On the Generalization of Optical Flow: Quantifying Robustness to Dataset Shifts

Katrin Bauer * Andrés Bruhn Jenny Schmalfuss *
University of Stuttgart, Computer Vision Group
first.last@vis.uni-stuttgart.de

Abstract

Optical flow models are commonly evaluated by their ability to accurately predict the apparent motion from image sequence data. Though not seen during training, this evaluation data generally shares the training data's characteristics because it stems from the same distribution, i.e.*, it is in-distribution (ID) with the training data. However, when models are applied in the real world, the test data characteristics may be shifted,* i.e.*, out-of-distribution (OOD), compared to the training data. For optical flow models, the generalization to dataset shifts is much less reported than the typical accuracy on ID data. In this work we close this gap and systematically investigate the generalization of optical flow models by disentangling accuracy and robustness to dataset shifts with a new effective robustness metric. We evaluate a testbed of 20 models on six established optical flow datasets. Across models and datasets, we find that ID accuracy can be used as a predictor for OOD performance, but certain models generalize better than this trend suggests. While our analysis reveals that model generalization capabilities declined in recent years, we also find that more training data and smart architectural choices can improve generalization. Across tested models, effective robustness to dataset shifts is high for models that avoid attention mechanisms and favor multi-scale designs. Code is available at https://github.com/cv-stuttgart/OF-EffectiveRobustness.*

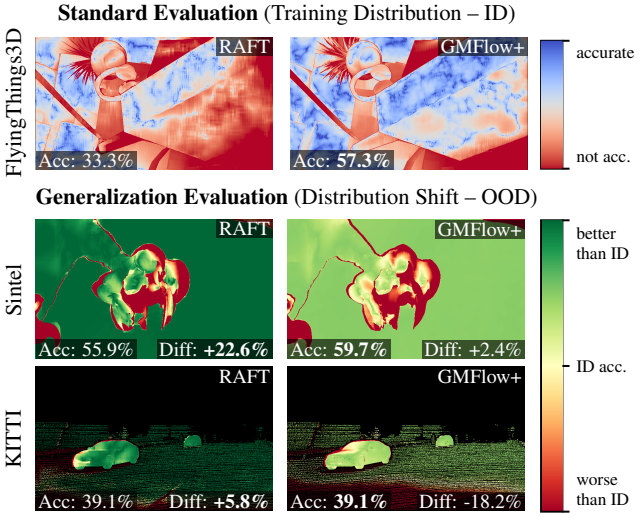

Figure 1. A high prediction accuracy on the training distribution does not imply good generalization. While GMFlow+ is more accurate than RAFT on the FlyingThings3D training data *[top]*, its quality stagnates or drops on new datasets relative to the training data *[middle, bottom]* (*Diff* low or negative). RAFT, however, improves on OOD data over its ID baseline (*Diff* positive), showing better generalization despite lower absolute accuracies. *Effective robustness* measures this performance difference on OOD data to the ID baseline performance. ID heatmaps *[top]* show EPE accuracy; OOD heatmaps *[middle, bottom]* show improvements or losses over average ID accuracy. Models use C+T checkpoints.

1. Introduction

Optical flow describes the motion in image sequences as a dense correspondence field between image pixels in subsequent frames. It is a low-level motion description used for a large variety of down-stream tasks including action recognition [59], video segmentation [65], robot navigation [18], and medical imaging [64]. As such, the optical flow estimations should be equally reliable on varying domains. The current literature evaluates optical flow models on various public benchmarks such as KITTI [37], MPI-Sintel [9] and Spring [36], which cover real-world and simulated scenes containing various challenges like changing illumination, blurs, and altering weather conditions.

While the abundance of existing optical flow datasets [9, 14, 31, 35–37, 48] creates an ideal testing ground for model generalization across domains, robustness comparisons across datasets are challenging for two main reasons: (i) incomparable performances for differently fine-tuned models and (ii) correlation of model performance across datasets. Firstly, new optical flow methods commonly report results *per dataset* after fine-tuning, which familiarizes the model with the dataset characteristics. For generaliza-

*Equal technical contribution.

Evaluation respects	Standard Evaluation [22, 24, 63, 72]	Robust Vision Challenge [26, 60]	Effective Robustness (ours)
(i) consistent finetuning	✗	✓	✓
(ii) ID-OOD correlation	✗	✗	✓

Table 1. Optical flow generalization: Evaluation challenges. Standard model evaluations use fine-tuned checkpoints per dataset and do not consider the ID-OOD performance correlation. Methods submitted to the robust vision challenge use a fixed ID checkpoint for all generalization evaluations, but only our effective robustness evaluation for optical flow corrects the ID-OOD correlation.

tion, however, one is interested in the model performance on completely shifted datasets which are out-of-distribution (OOD) with the training data, while the results of a model that was fine-tuned to a dataset are in-distribution (ID). Secondly, even if the same model checkpoint is used for the evaluation of multiple OOD datasets, the results on ID and OOD data are correlated, meaning that a model that did well on its training data distribution should typically perform well on a shifted dataset. For generalization, however, we search for models that can outperform this expectation on the shifted dataset, which is called *effective robustness* [62]. Fig. 1 illustrates the effect. Though RAFT [63] performs worse than GMFlow+ [72] on ID data, it maintains or improves its performance on the OOD data relative to its ID accuracy, meaning it generalizes well. In contrast, the OOD performance of GMFlow+ stagnates or drops compared to its ID accuracy, implying poor generalization.

In this work, we conduct the first evaluation of optical flow generalization that addresses both challenges: We explicitly select consistent model checkpoints and evaluate them with a carefully designed effective robustness metric to control for the ID-OOD correlation, which disentangles optical flow generalization from its accuracy.

Differences to Prior Generalization Studies. Prior generalization studies addressed the two challenges – consistent finetuning and ID-OOD correlation correction – to varying degrees, *cf*. Tab. 1. For optical flow, the *standard evaluation* evaluates new models after fine-tuning on training data of the respective dataset [12, 13, 22–27, 41, 61, 63, 69, 71], which addresses neither challenge. Many works additionally evaluate the performance of non-finetuned model checkpoints, but these studies are typically limited to few datasets including KITTI [37], Sintel [9], or special-property datasets [10], and furthermore disregard ID-OOD correlation. The *Robust Vision Challenge*[1] systematically investigates optical flow generalization by evaluating participating models with one consistent checkpoint for all evaluation datasets. However, the evaluation datasets are not truly OOD because training on their respective training splits is allowed, and the robustness evaluation uses model rankings across datasets, which does not correct ID-OOD correlation. Hence, no prior works evaluate optical flow generalization on true OOD dataset shifts in a manner that corrects the ID-OOD correlation.

Outside the optical flow domain, evaluations that account for the ID-OOD correlation were first introduced by Taori *et al*. [62] for classification. They propose *effective robustness* to identify models that outperform the linear correlation between accuracy on the training data (ID) and unseen evaluation data (OOD). In this work, we demonstrate an ID-OOD performance correlation for optical flow models, which motivates our optical flow adaptation of effective robustness to quantify generalization.

Contributions. In summary, we make four contributions.

(1) We systematically study the generalization to out-of-distribution dataset shifts for optical flow, demonstrating a linear correlation between ID and OOD accuracy.

(2) To separate generalization and accuracy, we propose *effective robustness for optical flow*, which quantifies robustness as improvement over the expected linear trend.

(3) We validate this new effective robustness formulation on 13 optical flow architectures (20 total variants) and across six diverse OOD evaluation datasets.

(4) Our analysis confirms that diverse training data improves generalization, but also finds that architectural details influence robustness more than pure model size.

2. Related Work

Traditionally, optical flow robustness focused on specific challenging conditions including varying illumination [30, 43, 75], large object motion [8, 70], occlusion [33], motion blur [45], image noise [7] or rain [34]. While these challenges remain, the advent of deep learning has led to more systematic studies on three robustness types: robustness to *adversarial attacks*, *image corruptions* and *dataset shifts*.

Adversarial Robustness. Adversarial robustness studies worst-case scenarios by optimizing image corruptions for a maximally perturbing effect on a model's performance. Several attacks were proposed for optical flow, ranging from local attacks with adversarial patches [46] over global [1, 51, 56] and object-constrained [32] image perturbations to attacks that simulate maximally distracting weather conditions [50, 52]. Adversarial attacks are a continuing security risk, as proposed defense strategies [3, 74] can be overcome by specialized attacks [49]. At the same time, for optical flow, the robustness to such attacks is already studied more extensively than other robustness types, and therefore not a focus of this work.

Robustness to Image Corruptions. Model performance can also strongly degrade under non-optimized

[1] www.robustvision.net

model-agnostic image corruptions. There are numerous image corruptions for general deep-learning tasks, including translations and rotations [15, 53], stylizations [17, 21], noises [19], blurs [19, 28, 42], or weather [19, 38, 52, 66]. Only few works study image corruptions on optical flow [2, 52, 54, 56], and most use the corruptions from Hendrycks and Dietterich [19]. While such synthetic corruptions can describe specific scenarios, robustness to image corruptions does not indicate robustness to natural distribution shifts, *i.e.*, other datasets, as observed for classification tasks [20, 62]. To study generalization of optical flow methods, we thus directly analyze robustness to dataset shifts rather than image corruptions.

Robustness to Dataset Shifts. To understand how the performance of optical flow methods generalizes to real-world scenarios without known ground-truth, we can assess accuracy changes across datasets. The Robust Vision Challenge is a notable step in this direction and compares model performance across various public benchmarks using the same model checkpoint. However, models can train on the benchmark training splits while our work focuses on analyzing generalization without fine-tuning. Also, it quantifies model robustness by rank relative to other models, which ignores the previously discussed correlation between ID and OOD accuracy. This correlation was discovered for classification tasks, where various works found an accuracy drop from in-distribution accuracy on ImageNet [11] to the out-of-distribution accuracy on challenging datasets [39, 47, 57]. Because the ID accuracy correlates with the accuracy on previously unseen data, Taori et al. [62] coined a model's ability to outperform this predictable trend as *effective robustness*. Later works extended the finding of predictable trends to more distribution shifts and model architectures [40, 73], unlabeled data [5], pre-trained models [4], models with varying training data [58] and multi-modal foundation models [67] – all for classification tasks. Here, we extend the concept of effective robustness to optical flow models to investigate their robustness to distribution shifts.

3. Effective Robustness for Optical Flow

This section establishes how we measure the generalization of optical flow models to data that is OOD compared to the training distribution. First, we empirically show that ID and OOD performance are correlated. Then we introduce effective robustness as a model's deviation from the expected trend. Finally, to transfer the effective robustness concept to optical flow, we identify a suitable accuracy measure and a baseline to describe the correlation.

Motivating Example. In Fig. 2 we compare the robustness to distribution shifts for optical flow models trained on Things [35], and evaluate them on the held-out test set of

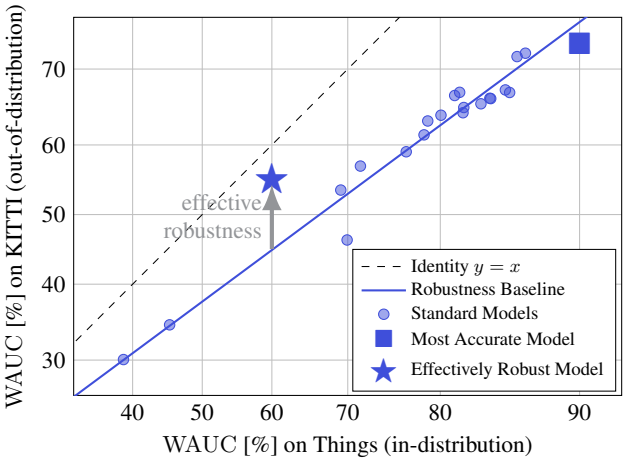

Figure 2. Model accuracies on a distribution shift are strongly correlated. The effective robustness controls for this correlation by measuring the gap between the predictable and observed out-of-distribution accuracy.

Things and KITTI [37], for ID and OOD evaluation, respectively. We see that the OOD performance of models (blue dots) follows from the ID performance with a predictable linear trend (blue line). The model marked with ★ performs worse on both datasets than the most accurate model on both datasets ■. Interestingly, ★ outperforms the trend of expected OOD performance based on its ID performance, meaning it generalizes better than ■ despite its worse accuracy on both datasets.

Effective Robustness Concept. Because of the strong correlation between ID and OOD accuracy as observed in the motivating example, comparing the plain OOD performances across models confounds robustness with accuracy. The effective robustness concept [62] decouples model robustness from accuracy by comparing the OOD accuracy to a robustness baseline β – the blue linear trend in Fig. 2 – which depends on the ID accuracy:

$$\mathrm{ER} = \mathrm{accuracy}_{\mathrm{OOD}} - \beta(\mathrm{accuracy}_{\mathrm{ID}}). \tag{1}$$

Models with effective robustness $\mathrm{ER} > 0$ generalize better to OOD data than the baseline trend suggests, while models with $\mathrm{ER} < 0$ do not generalize well. To adapt this generic definition to optical flow methods, we first discuss a suitable accuracy measure for $\mathrm{accuracy}_{\mathrm{ID}}$ and $\mathrm{accuracy}_{\mathrm{OOD}}$, and then define how to estimate the baseline β.

Accuracy Measure. As we analyze the robustness on various data distributions, the accuracy measure should take meaningful values on all of them. Moreover, an accuracy measure for optical flow that takes values in $[0, 1]$ improves resemblance to the effective robustness as introduced for image classification [62], where accuracies are typically in

$[0, 1]$. The WAUC [48], which was initially proposed for the VIPER benchmark, satisfies these two requirements. Unlike unbounded optical flow metrics like the end-point error EPE [6, 9], it takes values in a fixed range $[0, 1]$. At the same time, the WAUC is more flexible and fine-grained than single-threshold inlier rates IR_t [16, 36, 37] because it aggregates inlier rates over 100 thresholds with threshold-specific weights $w_k = 1 - \frac{k-1}{100}$. Its definition reads

$$\mathrm{WAUC} = \frac{1}{|w|} \sum_{k=1}^{100} w_k \cdot \mathrm{IR}_{\frac{k}{20}}, \tag{2}$$

where $\mathrm{IR}_{\frac{k}{20}} = \frac{1}{|\Omega|} \sum_{\boldsymbol{x} \in \Omega} [\![\|f_{\mathrm{gt}}(\boldsymbol{x}) - f(\boldsymbol{x})\|_2 \leq \frac{k}{20}]\!]$ is the inlier rate, f_{gt} the ground truth flow, f the predicted flow, $\Omega \subset \mathbb{R}^2$ the image domain and $|w| = \sum_{k=1}^{100} w_k$ a normalization factor. By definition, the WAUC takes values in $[0, 1]$, where larger values indicate a higher accuracy. Multiplying its value by 100 converts it to WAUC [%] with values in $[0, 100]$. Later in Tab. 2, we experimentally demonstrate a less noisy linear correlation between the ID and OOD accuracy when using the WAUC than when using the more commonly used accuracy measures EPE, IR_{1px}, IR_{3px} and IR_{5px}. This makes it the most suitable accuracy metric for optical flow generalization evaluations.

Robustness Baseline β. The robustness baseline formalizes the correlation between model accuracies on different datasets. As illustrated in Fig. 2, effective robustness compares a model's OOD accuracy to a robustness baseline which depends on the model's ID accuracy. This empirically fitted baseline is crucial to disentangle OOD robustness from model accuracy. We model the robustness baseline as a linear fit between the ID and OOD WAUC:

$$\beta(\mathrm{WAUC}_{\mathrm{ID}}, a, b) = \mathrm{expit}(a \cdot \mathrm{logit}(\mathrm{WAUC}_{\mathrm{ID}}) + b) \tag{3}$$

with parameters $a, b \in \mathbb{R}$, determined from the accuracies of all models in our testbed. Following the original approach by Taori *et al*. [62], we apply least squares linear regression in logit space. The logit transformation maps the the accuracies in $[0, 1]$ to the entire real axis which gives models with higher accuracy a higher weight in the regression than less accurate models. Additionally, it ensures robust models with ideal ID accuracy of 1 are also expected to have ideal OOD accuracy of 1. The transformation to logit space and its inverse are defined as $\mathrm{logit}(x) = \ln\left(\frac{x}{1-x}\right)$ and $\mathrm{expit}(x) = \frac{1}{1+\exp(-x)}$.

Effective Robustness for Optical Flow. With all those ingredients we define *effective robustness for optical flow* as the WAUC difference to the logit robustness baseline:

$$\mathrm{ER}_{\mathrm{WAUC}} = \mathrm{WAUC}_{\mathrm{OOD}} - \beta(\mathrm{WAUC}_{\mathrm{ID}}, a, b). \tag{4}$$

By taking the difference to the baseline trend, robustness is separated from model accuracy. Note that each distribution shift to a new OOD dataset requires re-fitting the baseline parameters a and b. The optical flow effective robustness takes values in $[-1, 1]$ for WAUC or $[-100, 100]$ for WAUC [%], where 0 indicates robustness that can be expected from the fitted baseline. Positive values indicate good robustness, negative values indicate bad robustness.

4. Experiments

The previously defined effective robustness metric enables a systematic analysis of optical flow generalization. In the following, we first describe the evaluated optical flow methods and datasets. Then, we systematically validate our choices on how to evaluate optical flow generalization, from ID and OOD datasets over the design of the effective robustness metric to its properties. With the identified setup, we finally evaluate generalization across all models, with a focus on architecture- and dataset-specific differences.

4.1. Models and Datasets

We evaluate generalization on a testbed of 13 optical flow models on six public datasets, which are described below.

Model Testbed. We compare 13 optical flow models, offering a total of 20 architectural variants as several models come with multiple versions, *cf*. Supp. Tab. A2. All models are state-of-the-art, supervised, and PyTorch-implemented.

Our analyzed models follow a curriculum learning schedule [24], which yields checkpoints from four training stages: First, pre-training on FlyingChairs [14]; Second, pre-training on FlyingThings3D [35]; Third, fine-tuning to Sintel [9] by combining data from FlyingThings3D, Sintel, KITTI [37] and HD1K [31]; And fourth, fine-tuning to and on KITTI. We refer to the training phases and resulting checkpoints as C, C+T, S, and K. The only exception to this curriculum is SEA-RAFT [69], which is additionally pre-trained on TartanAir [68] before following the curriculum schedule described above.

Evaluation Datasets. We evaluate the models on eight datasets from three different categories: Object datasets, movie datasets and automotive datasets. FlyingChairs (Chairs) [14] and FlyingThings3D (Things) [35] are large synthetic training datasets, featuring random 2D motion of chairs and 3D motion of diverse objects, respectively. Sintel [9] and Spring [36] are movie datasets extracted from short animated films with camera and character motion. KITTI [37], HD1K [31], Driving [35] and VIPER [48] are automotive datasets with scenes recorded by a car-mounted camera. Among those, KITTI and HD1K contain real camera feeds, while Driving and VIPER are rendered.

4.2. How to Evaluate Optical Flow Generalization

We evaluate optical flow generalization through effective robustness, which compares the OOD accuracy to a robustness baseline representing the predictable trend between the

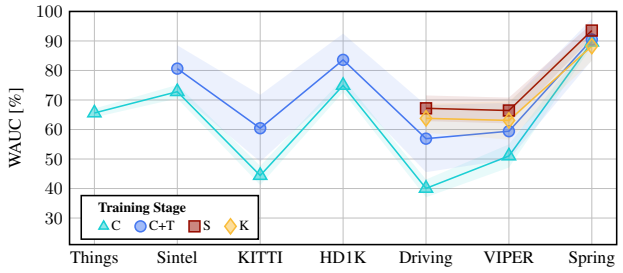

Figure 3. WAUC of our checkpoints averaged per training stage on all evaluated datasets. Each colored marker corresponds to the average over all model checkpoints at a fixed training stage. Sintel checkpoints achieve the best generalization accuracy.

ID and OOD accuracy. In the following, we justify choosing Things as ID dataset and validate the effective robustness design with its linear logit baseline and WAUC as accuracy metric. Then, we analyze how training data and training variability influence effective robustness.

Choosing ID and OOD Data. As we analyze model robustness to unseen data, we distinguish between in-distribution (ID) and out-of-distribution (OOD) data, which is similar to or distinct from the training data respectively. Rather than directly choosing ID and OOD data, we chose model checkpoints, which implicitly determine ID and OOD datasets. For a meaningful robustness analysis, models should have seen enough data to perform well, but too much training leaves no OOD data.

Fig. 3 visualizes the average OOD accuracy in WAUC [%] of our testbed models for each training stage. Note that early checkpoints, *e.g.* C+T, were not trained on many datasets and have all datasets except Things for OOD evaluation, while later checkpoints, *e.g.* S and K, were trained on most datasets and have only Driving, VIPER and Spring for OOD evaluation. On all datasets, the later checkpoints S and K have a better OOD performance than the earlier ones, because they were trained on more varied data. Interestingly, fine-tuning to KITTI degrades the OOD accuracy compared to the earlier S checkpoints. Across datasets, the per-checkpoint OOD performance varies strongly and indicates changing difficulties among the datasets. As even robust models rarely achieve the same WAUC across datasets, model robustness should only be compared per dataset.

Subsequently, we focus on the checkpoints trained on C+T, which yield good OOD accuracy despite being trained on limited data. Hence, Things is our typical ID dataset, which leaves Sintel, KITTI, HD1K, Driving, VIPER and Spring as the six OOD datasets.

Effective Robustness: Why Fit a Baseline? While the accuracy across datasets is not equal, it is correlated. This correlation is visualized in Fig. 4, which compares the per-checkpoint WAUC on the ID Things dataset to the WAUC

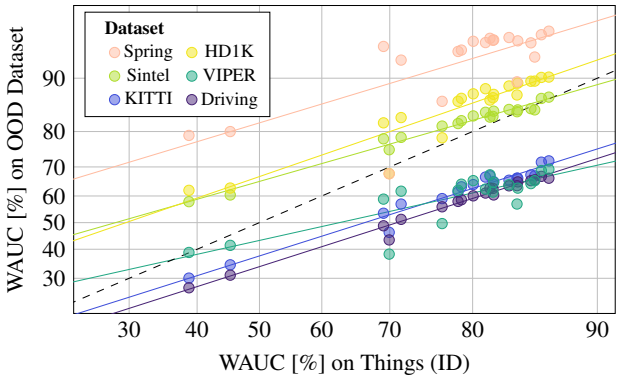

Figure 4. WAUC on each OOD dataset over the WAUC on Things for various models. Each point corresponds to one model checkpoint evaluated on one dataset. The color indicates the OOD dataset. For each dataset shift, the WAUC follows a linear trend. Lines denote the logit-linear fit for each dataset. Fitted baseline parameters are in Supp. Tab. A3.

on each of the OOD datasets. Here, the ID WAUC is a strong predictor for the OOD WAUC. Thus, comparing the plain OOD WAUC of models mixes model accuracy and generalization ability. We disentangle these two properties with the effective robustness, which compares the OOD WAUC of a model to the predictable trend that is modelled via a fitted robustness baseline. The robustness baselines per dataset are shown as colored lines in Fig. 4. A model's effective robustness is its slight deviation from the baseline.

Effective Robustness: Which Accuracy Measure? The observed correlation between ID and OOD performance is not equally strong for all optical flow accuracy measures. Tab. 2 quantifies the correlation between the ID and OOD accuracy for the popular measures WAUC, Inlier Rate (IR) with fixed thresholds and EPE. The Pearson correlation [44] measures linear correlation in $[-1, 1]$, where 1 indicates perfect linear correlation. Compared to all other accuracy measures, the WAUC yields the highest ID-OOD Pearson-correlation across datasets, and confirms our quantitative observations from Fig. 4. A strong correlation not only justifies fitting a baseline but also reduces the baseline's variance. This validates our choice of the WAUC as accuracy measure for optical flow effective robustness.

Influence of Model Training Data. Having established effective robustness for optical flow, we now investigate how model training data influences the OOD generalization and baseline trend. To this end, we visualize the ID-OOD correlation for different training stages, *i.e.*, for models trained on varying amounts of data. Fig. 5 shows the results on Things as ID and Driving as OOD data, results for other OOD data are in Supp. Fig. A2. Across datasets, OOD accuracy improves with more training data, *i.e.*, late checkpoints like S and K generalize better than the C+T checkpoint, which is a common notion in optical flow training. For each train-

Metric	**Correlation** of Things with					
	Sintel	KITTI	HD1K	Driving	VIPER	Spring
WAUC	0.993	**0.977**	**0.931**	**0.982**	**0.819**	0.667
IR_{1px}	**0.996**	0.954	0.928	0.964	0.739	0.369
IR_{3px}	0.995	0.947	0.891	0.966	0.769	**0.745**
IR_{5px}	0.993	0.936	0.870	0.944	0.764	0.706
EPE	0.807	0.884	0.793	0.850	0.459	0.043

Table 2. Pearson correlation [44] between the error of models on Things and each of the other datasets for the WAUC, inlier rate IR with thresholds 1, 3, and 5 pixels and average end-point error EPE, using C+T checkpoints. A Pearson correlation of 1 indicates a perfect linear correlation. A high correlation justifies the linear fit that we use as robustness baseline.

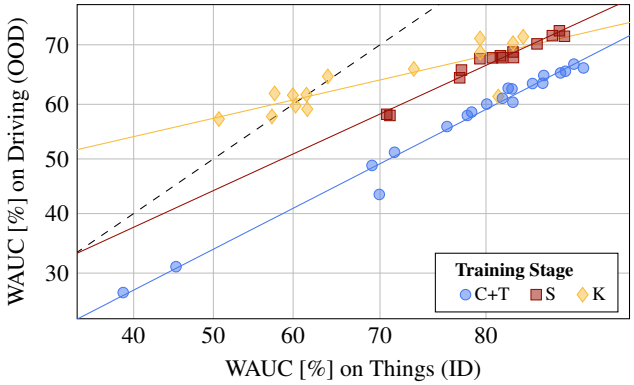

Figure 5. WAUC on Driving over the WAUC on Things for various models. Each point corresponds to one model checkpoint. The color indicates the training stage. Note that not all architectures offer a checkpoint at each training stage. Lines denote the logit-linear fit to the WAUC of model checkpoints at a fixed training stage. Model checkpoints trained on different data follow different trends. Fitted baseline parameters are in Supp. Tab. A3.

ing stage, the ID-OOD WAUC distribution follows a linear trend while checkpoints at different training stages follow different trends. Between the training stages C+T and S, the trend is offset with S checkpoints having a higher baseline robustness than the C+T one. Hence, we only fit effective robustness baselines for models trained on the same data, which separates robustness gains through model architecture from training data. Following our previous choice of Things as ID dataset, we only fit effective robustness baselines for the corresponding C+T model checkpoints.

Influence of Model Training Variability. The effective robustness is a model's accuracy deviation from the robustness baseline. Here, we investigate if this deviation – and thus effective robustness as metric – is stable and meaningful, given that model accuracies can vary depending on training randomness. To this end, we compare multiple instances of the RAFT architecture [63], all trained on Things but with varying batch sizes, learning rates, weight decays

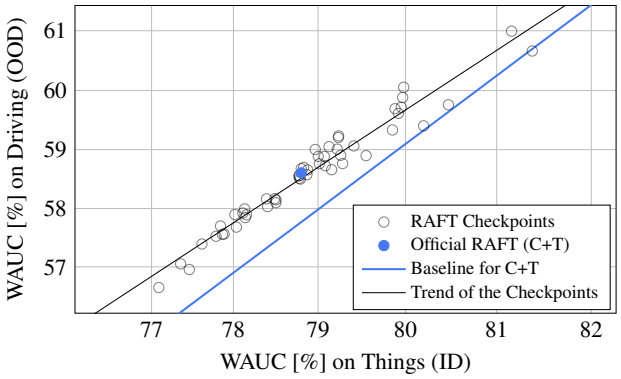

Figure 6. Accuracy of several RAFT checkpoints trained on Things. The robustness baseline is given by the logit-linear fit to the public C+T models (same blue line as in Fig. 5).

and training steps, to simulate training randomness. Fig. 6 shows the accuracies of those RAFT checkpoints on Things (ID) and Driving (OOD). While the checkpoint accuracies vary due to training randomness, they follow a linear trend that is nearly parallel to the robustness baseline. Therefore, all trained instances of RAFT have a similar effective robustness to the official RAFT checkpoint, even though their OOD accuracies vary. In Supp. Fig. A4, we show analogous results for other OOD datasets. While, in case of Sintel, KITTI and VIPER, the RAFT checkpoint trend does not completely parallel the robustness baseline, the robustness variance among checkpoints with similar accuracy is small, which still indicates the representativeness of effective robustness across model instances. Overall, this shows that effective robustness accounts for model accuracy variations due to model training variability, and demonstrates its expressiveness and stability.

4.3. Optical Flow Generalization Evaluated

After establishing effective robustness as an evaluation tool for optical flow generalization in the previous section, we now use it to analyze the OOD robustness in detail. Following the last section, we use model checkpoints after training on C+T, with the validation split of Things as ID dataset. Note that models are trained on Things' training set but not the validation set. Effective robustness scores for all model architectures across the six OOD datasets are listed in Tab. 3. Generalization is good for positive and bad for negative scores. By construction, effective robustness is uncorrelated with ID accuracy, *i.e.*, complexity differences among OOD datasets are equalized. Since models are trained on the same data, scores exclusively express differences in architectural robustness. All subsequent analysis is based on the data in Tab. 3. The first part focuses on optical flow model architectures, the second investigates OOD dataset similarities.

Method	Accuracy WAUC [%]	Effective Robustness ER_{WAUC} [%]							
	Things	Sintel	KITTI	HD1K	Driving	VIPER	Spring	Average	STD
FlowNet2S	38.73	−0.02	0.06	**3.86**	0.43	1.33	2.29	1.33	1.39
FlowNet2C	45.26	−1.87	0.13	−0.49	0.28	0.59	0.03	−0.22	0.81
FlowNet2	69.14	**1.89**	1.29	2.58	0.45	5.12	**4.61**	**2.66**	1.70
IRR-PWC	71.53	0.91	**2.63**	2.26	0.75	**6.53**	2.48	2.59	1.91
RAFT	78.80	0.64	1.84	1.56	0.84	4.03	1.22	1.69	1.12
GMA	78.44	0.22	0.33	1.25	0.63	2.14	1.20	0.96	0.66
SKFlow	80.06	0.91	1.28	1.56	0.87	4.32	1.62	1.76	1.18
MemFlow	83.52	0.09	−1.14	0.69	0.34	0.04	0.86	0.15	0.65
RPKNet	81.74	−0.36	2.42	−0.53	**1.71**	4.94	1.35	1.59	1.84
FlowFormer	82.03	−0.39	−0.58	−0.29	1.25	−1.32	1.10	−0.04	0.92
MatchFlow (R)	81.30	0.90	2.51	1.62	0.46	0.37	1.13	1.16	0.73
MatchFlow (G)	82.10	0.64	0.04	0.23	−1.18	2.13	1.10	0.49	1.01
GMFlow+ (s1)	69.92	−1.52	−6.65	−12.11	−5.58	−15.96	−21.47	−10.55	6.74
GMFlow+ (s2)	76.63	0.71	−0.19	−5.59	0.52	−8.98	−5.00	−3.09	3.66
GMFlow+	84.22	−0.80	−1.26	−1.16	−0.51	−7.58	−4.22	−2.59	2.54
SEA-RAFT (S)	84.29	−0.48	−1.39	0.64	0.73	−1.84	0.34	−0.33	0.99
SEA-RAFT (M)	85.39	−0.95	−1.64	0.24	−0.30	−1.26	−0.17	−0.68	0.66
SEA-RAFT (L)	85.69	−1.38	−2.36	0.07	−0.44	−0.53	−1.63	−1.04	0.82
MS-RAFT+	86.21	0.39	1.41	0.26	0.02	2.37	0.23	0.78	0.84
CCMR+	**86.78**	0.24	1.05	−0.05	−1.41	2.22	0.32	0.40	1.10

Table 3. Effective robustness score w.r.t. WAUC on different datasets. The baseline was fitted to all models listed in the table. Rows are roughly grouped by architectural similarity. Negative scores (bad robustness) are marked with a gray background. Best and second-best per column are **bold** and underlined respectively.

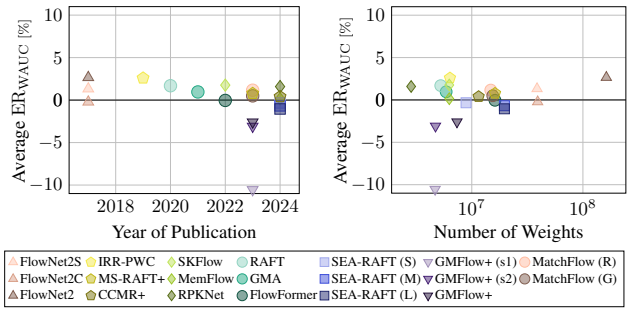

Figure 7. Effective robustness of each model over its publication year *[left]* and its number of weights *[right]*. Robustness decreased or stagnated the last years. The model size is not indicative of its robustness.

4.3.1. Influence of Model Architecture

In the following, we analyze generalization and its relation to model architecture. Initially, we assess broad model properties like publication date and size. Then, we focus on specifics, such as attention and receptive field sizes.

Model Publication Date and Size. In Fig. 7 we visualize the average effective robustness of all models with their publication date *[left]* and size *[right]*. Interestingly, the effective robustness of models stagnated or declined for the last half-decade, indicating no automatic generalization improvements for architectures that advance ID accuracy. Regarding model size, there is no apparent correlation with effective robustness. The smallest and largest models, RPKNet and FlowNet2, are both effectively robust. Hence, broad model properties like size as proxy for learning capacity do not explain generalization differences.

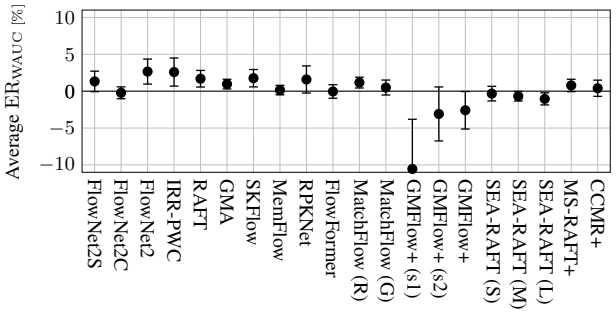

Figure 8. Average effective robustness and standard deviation of each model. Models with global motion aggregation (GMA, MatchFlow (G)) are less robust than similar models without it (RAFT, MatchFlow (R)). Larger convolution kernels (SKFlow, RPKNet) and multi-scale approaches (IRR-PWC, MS-RAFT+) provide good robustness.

Attention Mechanism. Since broad architectural properties do not explain generalization differences, we focus on an increasingly popular concept: the attention mechanism. Our model selection offers several model pairs with architectures that use and do not use attention: RAFT [63] and GMA [27], as well as MatchFlow (R) and MatchFlow (G) [13] are distinct through the insertion of the GMA module in the iterative refinement steps, and CCMR+ extends MS-RAFT+ mainly by adding an attention module. Fig. 8 visualizes the average effective robustness. In all three model pairs, attention or GMA modules reduce the effective robustness. Furthermore, transformers like FlowFormer [22] and GMFlow+ [72], which rely on self-attention, also have low effective robustness scores. Interestingly, SKFlow [61] also includes the GMA module but reaches a high effective robustness. We attribute this to its increased receptive field, resulting from larger convolution kernels, see below. There are two potential reasons for the observed poor generalization with attention: Global attention may distract models by attending to far-away details, and require more training data to generalize well.

Receptive Field Sizes. While attention seems to hinder generalization, we now turn to two potentially beneficial concepts: Large convolutions and multiscale architectures, which both effectively increase the receptive field. SKFlow and RPKNet apply larger convolution kernels than RAFT, improving their accuracy while preserving RAFT's good effective robustness, *cf*. Fig. 8. Also, the multi-scale methods IRR-PWC and MS-RAFT+ achieve high robustness scores within the testbed. Possibly related, the GMFlow+ variant (s2) improves the effective robustness of variant (s1) by adding a hierarchical matching refinement at a smaller scale. Models with large convolutions and multiple scales can process more context without distracting distant details and, hence, tend to improve effective robustness.

4.3.2. Optical Flow Dataset Similarities

After focusing on model architectures during the first part of this analysis, we finally turn to dataset-related aspects within our results. While the effective robustness baseline aims to remove ID-OOD correlation within the results, the effective robustness in Tab. 3 of each model still varies across datasets, *e.g.* FlowFormer is effectively robust on Driving and Spring, but less so on Sintel, KITTI, HD1K and VIPER. This inspires an analysis of dataset similarity to identify groups of OOD datasets that elicit similar generalization behavior from the tested models.

In Tab. 4 *[left]* we investigate whether the effective robustness rankings of models are correlated across datasets. We use Kendall's Tau correlation coefficient [29] as similarity measure, as it compares rankings rather than absolute values, making it more robust against outliers. Overall, there is a moderate correlation between the dataset's effective robustness rankings. However, there are three dataset pairs with stronger correlation, *i.e.*, where Kendall's Tau is larger than 0.6 – Spring with HD1K, VIPER with KITTI, and VIPER with Spring. Since VIPER and KITTI are automotive datasets, we exemplarily pick Spring and HD1K, where the scene content is unalike and hence does not explain the dataset correlation, for investigation. In this case, an explanation may be the distribution of flow vector length in each dataset, *cf*. Tab. 4 *[right]*, which visualizes the cumulative flow length distribution per dataset. There, two of the dataset pairs have very similar flow length distributions, with many long motion vectors in VIPER and KITTI, and predominantly short vectors in Spring and HD1K.

Overall, while dataset pairs with high correlation could be used as generalization-proxies for one another, the moderate values for the correlation coefficients suggest that the considered data sets only share some of their properties.

4.3.3. Discussion of Robustness Concepts

Previous literature discussed several different robustness concepts as outlined in Sec. 2. For optimized per-image perturbations (adversarial attacks) a trade-off between accuracy and robustness was observed [46, 51], which weakens for increasingly realistic perturbations [52]. For unoptimized realistic perturbations (*e.g.* the common corruptions by Hendrycks and Dietterich [19]), newer and more accurate methods are generally more robust [55]. This trend continues when transitioning from per-image corruptions to a joint optimization on different datasets in the Robust Vision Challenge. In contrast, our analysis finds neither a trade-off nor a systematic association between effective robustness and accuracy, thereby enabling comparisons of architectural concepts that are not confounded by model accuracy. For a more detailed comparison, we refer to Appendix B.

	Sintel	KITTI	HD1K	Driving	VIPER	Spring
Sintel	1					
KITTI	0.57	1				
HD1K	0.46	0.37	1			
Driving	0.25	0.35	0.26	1		
VIPER	0.47	**0.65**	0.46	0.27	1	
Spring	0.48	0.54	**0.66**	0.45	0.65	1

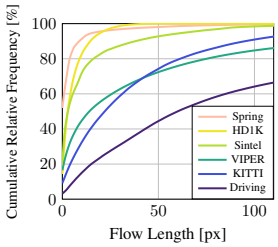

Table 4. Kendall's Tau correlation coefficient [29] comparing the ranking of models by the effective robustness on different datasets. That is, each entry in this table indicates the similarity between two columns of Tab. 3. Values are in the range $[-1, 1]$ where -1 indicates negative correlation, 0 no correlation and 1 positive correlation. The right figure shows the cumulative frequency of flow lengths in each dataset. Spring and HD1K yield the most similar rankings and also contain a similar length distribution.

5. Conclusion

We systematically analyzed the out-of-distribution generalization of current optical flow models to new datasets. For models trained on the same data distribution, we find a strong linear correlation between the ID and OOD accuracy. Therefore, we introduced the notion of effective robustness for optical flow that defines robustness as an improvement over the predictable trend and thus disentangles OOD robustness from accuracy. We experimentally validated our choices of ID and OOD datasets, the used model checkpoints, and our effective robustness definition for optical flow. Our evaluation of 20 optical flow architectures across six OOD datasets uncovers that the generalization capabilities of modern architectures steadily declined in recent years, independent of overall model size. However, we also find that improving robustness is possible with increased training data variety and smart architectural choices that avoid global motion aggregation (GMA) modules and favor large kernels or multi-scale concepts.

Limitations. A first limitation of this study is the moderate size of our optical flow model testbed. While we demonstrated a clear ID-OOD accuracy correlation, the limited testbed size can pronounce randomness in our logit-linear fit, and may skew the resulting effective robustness score. A second limitation is the baseline fit. As we only find a linear trend for models with the same ID, OOD and training data, any data change requires fitting a new baseline, potentially limiting the metric's scope. Finally, the effective robustness improves for reduced ID accuracy and maintained OOD accuracy. Hence, comprehensive model evaluations should always report plain accuracy alongside robustness.

Acknowledgements. Katrin Bauer and Andres Bruhn acknowledge support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 533085500. Jenny Schmalfuss thanks the International Max Planck Research School for Intelligent Systems IMPRS-IS.

References

[1] Shashank Agnihotri, Steffen Jung, and Margret Keuper. CosPGD: an efficient white-box adversarial attack for pixel-wise prediction tasks. In *Proc. International Conference on Learning Representations (ICML)*, pages 416–451, 2024. 2

[2] Shashank Agnihotri, Julian Caspary, Luca Schwarz, Xinyan Gao, Jenny Schmalfuss, Andrés Bruhn, and Margret Keuper. FlowBench: a robustness benchmark for optical flow estimation. *https://openreview.net/forum?id=S4jzvOBs9m*, 2025. 3

[3] Adithya Prem Anand, H. Gokul, Harish Srinivasan, Pranav Vijay, and Vineeth Vijayaraghavan. Adversarial patch defense for optical flow networks in video action recognition. In *Proc. IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1289–1296, 2020. 2

[4] Anders Andreassen, Yasaman Bahri, Behnam Neyshabur, and Rebecca Roelofs. The evolution of out-of-distribution robustness throughout fine-tuning. *arXiv preprint arXiv:2106.15831*, 2021. 3

[5] Christina Baek, Yiding Jiang, Aditi Raghunathan, and J Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, 35: 19274–19289, 2022. 3

[6] Simon Baker, Daniel Scharstein, J. P. Lewis, Stefan Roth, Michael J. Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision (IJCV)*, 92(1):1–31, 2011. 4

[7] Michael J. Black and Padmanabhan Anandan. A framework for the robust estimation of optical flow. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 231–236, 1993. 2

[8] Thomas Brox, Christoph Bregler, and Jitendra Malik. Large displacement optical flow. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 41–48, 2009. 2

[9] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *Proc. European Conference on Computer Vision (ECCV)*, pages 611–625, 2012. 1, 2, 4

[10] Adrien Courtois, Jean-Michel Morel, and Pablo Arias. Investigating neural architectures by synthetic dataset design. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4890–4899, 2022. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: a large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 3

[12] Qiaole Dong and Yanwei Fu. MemFlow: optical flow estimation and prediction with memory. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19068–19078, 2024. 2

[13] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Rethinking optical flow from geometric matching consistent perspective. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1337–1347, 2023. 2, 7

[14] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: learning optical flow with convolutional networks. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2758–2766, 2015. 1, 4

[15] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *Proc. International Conference on Learning Representations (ICML)*, pages 1802–1811, 2019. 3

[16] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012. 4

[17] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proc. International Conference on Learning Representations (ICLR)*, 2018. 3

[18] Juan J. Gómez-Rodríguez, José Lamarca, Javier Morlana, Juan D. Tardós, and José M. M. Montiel. SD-DefSLAM: semi-direct monocular slam for deformable and intracorporeal scenes. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 5170–5177, 2021. 1

[19] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 3, 8

[20] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8340–8349, 2021. 3

[21] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1501–1510, 2017. 3

[22] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. FlowFormer: a transformer architecture for optical flow. In *Proc. European Conference on Computer Vision (ECCV)*, pages 668–685, 2022. 2, 7

[23] Junhwa Hur and Stefan Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5754–5763, 2019.

[24] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: evolution of optical flow estimation with deep networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2462–2470, 2017. 2, 4

[25] Azin Jahedi, Maximilian Luz, Marc Rivinius, and Andrés Bruhn. CCMR: high resolution optical flow estimation via coarse-to-fine context-guided motion reasoning. In *Proc.*

IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 6899–6908, 2024.

[26] Azin Jahedi, Maximilian Luz, Marc Rivinius, Lukas Mehl, and Andrés Bruhn. MS-RAFT+: high resolution multi-scale RAFT. *International Journal of Computer Vision (IJCV)*, 132(5):1835–1856, 2024. 2

[27] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9772–9781, 2021. 2, 7

[28] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18963–18974, 2022. 3

[29] M. G. Kendall. Rank correlation methods. *Griffin*, 1948. 8

[30] Yeon-Ho Kim, Aleix M Martínez, and Avi C Kak. Robust motion estimation under varying illumination. *Image and Vision Computing (IVC)*, 23(4):365–375, 2005. 2

[31] Daniel Kondermann, Rahul Nair, Katrin Honauer, Karsten Krispin, Jonas Andrulis, Alexander Brock, Burkhard Gussefeld, Mohsen Rahimimoghaddam, Sabine Hofmann, Claus Brenner, et al. The HCI benchmark suite: stereo and flow ground truth with uncertainties for urban autonomous driving. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 19–28, 2016. 1, 4

[32] Tom Koren, Lior Talker, Michael Dinerstein, and Ran Vitek. Consistent semantic attacks on optical flow. In *Proc. Asian Conference on Computer Vision (ACCV)*, pages 1658–1674, 2022. 2

[33] Marius Leordeanu, Andrei Zanfir, and Cristian Sminchisescu. Locally affine sparse-to-dense matching for motion and occlusion estimation. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1721–1728, 2013. 2

[34] Ruoteng Li, Robby T Tan, and Loong-Fah Cheong. Robust optical flow in rainy scenes. In *Proc. European Conference on Computer Vision (ECCV)*, pages 288–304, 2018. 2

[35] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, 2016. 1, 3, 4

[36] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: a high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4981–4991, 2023. 1, 4

[37] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3061–3070, 2015. 1, 2, 3, 4

[38] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 3

[39] John P. Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. The effect of natural distribution shift on question answering models. In *Proc. International Conference on Learning Representations (ICML)*, pages 6905–6916, 2020. 3

[40] John P. Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *Proc. International Conference on Learning Representations (ICML)*, pages 7721–7735, 2021. 3

[41] Henrique Morimitsu, Xiaobin Zhu, Xiangyang Ji, and Xu-Cheng Yin. Recurrent partial kernel network for efficient optical flow estimation. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pages 4278–4286, 2024. 2

[42] Patrick Müller, Alexander Braun, and Margret Keuper. Classification robustness to common optical aberrations. *Proc. IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3632–3643, 2023. 3

[43] Nils Papenberg, Andrés Bruhn, Thomas Brox, Stephan Didas, and Joachim Weickert. Highly accurate optic flow computation with theoretically justified warping. *International Journal of Computer Vision (IJCV)*, 67:141–158, 2006. 2

[44] Karl Pearson. Note on Regression and Inheritance in the Case of Two Parents. *Proc. Royal Society of London Series I*, 58:240–242, 1895. 5, 6

[45] Travis Portz, Li Zhang, and Hongrui Jiang. Optical flow in the presence of spatially-varying motion blur. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1752–1759, 2012. 2

[46] Anurag Ranjan, Joel Janai, Andreas Geiger, and Michael J Black. Attacking optical flow. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2404–2413, 2019. 2, 8

[47] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proc. International Conference on Learning Representations (ICML)*, pages 5389–5400, 2019. 3

[48] Stephan R. Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2213–2222, 2017. 1, 4

[49] Erik Scheurer, Jenny Schmalfuss, Alexander Lis, and Andrés Bruhn. Detection defenses: An empty promise against adversarial patch attacks on optical flow. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6489–6498, 2024. 2

[50] Jenny Schmalfuss, Lukas Mehl, and Andrés Bruhn. Attacking motion estimation with adversarial snow. *arXiv preprint arXiv:2210.11242*, 2022. 2

[51] Jenny Schmalfuss, Philipp Scholze, and Andrés Bruhn. A perturbation-constrained adversarial attack for evaluating the robustness of optical flow. In *Proc. European Conference on Computer Vision (ECCV)*, pages 183–200, 2022. 2, 8

[52] Jenny Schmalfuss, Lukas Mehl, and Andrés Bruhn. Distracting downpour: adversarial weather attacks for motion estimation. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10106–10116, 2023. 2, 3, 8

[53] Jenny Schmalfuss, Nadine Chang, Vibashan VS, Maying Shen, Andrés Bruhn, and Jose M. Alvarez. PARC: A quantitative framework uncovering the symmetries within vision language models. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25081–25091, 2025. 3

[54] Jenny Schmalfuss, Victor Oei, Lukas Mehl, Madlen Bartsch, Shashank Agnihotri, Margret Keuper, and Andrés Bruhn. RobustSpring: Benchmarking robustness to image corruptions for optical flow, scene flow and stereo. *arXiv preprint arXiv:2505.09368*, 2025. 3

[55] Jenny Schmalfuss, Victor Oei, Lukas Mehl, Madlen Bartsch, Shashank Agnihotri, Margret Keuper, and Andrés Bruhn. RobustSpring: Benchmarking robustness to image corruptions for optical flow, scene flow and stereo. *arXiv preprint arXiv:2505.09368*, 2025. 8

[56] Simon Schrodi, Tonmoy Saikia, and Thomas Brox. Towards understanding adversarial robustness of optical flow networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8916–8924, 2022. 2, 3

[57] Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do image classifiers generalize across time? In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9661–9669, 2021. 3

[58] Zhouxing Shi, Nicholas Carlini, Ananth Balashankar, Ludwig Schmidt, Cho-Jui Hsieh, Alex Beutel, and Yao Qin. Effective robustness against natural distribution shifts for models with different training data. *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, 36, 2024. 3

[59] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Proc. Conference on Neural Information Processing Systems (NIPS)*, 27, 2014. 1

[60] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Models matter, so does training: An empirical study of CNNs for optical flow estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(6): 1408–1423, 2019. 2

[61] Shangkun Sun, Yuanqi Chen, Yu Zhu, Guodong Guo, and Ge Li. SKFlow: learning optical flow with super kernels. *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, 35:11313–11326, 2022. 2, 7

[62] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, 33:18583–18599, 2020. 2, 3, 4

[63] Zachary Teed and Jia Deng. RAFT: recurrent all-pairs field transforms for optical flow. In *Proc. European Conference on Computer Vision (ECCV)*, pages 402–419, 2020. 2, 6, 7

[64] Dinh-Hoan Trinh and Christian Daul. On illumination-invariant variational optical flow for weakly textured scenes. *Computer Vision and Image Understanding (CVIU)*, 179:1–18, 2019. 1

[65] Yi-Hsuan Tsai, Ming-Hsuan Yang, and Michael J. Black. Video segmentation via object flow. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3899–3908, 2016. 1

[66] Alexander von Bernuth, Georg Volk, and Oliver Bringmann. Simulating photo-realistic snow and fog on existing images for enhanced CNN training and evaluation. *Proc. IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 41–46, 2019. 3

[67] Chenguang Wang, Ruoxi Jia, Xin Liu, and Dawn Song. Benchmarking zero-shot robustness of multimodal foundation models: A pilot study. *arXiv preprint arXiv:2403.10499*, 2024. 3

[68] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: a dataset to push the limits of visual slam. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916, 2020. 4

[69] Yihan Wang, Lahav Lipson, and Jia Deng. SEA-RAFT: simple, efficient, accurate RAFT for optical flow. In *Proc. European Conference on Computer Vision (ECCV)*, pages 36–54, 2024. 2, 4

[70] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. DeepFlow: large displacement optical flow with deep matching. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1385–1392, 2013. 2

[71] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. GMFlow: learning optical flow via global matching. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8121–8130, 2022. 2

[72] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023. 2, 7

[73] Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. *arXiv preprint arXiv:2302.12254*, 2023. 3

[74] Lingyu Zhang, Chengzhi Mao, Junfeng Yang, and Carl Vondrick. Adversarially robust video perception by seeing motion. *arXiv preprint arXiv:2212.07815*, 2022. 2

[75] Henning Zimmer, Andrés Bruhn, Joachim Weickert, Levi Valgaerts, Agustín Salgado, Bodo Rosenhahn, and Hans-Peter Seidel. Complementary optic flow. In *Proc. International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*, pages 207–220, 2009. 2