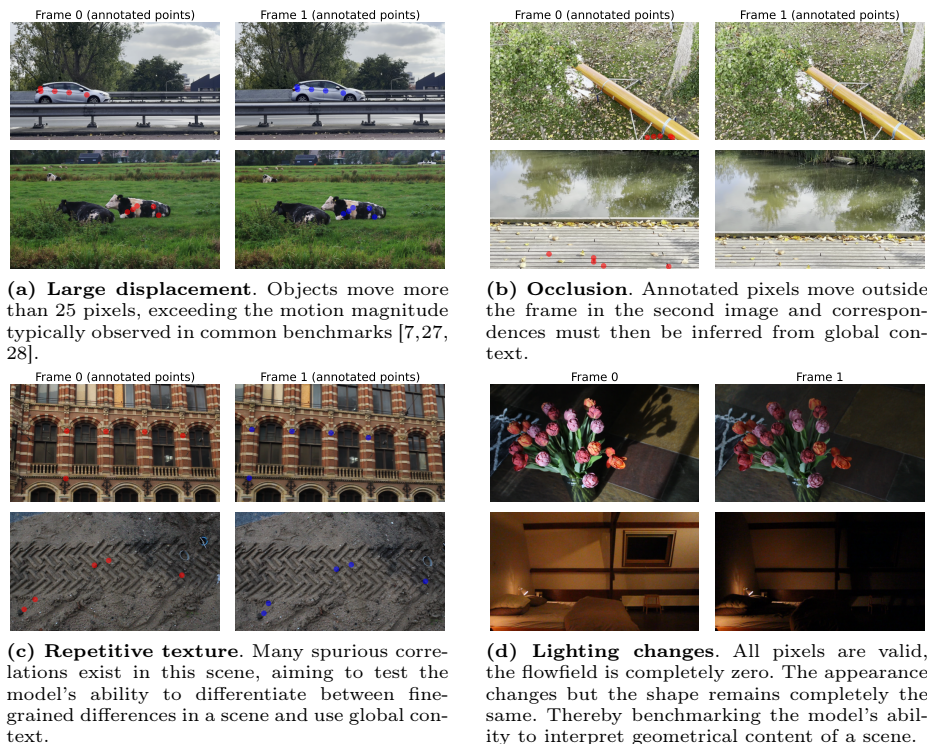


Fig. 3: Example frame-pairs from the FlowFactor dataset. Annotated points are best viewed zoomed in.

Closest to our work are efforts that analyse cross-dataset generalisation of modern deep optical flow models [2, 36, 43]. However, these datasets are often narrow in domain [7, 21] or synthetic in nature [7]. Moreover, these evaluations typically aggregate accuracy over entire datasets and do not systematically isolate individual failure factors in real scenes. Also, they often rely on a single real-world benchmark (e.g., KITTI [28] or Slow Flow [18]) and use test data that is closely related to the training data [2]. In contrast, our work focuses on real-world evaluation suites that are unseen during training and are explicitly designed (in the case of FlowFactor) to factorise evaluation into four isolated challenges: occlusions, lighting changes, repetitive textures, and large displacements. This allows us to move beyond aggregate error measures and towards a more diagnostic understanding of the root failure causes of current optical flow models.

3 Datasets for real-world generalisation

We construct three datasets to benchmark how well optical flow models generalise to the real-world. The *FlowFactor* dataset allows for precise controlled

settings, while *TAP-flow* [9] captures an uncontrolled setting, with multiple challenging confounders, Slow Flow [18] allows for a dataset that also incorporates motion blur and has a dense flow field.

3.1 FlowFactor: Isolated Category Benchmark

To gain insight into failure cases we sparsely annotated 1,000 frame pairs at a resolution of 1280×720 , grouped into four distinct categories: large displacements, repetitive textures, occlusions, and lighting variation. These factors are selected based on common failure modes in optical flow [4] and target different challenges, which can be encountered in real-world use. The isolated settings allow for controlled model testing. Examples can be seen in Fig. 3.

Large displacements. Large displacements evaluate whether a model can accurately track substantial motion. We select cases where the Euclidean distances between pixel pairs are ≥ 25 pixels, but typically is much larger. This displacement can be achieved by object motion, camera motion, or both object motion and camera motion.

Occlusions. Occlusions assess whether a model can reason about objects that become partially hidden, either by other objects or by the image boundary. We include 125 pairs with out-of-frame occlusions, where the object leaves the image boundary, and 125 pairs with inter-object occlusions. For the inter-object case we take visually flat surfaces that get partially occluded (e.g. a tree in front of a car) and we linearly interpolate occluded points on the surface.

Repetitive Textures. Repetitive textures test the ability to resolve visual ambiguity when local appearance is insufficient. As the local region of a pixel remains visually similar, it requires the model to use global context to solve the task, as it cannot be solved by local context. We select 250 frame pairs containing visually similar but spatially shifted textures. This produces ambiguous local appearance, requiring global context for correct correspondence.

Lighting Variation. Lighting variation tests robustness to changes in illumination that alter appearance while leaving the underlying motion unchanged. We select 250 static frame pairs where lighting intensity, shadows, or glare change across frames while all objects remain stationary. These pairs contain dense zero-flow fields but look visually different.

Pixel selection strategy. We follow the annotation strategy of Doersch *et al.* [9] by manually selecting five informative points per pair. Annotators are free to choose points of interest, resulting in natural diversity in difficulty, location, and appearance. The points that are chosen are easily annotatable such as bright spots and edges, examples can be seen in Fig. 4. We show the correctness of our annotation by asking five people to annotate up to 20 points per person resulting

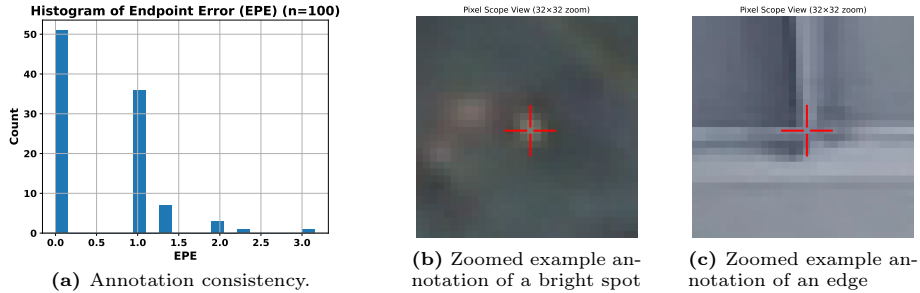


Fig. 4: Feasibility and accuracy of hand-annotated points. (a) Histogram of Euclidean distances, when 100 of our points are annotated by others. The average distance is 0.58, which falls within our used 3-pixel error metric. (b and c) We annotate easily recognisable pixels like edges and bright spots.

in an average 0.58 EPE to our chosen points, way below the three point threshold, which is the margin in our FI-error metric. The results can be seen in Fig. 4a.

A custom annotation interface exports the results in KITTI-style flow maps [8, 12, 20, 30, 37]. We use 5 points per scene as this gives a strong indication of the accuracy of a model on the entirety of the scene relative to other models.

We verify this hypothesis by using the first two frame pairs of each TAP-Vid-DAVIS [9] video and sampling 2 sets of 5 non-overlapping annotated points from the same frame. This leads to 2 datasets with 60 frame pairs each. Each dataset uses the same scenes but different points. We hypothesise that the relative ordering of the models remain roughly the same between both datasets, thereby showing that 5 points can be used as a strong proxy for the relative accuracy on the entire scene. This appears to be true as the spearman correlation coefficient is 0.95 ($p = 3.51e-22$), which suggests a strongly positive correlation between the ranking of both point sets. Thereby providing substance to the claim that 5 points are a good proxy of relative scene accuracy.

3.2 TAP-Flow: Uncontrolled benchmark

We depart from the high-quality TAP-Vid-DAVIS point-tracking benchmark [9]. This dataset contains 30 videos with up to 5 objects with 5 point annotations per object labelled.

Optical Flow Repurposing. The original TAP-Vid dataset [9] includes varying lighting, large displacements, varied scenery and a diverse flow field, and thus is well-suited for evaluating real-world optical flow generalisation. Because TAP-Vid consists of sparse point tracks rather than flow fields, we repurpose it for optical flow evaluation, we use all the 30 TAP-Vid-DAVIS [9] videos and pick the points that appear in both frames, so no occluded points are used, and use an interval between frames ranging from 3 to 6 for every frame. Resulting in 3757

frames across 30 videos. This benchmark allows for a balanced marker of real-world accuracy. These sparse annotations are exported in KITTI-style format, enabling standardised optical flow evaluation.

3.3 Slow Flow: uncontrolled benchmark

We also introduce the Slow Flow dataset [18]. Slow Flow, has 3 different categories, in which the 90th percentile displacement is 100, 200, and 300 pixels, with blur lengths of 0, 1, 3, 5, 9 frames. This results in 3447 dense ground-truth frames [18]. Thereby being complementary to the TAP-Flow [9] dataset as TAP-Flow does not contain motion blur and occlusions, leading to a more diverse benchmark.

4 Evaluation methodology

4.1 Model selection

We use the models provided by the PTLFlow framework [29]. To establish correctness of the implementation we replicate reported benchmark accuracies on the training set of both KITTI [28] and Sintel [7]. We eliminate models which differ in our reproduction more than 2.0 epe on Sintel [7] or 2.0 Fl-all on KITTI [28], than reported by their respective authors. Because of this reason we eliminated the HD3 model, [44], the RAFT_small model [36], and the MaskFlowNet_small model [47]. The deviations between reported errors by their authors and our implementation can be found in the supplemental material.

4.2 Checkpoint selection

We use the Things (C+T) [36, 48] checkpoint for all models corresponding to a training regime of first training on FlyingChairs [10] and then on FlyingThings3D [26]. The checkpoints are provided by the authors of their respective papers and hence training conditions may have differed. Training all models is however infeasible and would take up substantial computational resources. We use the Things checkpoint because it's the checkpoint that nearly every model provides and is the one most used to measure OOD generalisability [16, 32, 36, 40].

4.3 Error metric

We use the Fl-all metric propagated by the authors of KITTI [28]. Fl-all counts a pixel as correct if the Euclidean distance is less than 3 pixels or 5% to the ground truth pixel. The Fl-all score is computed by the PTLFlow framework [29]. The rationale for the Fl-all error is to account for human error. This is based on the assumption that human error falls within a range of 3 pixels, see Fig. 4. And that to have a more specific metric, such as the End-Point-Error which is the Euclidean distance, would suggest a level of detail that is not present in the ground truth. Since we also make use of human annotations, we apply the Fl-all metric consistently across all of our results.

5 Experiments

5.1 Exp. 1: How well does optical flow quality on current benchmarks generalise to real world optical flow quality?

We first evaluate all models on three real-world benchmarks: FlowFactor, Slow Flow [18] and TAP-Flow [9] to obtain a balanced estimate of real-world accuracy, the results are shown in Figure 5. When considering all models, the overall trend is downward on all benchmarks, indicating that a lower error on KITTI [28], Sintel [7] and Spring [27] generally results in a lower error on real-world datasets.

However, when we restrict our attention to more modern models that outperform the seminal RAFT [36] method on the training sets of KITTI [28], Sintel [7], and Spring [27] a different picture emerges: these stronger current-benchmark models tend to stagnate or even perform worse on real-world data (Fig. 7). We choose RAFT as a reference point because after that more and more models have implemented attention-based features and increased model capacity, which has repeatedly been observed to increase the risk of overfitting on narrow synthetic distributions [42]. As the displacement distributions of FlyingChairs and Sintel are closely aligned [25], we hypothesise that higher accuracy on the Sintel training set further may come at the cost of accuracy on other motion distributions, such as those in our real-world benchmarks.

Overall, we find that real-world accuracy has stagnated despite continued improvements on KITTI [28], Sintel [7], and Spring [27]. This is consistent with Bauer *et al.* [6], who observed a similar trend for a mostly synthetic uncontrolled setting. These results call into question whether KITTI [28], Sintel [7] and Spring [27] are still sufficient markers of generalisability: are the latest reductions in error on these benchmarks indicative of real-world progress, or do they mainly reflect increased specialisation to these domains? In Fig. 6 we show a qualitative comparison between models, where we show that SEA-RAFT [40], a model that scores strongly on benchmarks such as Spring, fails to generalise to the real world.

5.2 Exp 2: isolated categories vs real-world datasets

We next investigate which factors in the FlowFactor dataset most strongly correlate with real-world accuracy on Slow Flow [18]. In Fig. 8 we show the Spearman rank correlation coefficient between accuracy on each isolated FlowFactor category and accuracy on Slow Flow [18]. The varying-lighting category and the large-displacement category exhibit the strongest positive correlations with Slow Flow accuracy, indicating that these two factors are particularly predictive of real-world behaviour. Notice that some models deviate from the trend as they tend to overfit on certain aspects. We show that in general, all our categories positively correlate with real-world accuracy. We also show that the confounders correlate similarly with the TAP-Flow dataset; this can be found in the supplementary material.

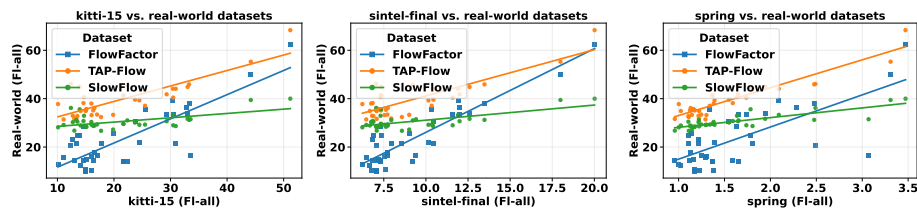


Fig. 5: Correlation between real-world datasets and KITTI [28], Sintel-Final [7] and Spring [27] (all models). We plot the FI-error of every Things model and the Ordinary Least Squares (OLS) fit on the KITTI [28], Sintel-Final [7] and Spring [27] training sets against three real-world benchmarks: FlowFactor, Slow Flow [18] and TAP-Flow dataset [9]. The overall downward trend indicates that, when considering all models, these benchmarks still serve as a reasonable proxy for estimating real-world applicability.

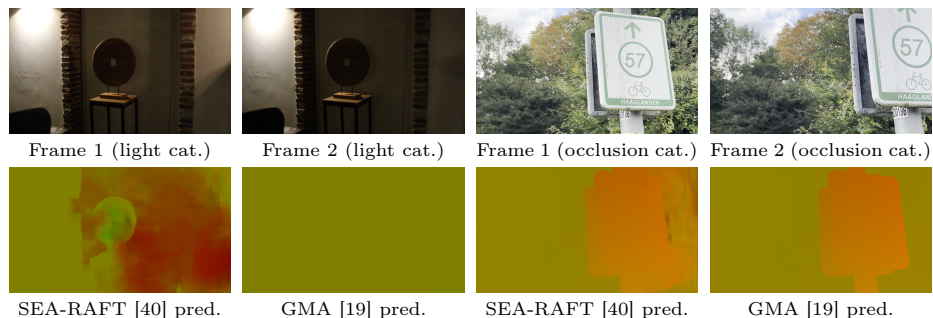


Fig. 6: Qualitative analysis. The GMA model [19] demonstrates superior stability under conditions of extreme photometric volatility. Specifically, it maintains a near-zero motion field when subjected to drastic lighting changes, whereas the SEA-RAFT small model [40] exhibits sensitivity to these variations, erroneously predicting motion. A similar performance gap is observed in the occlusion category. GMA [19]: Robustly leverages global motion aggregation to mitigate the loss of local visual cues. SEA-RAFT small [40]: Undergoes significant performance degradation when spatial correspondence is interrupted. These discrepancies in real-world robustness are notably obscured by standard evaluation metrics. On the KITTI benchmark [28], the primary metric for assessing real-world generalisation, both GMA [19] (C+T) and SEA-RAFT small [40] (C+T) yield comparable accuracy scores. This suggests that current benchmarks may lack the diversity required to expose the underlying vulnerabilities of certain architectures.

5.3 Exp. 3: Inter-correlation of isolated categories

To study how different factors interact across models, we compute the Spearman rank correlation between model rankings for each pair of FlowFactor categories. This allows us to understand whether models that perform well on one factor tend to perform well or poorly on another. The results are shown in Fig. 9.

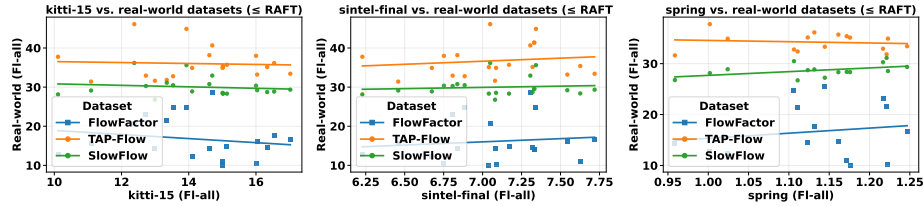


Fig. 7: Correlation between real-world datasets and KITTI, Sintel-Final and Spring for models outperforming RAFT. We plot the F1-error of every Things model on KITTI [28] Sintel-Final [7] and Spring [27] training sets against FlowFactor, Slow Flow [18] and TAP-Flow [9], restricting to models that surpass RAFT [36] on KITTI [28], Sintel [7] and Spring [27]. The trendline is the OLS fit. While newer architectures achieve lower errors on these benchmarks, real-world accuracy stagnates or even degrades. This suggests that below RAFT-level F1-all errors, current benchmarks no longer reflect real-world progress, highlighting the need for stronger real-world generalisation benchmarks.

The left heatmap uses all models and reveals that, globally, categories are positively correlated: models that are strong on one factor tend to be strong on others as well. The right heatmap is restricted to models that appear in the top-10 for at least one category. In this high-performing regime, we observe that the displacement- and repetitive-texture categories correlate negatively with the lighting and occlusion categories. This suggests a trade-off: models that are highly optimised for large displacements, as in the repetitive textures and large displacements category, tend to perform worse on the small-displacement, near-stationary scenes used for lighting and occlusion. In other words, robustness to photometric distortions in stationary scenes can conflict with challenging motion patterns, hinting at an inherent tension between robustness and accuracy.

5.4 Exp. 4: To what extent does extra training data reduce the generalisation gap to the real world?

Throughout this paper we have used the same C+T things checkpoints to keep a fair comparison between models. However, in the real-world one could use as much data as one would like. Recently, models have started to incorporate the synthetic dataset TartanAir [38] into their pre-training regime [39, 40, 46], which contains more than 1 million frame pairs. Others have changed the entire protocol by now mixing different labeled and unlabeled datasets into their training regime [23]. Does this increase in pre-training dataset size help with generalisation to the real-world? In Fig. 10 we show that extra synthetic data on top of C+T does not necessarily improve performance. The models sporting the highest accuracy on FlowFactor use either C+T or C+T in combination with a small amount of more realistic data (C+T+S+K+H). We show a similar pattern for the other 2 datasets in the supplementary material.

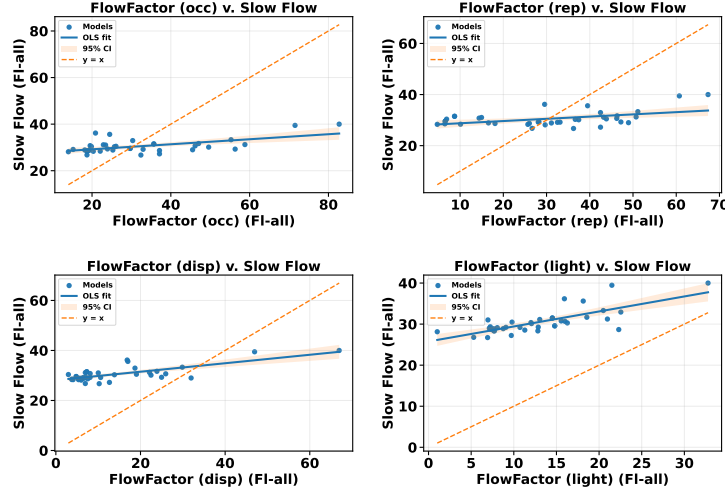


Fig. 8: Correlation between Slow Flow and isolated categories in FlowFactor. Each panel shows the correlation between accuracy on a single FlowFactor category and accuracy on Slow Flow [18]. All categories exhibit positive correlation, but lighting ($\rho = 0.75$) and large displacements ($\rho = 0.74$) correlate most strongly, suggesting that improving robustness to these factors is particularly important for real-world accuracy.

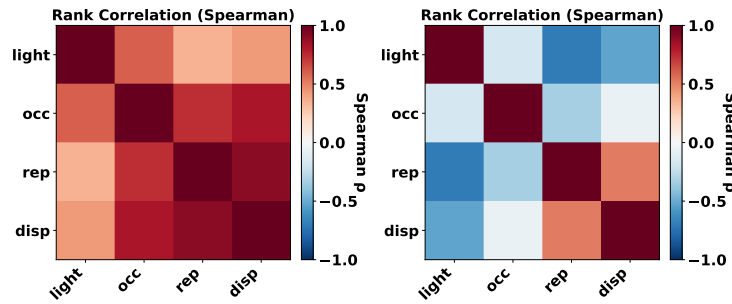


Fig. 9: Correlation between rankings on individual categories of the FlowFactor dataset. Spearman rank correlation between model rankings on each pair of FlowFactor categories. Left: all models. Right: restricted to models that appear in the top-15 for at least one category, highlighting emerging trade-offs between displacement-heavy and stationary, distortion-based factors. Showing that there is no clear model, that works in every scenario.

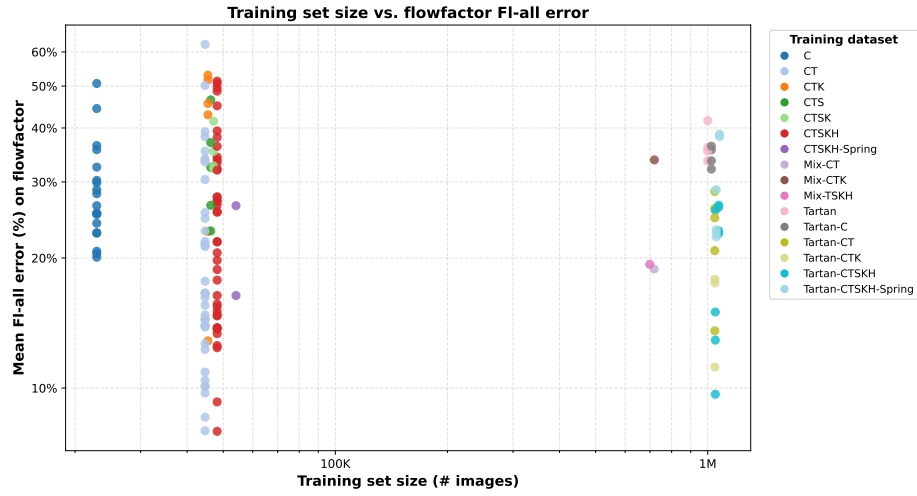


Fig.10: Size training dataset vs. accuracy on FlowFactor. Here we plot for all models all possible checkpoints. For every checkpoint we plot the accompanying training dataset size and accuracy on our real-world dataset FlowFactor. Increasing the pre-training dataset size with more synthetic data, does not necessarily improve accuracy on the FlowFactor dataset, compared to C+T training.

6 Conclusion

We created a factorised real-world dataset containing 1,000 frame pairs. We showed that the generalisability of recent models to the real world has stagnated and that current benchmarks do not measure real-world accuracy. We further demonstrated that accuracy across all factors in our factorised dataset correlates positively with accuracy on real-world benchmarks. In addition, we found that high accuracy on large displacements and repetitive textures correlates negatively with accuracy under varying lighting conditions and occlusions. Furthermore, we show that training on extra synthetic data does not necessarily reduce the generalisability gap to the real world.

Acknowledgements This project is supported by NWO (project VI.C.242.088). We would like to thank the following people for their contributions: Iris Petre, Jesse Klijnsma, Zhuoyue Ge, Marijn Timmerije, and Sachhyam Dahal. Furthermore, we would like to thank the 5 people that participated in the annotation survey.

Environmental statement During this study, an NVIDIA A40 was used for approximately 553.5 hours, resulting in an estimated 44.86 kg CO_2 -equivalent. These emissions were offset via certified (re)forestation credits ¹.

References

1. Robust vision challenge. <http://www.robustvision.net/>, accessed: 2025-11-13
2. Agnihotri, S., Caspary, J.Y., Schwarz, L., Gao, X., Schmalfluss, J., Bruhn, A., Keuper, M.: Flowbench: Benchmarking optical flow estimation methods for reliability and generalization. *Transactions on Machine Learning Research* (2025)
3. Alfarano, A., Maiano, L., Papa, L., Amerini, I.: Estimating optical flow: A comprehensive review of the state of the art. *Computer vision and image understanding* **249**, 104160 (2024)
4. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. *International journal of computer vision* **92**(1), 1–31 (2011)
5. Barron, J.L., Fleet, D.J., Beauchemin, S.S.: Performance of optical flow techniques. *International journal of computer vision* **12**(1), 43–77 (1994)
6. Bauer, K., Bruhn, A., Schmalfluss, J.: On the generalization of optical flow: Quantifying robustness to dataset shifts
7. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: *European conference on computer vision*. pp. 611–625. Springer (2012)
8. Dahal, S.: Going against the flow: Evaluating optical flow estimation models on real-world non-rigid motion (2025), <https://resolver.tudelft.nl/uuid:ab2db1e7-9122-43c3-9377-f188fceb30b6>
9. Doersch, C., Gupta, A., Markeeva, L., Recasens, A., Smaira, L., Aytar, Y., Carreira, J., Zisserman, A., Yang, Y.: Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems* **35**, 13610–13626 (2022)
10. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: FlowNet: Learning optical flow with convolutional networks. In: *Proceedings of the IEEE international conference on computer vision*. pp. 2758–2766 (2015)
11. Eslami, N., Arefi, F., Mansourian, A.M., Kasaei, S.: Rethinking raft for efficient optical flow. In: *2024 13th Iranian/3rd International Machine Vision and Image Processing Conference (MVIP)*. pp. 1–7. IEEE (2024)
12. Ge, Z.: Real-world evaluation of optical flow with varying lighting conditions (2025), <https://resolver.tudelft.nl/uuid:7ad41253-902a-4d78-a86a-c7bee94b9e0c>
13. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: *2012 IEEE conference on computer vision and pattern recognition*. pp. 3354–3361. IEEE (2012)
14. Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., Fleet, D.J., Gnanapragasam, D., Golemo, F., Herrmann, C., et al.: Kubric: A scalable dataset generator. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 3749–3761 (2022)

¹ https://treesforall.nl/app/uploads/certificates/Certificaat_Trees_for_All_1615061d.pdf

15. Heeger, D.J.: Model for the extraction of image flow. *Journal of the Optical Society of America A* **4**(8), 1455–1471 (1987)
16. Huang, Z., Shi, X., Zhang, C., Wang, Q., Cheung, K.C., Qin, H., Dai, J., Li, H.: Flowformer: A transformer architecture for optical flow. In: *European conference on computer vision*. pp. 668–685. Springer (2022)
17. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2462–2470 (2017)
18. Janai, J., Guney, F., Wulff, J., Black, M.J., Geiger, A.: 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*. pp. 3597–3607 (2017)
19. Jiang, S., Campbell, D., Lu, Y., Li, H., Hartley, R.: Learning to estimate hidden motions with global motion aggregation. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 9772–9781 (2021)
20. Klijnsma, J.B.: Real-world evaluation of optical flow on repetitive patterns (2025), <https://resolver.tudelft.nl/uuid:dc709377-e116-48eb-91a5-4b6812902034>
21. Kondermann, D., Nair, R., Honauer, K., Krispin, K., Andrulis, J., Brock, A., Gusefeld, B., Rahimimoghaddam, M., Hofmann, S., Brenner, C., 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*. pp. 19–28 (2016)
22. Li, H., Luo, K., Liu, S.: Gyroflow: Gyroscope-guided unsupervised optical flow learning. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 12869–12878 (2021)
23. Liang, Y., Fu, Y., Hu, Y., Shao, W., Liu, J., Zhang, D.: Flow-anything: Learning real-world optical flow estimation from large-scale single-view images. *arXiv preprint arXiv:2506.07740* (2025)
24. Liu, C., Freeman, W.T., Adelson, E.H., Weiss, Y.: Human-assisted motion annotation. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1–8. IEEE (2008)
25. Mayer, N., Ilg, E., Fischer, P., Hazirbas, C., Cremers, D., Dosovitskiy, A., Brox, T.: What makes good synthetic training data for learning disparity and optical flow estimation? *International Journal of Computer Vision* **126**(9), 942–960 (2018)
26. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: 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*. pp. 4040–4048 (2016)
27. Mehl, L., Schmalfluss, J., Jahedi, A., Nalivayko, Y., Bruhn, A.: 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*. pp. 4981–4991 (2023)
28. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 3061–3070 (2015)
29. Morimitsu, H.: Pytorch lightning optical flow. <https://github.com/hmorimitsu/ptlflow> (2021)
30. Petre, I.A.: Performance of optical flow models on real-world occluded regions (2025), <https://resolver.tudelft.nl/uuid:c25761a4-939d-4677-b790-b3fb9bc4bc28>

31. Schmalfluss, J., Oei, V., Mehl, L., Bartsch, M., Agnihotri, S., Keuper, M., Bruhn, A.: Robustspring: Benchmarking robustness to image corruptions for optical flow, scene flow and stereo. *arXiv preprint arXiv:2505.09368* (2025)
32. Shi, X., Huang, Z., Li, D., Zhang, M., Cheung, K.C., See, S., Qin, H., Dai, J., Li, H.: Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 1599–1610 (2023)
33. Stürzel, F., Spillmann, L.: Perceptual limits of common fate. *Vision research* **44**(13), 1565–1573 (2004)
34. Sun, D., Vlasic, D., Herrmann, C., Jampani, V., Krainin, M., Chang, H., Zabih, R., Freeman, W.T., Liu, C.: Autoflow: Learning a better training set for optical flow. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10093–10102 (2021)
35. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 8934–8943 (2018)
36. Teed, Z., Deng, J.: 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*. pp. 402–419. Springer (2020)
37. Timmerije, M.: Bridging the gap: A real-world dataset and evaluation of optical flow models in large displacement scenarios (2025), <https://resolver.tudelft.nl/uuid:d7c4cc36-6582-47fc-94cd-cc0dc393725b>
38. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual slam. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 4909–4916. IEEE (2020)
39. Wang, Y., Deng, J.: Waft: Warping-alone field transforms for optical flow. *arXiv preprint arXiv:2506.21526* (2025)
40. Wang, Y., Lipson, L., Deng, J.: Sea-raft: Simple, efficient, accurate raft for optical flow. In: *European Conference on Computer Vision*. pp. 36–54. Springer (2024)
41. Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C.: Learning to detect motion boundaries. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2578–2586 (2015)
42. Weng, W., Zhu, X., Jing, L., Dong, M.: Attention mechanism trained with small datasets for biomedical image segmentation. *Electronics* **12**(3), 682 (2023)
43. Xu, H., Zhang, J., Cai, J., Rezaatofghi, H., Tao, D.: Gmflow: Learning optical flow via global matching. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 8121–8130 (2022)
44. Yin, Z., Darrell, T., Yu, F.: Hierarchical discrete distribution decomposition for match density estimation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 6044–6053 (2019)
45. Zhai, M., Xiang, X., Lv, N., Kong, X.: Optical flow and scene flow estimation: A survey. *Pattern Recognition* **114**, 107861 (2021)
46. Zhang, D., Wang, F., Pollefeys, M., Xu, H.: Megaflow: Zero-shot large displacement optical flow. *arXiv preprint arXiv:2603.25739* (2026)
47. Zhao, S., Sheng, Y., Dong, Y., Chang, E.I., Xu, Y., et al.: Maskflownet: Asymmetric feature matching with learnable occlusion mask. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 6278–6287 (2020)
48. Zhao, S., Zhao, L., Zhang, Z., Zhou, E., Metaxas, D.: Global matching with overlapping attention for optical flow estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 17592–17601 (2022)

49. Zheng, Y., Zhang, M., Lu, F.: Optical flow in the dark. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6749–6757 (2020)
50. Zhou, H., Chang, Y., Shi, Z., Yan, W., Chen, G., Tian, Y., Yan, L.: Adverse weather optical flow: Cumulative homogeneous-heterogeneous adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)