A Perturbation-Constrained Adversarial Attack for Evaluating the Robustness of Optical Flow

Jenny Schmalfuss, Philipp Scholze, and Andrés Bruhn

Institute for Visualization and Interactive Systems, University of Stuttgart, Germany
{jenny.schmalfuss,andres.bruhn}@vis.uni-stuttgart.de
philipp.scholze@simtech.uni-stuttgart.de

Abstract. Recent optical flow methods are almost exclusively judged in terms of accuracy, while their robustness is often neglected. Although adversarial attacks offer a useful tool to perform such an analysis, current attacks on optical flow methods focus on real-world attacking scenarios rather than a *worst case* robustness assessment. Hence, in this work, we propose a novel adversarial attack – the Perturbation-Constrained Flow Attack (PCFA) – that emphasizes destructivity over applicability as a real-world attack. PCFA is a global attack that optimizes adversarial perturbations to shift the predicted flow towards a specified target flow, while keeping the L_2 norm of the perturbation below a chosen bound. Our experiments demonstrate PCFA's applicability in white- and black-box settings, and show it finds stronger adversarial samples than previous attacks. Based on these strong samples, we provide the first joint ranking of optical flow methods considering both prediction quality and adversarial robustness, which reveals state-of-the-art methods to be particularly vulnerable. Code is available at https://github.com/cv-stuttgart/PCFA.

Keywords: Optical Flow · Robustness · Global Adversarial Attack · L_2 Constrained Perturbation

1 Introduction

Optical flow describes the apparent motion between subsequent frames of an image sequence. It has numerous applications ranging from action recognition [42] and video processing [46] to robot navigation [53] and epidemic spread analysis [34]. Over the past decade, the quality of optical flow methods has improved dramatically due to methodological advances: While early optical flow methods were mostly variational [4, 6, 7, 16, 36], today's top methods are based on neural networks [17, 19, 29, 37, 39, 49, 50, 52].

Up to date, theses methodological advances are mainly driven by the quality scores on a few major benchmarks [2,3,9,14,18,23] that measure how well the calculated flow matches a known ground truth. Given that optical flow is also used in the context of medical applications [40,51] and autonomous driving [10,45], it is surprising that robustness plays only a subordinated role in the development of new methods. In fact, robustness is rarely assessed in the literature and only few methods were developed with robustness as explicit goal [4, 21, 22, 35, 43].

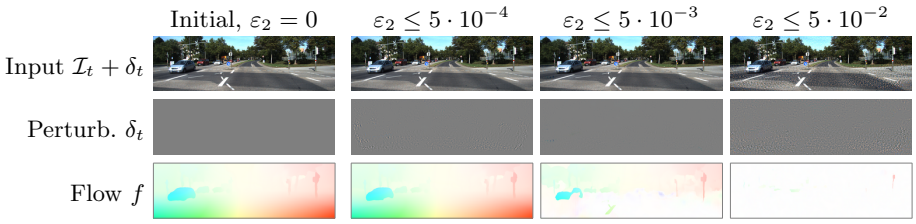

Fig. 1. Robustness evaluation for RAFT [39]. Our Perturbation-Constrained Flow Attack trains flow-erasing *perturbations* δ_t, whose L_2 *norm* is controlled via ε_2.

A possible explanation for this blind spot is the ambiguity of the term *robustness* as well as its challenging quantification for optical flow methods. In this work we therefore focus on an improved measure for quantifying the robustness by means of adversarial attacks. Our choice is motivated by the recently demonstrated vulnerability of optical flow networks to malicious input changes [30]. This vulnerability clearly suggests that adversarial robustness should complement the qualitative performance when evaluating optical flow methods.

Adversarial attacks for optical flow are a very recent field of research with only two attacks available so far. While Ranjan *et al.* [30] proposed a local attack in terms of a patch-based approach, Schrodi *et al.* [32] introduced a global attack inspired by attacks for classification. This raises the question whether these two attacks are sufficiently strong to meaningfully measure adversarial robustness. Answering it is difficult, as clear definitions for *attack strength* and *adversarial robustness* are currently missing in the context of optical flow.

In the context of classification, however, these quantities are already defined. There, adversarial networks aim to find small perturbations to the input that lead to its misclassification [15]. Hence, stronger attacks need smaller input perturbations to cause an incorrect class. While classification networks output a finite amount of discrete classes, optical flow methods predict a field of 2D flow vectors. Using a small perturbation makes it unlikely that one can create an arbitrarily large deviation from the unperturbed flow. Therefore, a sensible definition for a strong attack is "an attack that finds the most destructive adversarial perturbation from all perturbations under a specified bound".

This subtle change in the definition of attack strength for optical flow significantly influences the attack design: For a strong attack, an efficient way to bound the input perturbation is required; see Fig 1. Previous attacks for optical flow either lack effective bounds for their adversarial perturbations [32], or provide a weak adversarial attack [30] to enable real-world applicability. More effective attacks are therefore required for a rigorous quantification of adversarial robustness for optical flow. In this context, we make the following contributions:

1. We formalize a generic *threat model* for optical flow attacks and propose measures for *attack strength* and *adversarial robustness*, to improve the comparability of robustness evaluations and among adversarial attacks.

2. We present the *Perturbation-Constrained Flow Attack (PCFA)*, a strong, global adversarial attack for optical flow that is able to limit the perturbation's L_2 norm to remain within a chosen bound.
3. With PCFA, we generate *joint* and *universal* global perturbations.
4. We experimentally demonstrate that PCFA finds *stronger adversarial samples* and is therefore better suited to quantify adversarial robustness than previous optical flow attacks.
5. We provide the first *ranking* of current optical flow methods that combines their *prediction quality* on benchmarks [9, 23] with their *adversarial robustness* measured by the strongest configuration of PCFA.

2 Related Work

In the following, we mainly focus on related work in the field of optical flow. Thereby, we cover the assessment of robustness, the use of adversarial attacks as well as the design of neural networks. Further related work, also including adversarial attacks for classification, is discussed in our short review in Sec. 3.

Robustness Assessment for Optical Flow. For assessing the robustness of optical flow methods, different concepts have been proposed in the literature. On the one hand, early optical flow methods investigated robustness with regard to outliers [4], noise [6, 8] or illumination changes [43]. On the other hand, the Robust Vision Challenge[1] quantifies robustness as the generalization of a method's qualitative performance across datasets. In contrast, our work relies on a different concept. We consider *adversarial robustness* [15, 38], which is motivated by the Lipschitz continuity of functions. The Lipschitz constant is frequently used as a robustness measure for neural networks, where a small Lipschitz constant implies that the output does only change to the same extend as the input. While finding the exact Lipschitz constant for a neural network is NP-hard [44], bounding it is feasible. For upper bounds on the Lipschitz constant, analytic architecture-dependent considerations are required that are generally difficult – especially for such diverse architectures as in current flow networks. In contrast, finding lower bounds is possible by performing adversarial attacks [11]. Hence, this work uses adversarial attacks to quantify robustness. Thereby, we aim to find input perturbations that cause particularly strong output changes.

Adversarial Attacks for Optical Flow. To the best of our knowledge, there are only two works that propose adversarial attacks tailored to optical flow networks. Ranjan *et al.* [30] developed the first adversarial attack, which causes wrong flow predictions by placing a colorful circular patch in both input frames. To be applicable in the real world, their patches are trained with many constraints (e.g. location and rotation invariances, patches are circular, coherent regions), which comes at the cost of a reduced attack strength. Recently, Schrodi *et al.* [32] introduced a less constrained global attack on optical flow. It is based on the I-FGSM [20] attack for classification that was developed for speed rather

[1] http://www.robustvision.net/

than attack strength and does not effectively limit the perturbation size[2]. Both attacks have their own merits, i.e. speed or real world applicability, but are consequently not fully suitable for a rigorous robustness assessement. In contrast, our novel PCFA has a different purpose: Unlike previous attacks, it does not compromise attack strength, thus enabling an effective robustness quantification.

In the context of vision problems similar to optical flow, Wong *et al.* [47] successfully attacked stereo networks with I-FGSM [20] and its momentum variant MI-FGSM [13]. Moreover, Anand *et al.* [1] proposed an approach to secure optical flow networks for action recognition against adversarial patch attacks by a preceding filtering step that detects, removes and inpaints the attacked location.

Neural Networks for Optical Flow. Regarding neural networks for optical flow, related work is given by those approaches for which we later on evaluate the robustness, i.e. the methods in [17, 19, 29, 37, 39]. These approaches are representatives of the following three classes of networks: classical, pyramidal and recurrent networks. *Classical networks* such as FlowNet2 [17] rely on a stacked encoder-decoder architecture with a dedicated feature extractor and a subsequent correlation layer. More advanced *pyramidal networks* such as SpyNet [29] and PWCNet [37] estimate the optical flow in coarse-to-fine manner using initializations from coarser levels, by warping either the input frames or the extracted features. Finally, state-of-the-art *recurrent networks* such as RAFT [39] and GMA [19] perform iterative updates that rely on a sampling-based hierarchical cost volume. Thereby GMA additionally considers globally aggregated motion features to improve the performance at occlusions.

3 Adversarial Attacks: Foundations and Notations

Adversarial attacks uncovered the brittle performance of neural networks on slightly modified input images, so called *adversarial samples* [15, 38]. Different ways to train adversarial samples exist, which lead to different attack types. *Targeted attacks* perturb the input to induce a specified target output. Compared to *untargeted attacks*, they are considered the better choice for strong attacks, since they can simulate the former ones by running attacks on all possible targets and taking the most successful perturbation [11]. *Global attacks* [11, 13, 15, 20] allow any pixel of the image to be disturbed within a norm bound, while *patch attacks* [5, 30] perturb only pixels within a certain neighborhood. So called *universal perturbations* are particularly transferable and have a degrading effect on multiple images rather than being optimized for a single one [12,24,30,32,33]. As they affect a class of input images, their effect on a single image is often weaker.

[2] FGSM [15] and I-FGSM [20] limit the perturbation size below ε_∞ by performing only so many steps of a fixed step size τ, that exceeding the norm bound is impossible. To this end, the number of steps is fixed to $N = \lfloor \frac{\varepsilon_\infty}{\tau} \rfloor$, which comes down to a one-shot optimization. Additionally, this "early stopping" reduces the attack strength as it prevents optimizing in the vicinity of the bound.

Since adversarial attacks were first used in the context of classification networks [15,38], many concepts go back to this field. Hence, we briefly review those attacks before we discuss attacks for optical flow in more detail.

Classification. Szegedy *et al.* [38] provided the first optimization formulation for a targeted adversarial attack to find a small perturbation $\delta \in \mathbb{R}^m$ to the input $x \in \mathbb{R}^m$, such that a classifier $\mathcal{C}$ outputs the incorrect label t:

$$\min \|\delta\|_2 \quad \text{s.t.} \quad \mathcal{C}(x+\delta) = t\,, \quad \text{and} \quad x+\delta \in [0,1]^m\,. \tag{1}$$

Many adversarial attacks were proposed to solve problem (1), with varying focus and applicability [5, 11, 13, 15, 20]. Among them, the following two methods are most relevant to our work: The Fast Gradient Sign Method (FGSM) [15] method is used for a fast generation of adversarial samples; More recent variations include multiple iterations [20] and momentum [13]. In contrast, the C&W attack [11] emphasizes perturbation destructivity by encouraging the target over all other labels in the optimization (1), while minimizing the adversarial perturbation's L_2 norm. For a broader overview of classification attacks, we refer to the review article of Xu *et al.* [48].

Optical Flow. Given two subsequent frames $\mathcal{I}_t$ and $\mathcal{I}_{t+1} \in \mathbb{R}^I$ of an image sequence, optical flow describes the apparent motion of corresponding pixels in terms of a displacement vector field over the image domain $I = M \times N \times C$, where C is the number of channels per frame. Subsequently, we specify three flow fields: the *ground truth* flow field $f^g = (u^g, v^g) \in \mathbb{R}^{M\times N\times 2}$, the *unattacked* or *initial optical flow* $f \in \mathbb{R}^{M\times N\times 2}$ that comes from a given flow network with two input frames, and the *perturbed* or *adversarial flow* $\check{f} \in \mathbb{R}^{M\times N\times 2}$ from the same network after adding adversarial perturbations $\delta_t, \delta_{t+1} \in \mathbb{R}^I$ to the inputs.

The *local* Patch Attack by Ranjan *et al.* [30] optimizes universal patches

$$\operatorname*{argmin}_{\delta} \mathcal{L}(\check{f}, f^t) \quad \text{s.t.} \quad \delta_t = \delta_{t+1} = \delta\,, \quad \delta \text{ is patch} \quad \text{and} \quad \delta \in [0,1]^I \tag{2}$$

over multiple frames. The cosine similarity serves as loss function $\mathcal{L}$ to minimize the angle between $\check{f}$ and f^t, targeting the negative initial flow $f^t = -f$. To make the circular perturbations location and rotation invariant, the respective transformations are applied before adding the perturbation to the frames.

In contrast, the *global* flow attack by Schrodi *et al.* [32] uses J steps of I-FGSM [20] to generate adversarial perturbations $\|\delta_t, \delta_{t+1}\|_\infty \leq \varepsilon_\infty$ for single frames as

$$\delta_z^{(j+1)} = \delta_z^{(j)} - \frac{\varepsilon_\infty}{J} \cdot \operatorname{sgn}(\nabla_{\mathcal{I}_z + \delta_z^{(j)}} \mathcal{L}(\check{f}, f^t))\,, \quad z = t,\, t+1\,, \quad j = 1, \ldots, J. \tag{3}$$

For universal perturbations, it uses the loss function and training from [30].

While we also develop a global attack like [32], we do not base it on I-FGSM variants as they quickly generate non-optimal perturbations. Instead, we develop a novel attack for optical flow, and explicitly optimize it for attack strength.

4 A Global Perturbation-Constrained Adversarial Attack for Optical Flow Networks

As motivated in the introduction, strong flow attacks require refined notions for *attack strength* and *adversarial robustness*, which are discussed first. Based on these refined notions, we present the Perturbation-Constrained Flow Attack (PCFA) that optimizes for strong adversarial perturbations while keeping their L_2 norm under a specified bound. Moreover, with joint and universal perturbations we discuss different perturbation types for optical flow attacks.

4.1 Attack Strength and Adversarial Robustness for Optical Flow

In the context of classification networks, a strong adversarial sample is one that causes a misclassification while being small, see Problem (1). As optical flow methods do not produce discrete classes but flow fields $f \in \mathbb{R}^{M \times N \times 2}$, significantly larger adversarial perturbations δ_t, δ_{t+1} would be required to induce a specific target flow f^t. Further it is unclear whether a method *can* output a certain target. What can be controlled, however, is the perturbation size. Therefore, a useful *threat model for optical flow* is one that limits the perturbation size to ε and minimizes the distance between attacked flow and target at the same time:

$$\underset{\delta_t, \delta_{t+1}}{\operatorname{argmin}} \mathcal{L}(\check{f}, f^t) \;\text{ s.t. }\; \|\delta_t, \delta_{t+1}\| \leq \varepsilon\,, \;\; \mathcal{I}_z + \delta_z \;\in [0,1]^I\,, \;\; z = t,\, t+1\,. \tag{4}$$

Because the used norms and bounds are generic in this formulation, previous flow attacks fit into this framework: The Patch Attack by Ranjan *et al.* [30] poses a L_0 bound on the patch by limiting the patch size, while the I-FGSM flow attack by Schrodi *et al.* [32] can be seen as a L_∞ bound on the perturbation.

Quantifying Attack Strength. Given that two attacks use the same targets, norms and bounds, they are fully comparable in terms of effectiveness. For stronger attacks, the adversarial flow $\check{f}$ has a better resemblance to the target f^t. As the average endpoint error AEE (see Eq. (7)) is widely used to quantify distances between optical flow fields, we propose to quantify *attack strength* as

$$\text{AEE}(\check{f}, f^t) \;\text{ for }\; \|\delta_t, \delta_{t+1}\| \leq \varepsilon \quad = \quad \text{Attack Strength.}$$

Quantifying Adversarial Robustness. To assess the general robustness of a given method, the specific target is of minor importance. A robust method should not produce a largely different flow prediction for slightly changed input frames, which makes the distance between adversarial flow $\check{f}$ and unattacked flow f a meaningful metric. This distance is small for robust methods and can be quantified with the AEE. Therefore we propose to quantify the *adversarial robustness* for optical flow as

$$\text{AEE}(\check{f}, f) \;\text{ for }\; \|\delta_t, \delta_{t+1}\| \leq \varepsilon \quad = \quad \text{Adversarial Robustness.}$$

This definition of adversarial robustness does intentionally not include a comparison to the ground truth flow f^g as in [29, 32] for two reasons. First, a ground truth comparison measures the flow quality, which should be kept separate from robustness because these quantities likely hinder each other [41]. Secondly, changed frame pairs will have another ground truth, but it is unclear what this ground truth for the attacked frame pairs looks like. While constructing a pseudo ground truth for patch attacks can be possible[3], the required ground truth modifications for global attacks are in general unknown. For those reasons, we suggest to report quality metrics and adversarial robustness separately.

4.2 The Perturbation-Constrained Flow Attack

Starting with the threat model for optical flow (4), we opt for global perturbations to generate strong adversarial samples. To obtain a differentiable formulation, the perturbation is bounded in the L_2 norm. Our *Perturbation-Constrained Flow Attack (PCFA)* then solves the inequality-constrained optimization

$$\underset{\delta_t,\delta_{t+1}}{\operatorname{argmin}}\ \mathcal{L}(\check{f}, f^t) \quad \text{s.t.} \quad \|\delta_t, \delta_{t+1}\|_2 \leq \varepsilon_2\sqrt{2I}, \quad \mathcal{I}_z + \delta_z \in [0,1]^I, \quad z = t, t{+}1. \tag{5}$$

We use the additional factor $\sqrt{2I}$ to make the perturbation bound independent of the image size $I = M \times N \times C$. This way, $\varepsilon_2 = 0.01$ signifies an average distortion of 1% of the frames' color range per pixel. To solve (5), four aspects need consideration: (i) How to implement the inequality constraint $\|\delta_t, \delta_{t+1}\|_2 \leq \varepsilon\sqrt{2I}$, (ii) how to choose the loss function $\mathcal{L}$, (iii) how to choose the target f^t and (iv) how to ensure the box constraint $\mathcal{I}_z + \delta_z \ \in [0,1]^I$, $z = t$, $t{+}1$.

Inequality Constraint. We use a penalty method with exact penalty function [27] to transform the inequality-constrained problem (5) into the following unconstrained optimization problem for $\hat{\delta} = \delta_t, \delta_{t+1}$:

$$\underset{\hat{\delta}}{\operatorname{argmin}}\ \phi(\hat{\delta}, \mu)\,, \quad \phi(\hat{\delta}, \mu) = \mathcal{L}(\check{f}, f^t) + \mu|c(\hat{\delta})|\,. \tag{6}$$

The penalty function c linearly penalizes deviations from the constraint $\|\hat{\delta}\|_2 \leq \hat{\varepsilon}_2 = \varepsilon_2\sqrt{2I}$ and is otherwise zero: $c(\hat{\delta}) = \max(0, \|\hat{\delta}\|_2 - \hat{\varepsilon}_2) = \mathrm{ReLU}(\|\hat{\delta}\|_2 - \hat{\varepsilon}_2)$. If the penalty parameter $\mu \in \mathbb{R}$ approaches infinity, the unconstrained problem (6) will take its minimum within the specified constraint. In practice, it is sufficient to choose μ large. The selected exact penalty function ϕ is nonsmooth at $\|\hat{\delta}\|_2 = \hat{\varepsilon}_2$, which makes its optimization potentially problematic. However, formulating the problem with a smooth penalty function would require to solve a series of optimization problems and is therefore computationally expensive [27]. We solve (6) directly with the L-BFGS [26] optimizer, which worked well in practice. Moreover, in our implementation we use the squared quantities $\|\hat{\delta}\|_2^2 \leq \hat{\varepsilon}_2^2$ for the constraint to avoid a pole in the derivative of $\|\hat{\delta}\|_2$ at $\hat{\delta} = 0$.

[3] Ranjan *et al.* [30] generate a pseudo ground truth for their attack with static patches by prescribing a zero-flow at the patch locations.

Loss Functions. The loss function $\mathcal{L}$ should quantify the proximity of adversarial and target flow. The *average endpoint error (AEE)* is a classical measure to quantify the quality of optical flow as

$$\mathrm{AEE}(\check{f}, f^t) = \frac{1}{MN} \sum_{i \in M \times N} \|\check{f}_i - f_i^t\|_2 \,. \tag{7}$$

However, its derivative is undefined if single components of the adversarial- and target flow coincide, i.e. $\check{f}_i = f_i^t$. In practice, we rarely observed problems for reasonably small perturbations bounds $\hat{\varepsilon}_2$, which prevent a perfect matching. The *mean squared error (MSE)* circumvents this issue due to its squared norm

$$\mathrm{MSE}(\check{f}, f^t) = \frac{1}{MN} \sum_{i \in M \times N} \|\check{f}_i - f_i^t\|_2^2 \,, \tag{8}$$

but is less robust to outliers as deviations are penalized quadratically. Previous optical flow attacks [30, 32] use the *cosine similarity (CS)*

$$\mathrm{CS}(\check{f}, f^t) = \frac{1}{MN} \sum_{i \in M \times N} \frac{\langle \check{f}_i, f_i^t \rangle}{\|\check{f}_i\|_2 \cdot \|f_i^t\|_2} \,. \tag{9}$$

Since this loss only measures angular deviations between flows, it fails to train adversarial perturbation where the adversarial flow or the target are zero.

Target Flows. In principle, any flow field can serve as target flow. Ranjan *et al.* [30] flip the flow direction for a *negative-flow attack* $f^t = -f$. However, this target strongly depends on the initial flow direction. As input agnostic alternative, we propose the *zero-flow attack* with $f^t = 0$. It is especially useful to train universal perturbations that are effective on multiple frames.

Ensuring the Box Constraint. During the optimization, the perturbed frames should remain within the allowed value range, e.g. return a valid color value. All previous flow attacks use *clipping* that crops the perturbed frames to their allowed range after adding δ_t, δ_{t+1}. The *change of variables (COV)* [11] is an alternative approach that optimizes over the auxiliary variables w_t, w_{t+1} instead of δ_t, δ_{t+1} as

$$\delta_z = \frac{1}{2}(\tanh(w_z) + 1) - \mathcal{I}_z \,, \quad z = t,\, t+1 \,. \tag{10}$$

This optimizes $w_t, w_{t+1} \in [-\infty, \infty]^I$, and afterwards maps them into the allowed range $[0, 1]^I$ for the perturbed frames. Our evaluation considers both approaches.

4.3 Joint and Universal Adversarial Perturbations

Out of the box, the optimization problem (6) holds for *disjoint* perturbations, resulting in two perturbations δ_t, δ_{t+1} for an input frame pair $\mathcal{I}_t$, $\mathcal{I}_{t+1}$. Let us now discuss optimizing *joint perturbations* for both input frames and *universal perturbations* for multiple input pairs; their difference is illustrated in Fig. 2. By at-

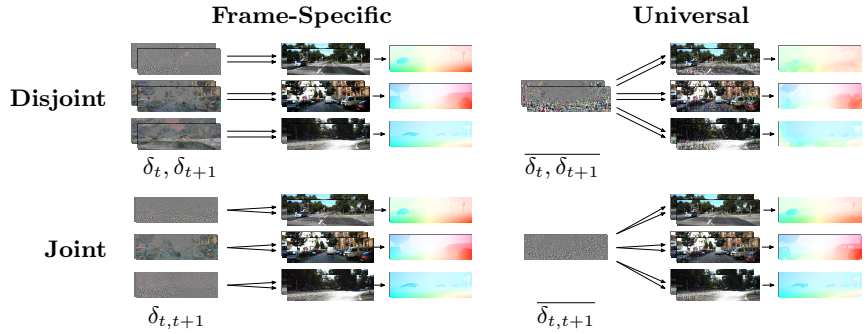

Fig. 2. Illustration of the differences between *disjoint* and *joint* as well as *frame-specific* and *universal* adversarial perturbations for attacking optical flow networks.

tacking both frames or several frame pairs simultaneously, these perturbations have to fulfill more constraints and hence typically offer a weaker performance.

Joint Adversarial Perturbations. In case of *joint* adversarial perturbations, a common perturbation $\delta_{t,t+1}$ is added to both input frames. In its current formulation, the COV box constraint is only possible for disjoint perturbations.

Universal Adversarial Perturbations. Training *universal* instead of *frame-specific* perturbations is straightforward using our optimization (6). Additional projection operations to ensure the norm bound as in other schemes [12,24,30,32, 33] are unnecessary, because PCFA directly optimizes perturbations of limited size. Similar to [33], we refine adversarial perturbations on minibatches. With this scheme, we train disjoint $\overline{\delta_t, \delta_{t+1}}$ and joint $\overline{\delta_{t,t+1}}$ universal perturbations.

4.4 Design Overview and Comparison to Literature

Tab. 1 summarizes our method design-wise and compares it to the other optical flow attacks from the literature. PCFA is the first attack that allows an effective L_2 norm bound for the perturbation. In the evaluation, we provide an extensive analysis of its performance for different perturbation types, losses and targets.

Table 1. Comparison of adversarial optical flow attacks, configurations as stated in the respective publications. *Clip* = Clipping, A = AAE, M = MSE, C = CS.

Attack	Type	$\|\delta\|_*$	Perturbation Types				Losses	Box Constr.
			δ_t, δ_{t+1}	$\delta_{t,t+1}$	$\overline{\delta_t, \delta_{t+1}}$	$\overline{\delta_{t,t+1}}$		
Patch Att. [30]	Patch	L_0	–	–	–	✓	C	Clip
I-FGSM [32]	Global	L_∞	✓	–	✓	–	C	Clip
PCFA (ours)	Global	L_2	✓	✓	✓	✓	A, M, C	Clip, COV

5 Experiments

Our evaluation addresses three distinct aspects: First, we identify the strongest PCFA configuration by evaluating loss functions, box constraints and targets, and compare the resulting approach to I-FGSM [32]. Secondly, we assess the strength of PCFA's joint and universal perturbations in white- and black box attacks, which includes a comparison with the Patch Attack from [29]. Finally, based on PCFA's strongest configuration, we perform a common evaluation of optical flow methods regarding estimation quality *and* adversarial robustness.

At https://github.com/cv-stuttgart/PCFA we provide our PCFA implementation in PyTorch [28]. It is evaluated it with implementations of FlowNet2 [17] from [31], PWCNet [37], SpyNet [29] from [25], RAFT [39] and GMA [19] on the datasets KITTI 2015 [23] and MPI-Sintel final [9]. A full list of parameters and configurations for all experiments is in the supplementary material, Tab. A1.

5.1 Generating Strong Perturbations for Individual Frame Pairs

In the following we consider *disjoint non-universal* perturbations δ_t, δ_{t+1} on the KITTI test dataset. This allows us to (i) identify the strongest PCFA configuration, to (ii) show that PCFA can be used to target specific flows, and to (iii) compare its strength to I-FGSM [32]. We solve PCFA from Eq. (6) with 20 L-BFGS [26] steps per frame pair.

Loss and Box Constraint. Tab. 2 summarizes the attack strength AEE($\check{f}, f^t$) for all combinations of losses and box constraints on the targets $f^t \in \{0, -f\}$ with $\varepsilon_2 = 5 \cdot 10^{-3}$ for RAFT. Compared to clipping, the change of variables (COV) always yields a stronger attack, i.e. a smaller distance to the target when using the same loss. Despite its problematic derivative, the average endpoint error (AEE) reliably outperforms the other losses, while the cosine similarity (CS) that is used in all previous flow attacks [30,32] performs worst. Also, the CS loss fails on the zero-flow target where perturbations keep their initial values (*cf.* Supp. Tab. A2). Since AEE with COV yields the strongest attack independent of the target, we select this configuration for the remaining experiments.

Targets. Next, we investigate how well PCFA can induce a given target flow for different perturbation sizes. Fig. 1 depicts the perturbed input frames, normalized adversarial perturbations and resulting flow fields for a zero-flow attack

Table 2. PCFA attack strength AEE($\check{f}, f^t$) on the KITTI test dataset for different *loss functions*, *targets* and *box constraints* on RAFT. Small values indicate strong attacks.

	$f^t = 0$			$f^t = -f$		
	AEE	MSE	CS	AEE	MSE	CS
Clipping	3.76	10.51	32.37	22.48	38.57	129.44
COV	**3.54**	7.46	32.37	**18.84**	34.82	86.00

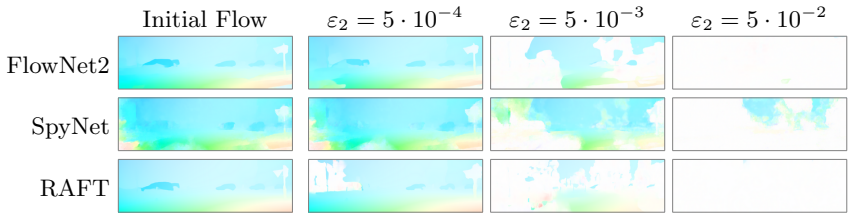

Fig. 3. Visual comparison of PCFA with zero-flow target on different *optical flow methods* for increasing *perturbation sizes* ε_2. White pixels represent zero flow.

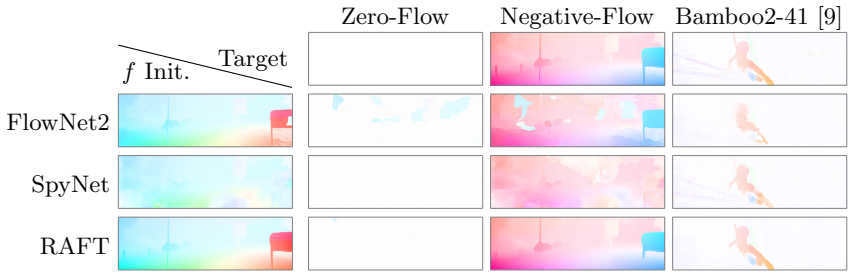

Fig. 4. Visual comparison of PCFA attacked flows with different *targets* for multiple *optical flow methods*. Choosing $\varepsilon_2 = 10^{-1}$ allows to come close to the respective target.

with increasing perturbation size on RAFT. Similarly, Fig. 3 and Supp. Fig. A1 show resulting flow fields for the other networks. Evidently, a perturbation with larger L_2 norm ε_2 reaches a better resemblance between the adversarial flow prediction and the all-white zero-flow target. This is expected, as larger deviations in the input should result in larger output changes. However, it is remarkable that changing color values by 5% on average ($\varepsilon_2 = 5 \cdot 10^{-2}$) suffices to erase the predicted motion. Moreover, not only the zero-flow but also other targets can be induced with PCFA. This is illustrated in Fig. 4 and Supp. Fig. A2. Note that the final proximity to the target mainly depends on its initial distance to the predicted flow field, as close targets are easier to reach.

Comparison of PCFA and I-FGSM. Next, we compare the performance of PCFA and I-FGSM [32] on all networks over a range of perturbations $\varepsilon_2 \in \{5{\cdot}10^{-4}, 10^{-3}, 5{\cdot}10^{-3}, 10^{-2}, 5{\cdot}10^{-2}\}$. We configure I-FGSM as in [32] and iterate until ε_2 is reached, even though it still optimizes for L_∞. Fig. 5 shows the attack strength over the average perturbation norm, Supp. Fig. A3 the corresponding adversarial robustness. While small perturbations δ_t, δ_{t+1} hardly minimize the distance between adversarial and zero-target flow in Fig. 5, large perturbations produce almost perfect target matches with a distance of 0. For each tested optical flow network (color coded) our PCFA (solid) achieves smaller distances than I-FGSM (dashed) – independent of the perturbation size. Hence, PCFA is the global attack of choice to generate strong disjoint image-specific perturbations for optical flow networks.

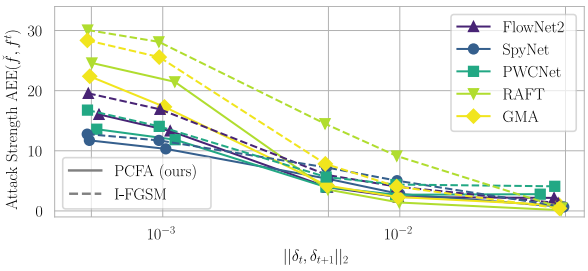

Fig. 5. *Attack strength* with zero-flow target over *perturbation size*, for *PCFA* (solid) and *I-FGSM* [32] (dashed) on different flow networks. Smaller is stronger.

Table 3. Zero-target proximity for different *perturbations*, examples in Supp. Fig. A4.

Perturbation Type		FlowNet2	SypNet	PWCNet	RAFT	GMA
Frame-Specific	δ_t, δ_{t+1}	**3.22**	**4.54**	4.28	**3.76**	**3.59**
	$\delta_{t,t+1}$	4.16	5.84	**3.82**	5.35	4.78
Universal	$\overline{\delta_t, \delta_{t+1}}$	22.03	14.57	19.13	28.88	28.49
	$\overline{\delta_{t,t+1}}$	**20.49**	**14.19**	**18.99**	**28.53**	**27.17**

5.2 Joint and Universal Perturbations

Next we investigate PCFA's potential to generate more general, i.e. joint and universal, perturbations. Moreover, to assess the transferability of network-specific joint universal perturbations, we apply them to all tested networks.

Joint and Universal Perturbations (White Box). In the white box setting, the perturbations are trained and tested on the same model and data. We evaluate the attack strength of perturbation types as target proximity AEE($\check{f}, f^t$) for a zero-flow attack with $\varepsilon_2 = 5 \cdot 10^{-3}$ trained on the KITTI test dataset. For comparability, clipping is used as COV only works for disjoint perturbations. In Tab. 3, frame-specific perturbations clearly show a better target resemblance than universal ones for all networks, as they are optimized for frame-specific destructivity rather than transferability. Further, disjoint perturbations are more effective than joint ones when they are frame-specific. However, for universal perturbations the situation is reversed. This is surprising as disjoint perturbations can adapt to both inputs, which should allow an even better target match. While joint perturbations might add static structures to simulate zero-flow, we reproduced this observation also for a negative-flow target (*cf.* Supp. Tab. A3).

Transferability of Adversarial Perturbations (Black Box). To conclude the PCFA assessment, we train universal joint perturbations in a black box manner. Per network, we optimize $\overline{\delta_{t,t+1}}$ for 25 epochs (batch size 4) on the *training* set, before they are applied to the *test* set for every network. Tab. 4 shows the ad-

Table 4. Transferability of KITTI universal perturbations between *training* and *test* dataset and between different *networks*, measured as adversarial robustness AEE($\check{f}, f$). Large values denote a better transferability, smaller values indicate higher robustness.

Test \ Train	FlowNet2	SpyNet	PWCNet	RAFT	GMA
FlowNet2 [17]	**3.29**	**2.69**	2.22	1.17	1.12
SpyNet [29]	0.60	2.25	0.57	0.46	0.42
PWCNet [37]	1.53	2.19	**2.99**	0.85	0.75
RAFT [39]	2.88	1.87	2.52	3.52	3.19
GMA [19]	3.12	2.14	2.97	**3.95**	**3.81**

Table 5. Adversarial robustness with universal perturbations from *Patch Attack* [30] and *PCFA*, with the setup from Tab. 4. Perturbations from the KITTI training set are applied to the generating network on the KITTI test set, *cf.* Fig. 6 for perturbations.

Attack	FlowNet2	SpyNet	PWCNet	RAFT	GMA
Patch Attack [30]	0.99	1.38	1.37	0.76	0.95
PCFA (ours)	**3.29**	**2.25**	**2.99**	**3.52**	**3.81**

versarial robustness w.r.t. universal perturbations for KITTI (see Supp. Tab. A4 for Sintel). Here, we observe great differences between the transferability. While SpyNet's perturbation reliably disturbs the predictions for all networks, perturbations for RAFT or GMA mutually cause strong deviations but hardly affect other networks. Fig. 6 and Supp. Fig. A5 further suggest that networks with transferable perturbations mainly consider well generalizing, robust features. In contrast, fine-scaled, non-transferable patterns only affect today's top methods RAFT and GMA. Finally, we compare the effectiveness of the global PCFA to the Patch Attack by Ranjan *et al.* [30] with a diameter of 102 px in Tab. 5. As both attack setups perturb a similar amount of information per frame (see supplementary material for details), this supports the initial conjecture that fewer constraints help to increase PCFA's effectiveness.

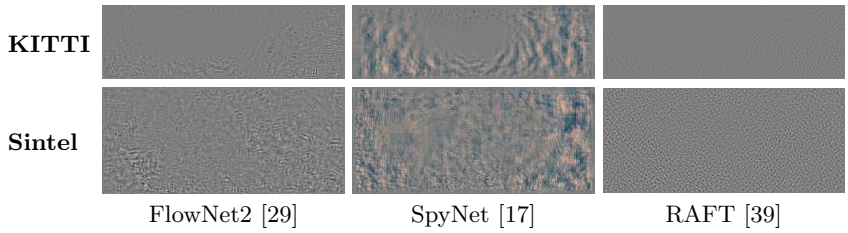

Fig. 6. Normalized *universal perturbations* for different network architectures learned from the respective training datasets. Top row: KITTI. Bottom row: Sintel.

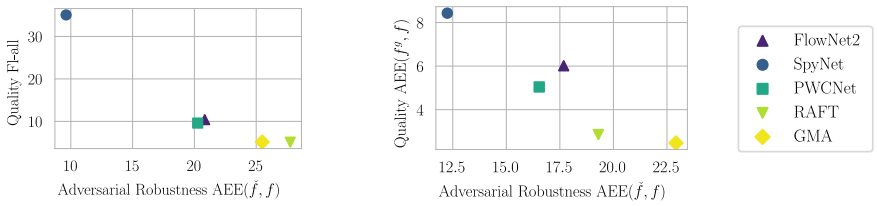

Fig. 7. Joint evaluation of optical flow methods by *prediction quality* and *adversarial robustness* on KITTI (left) and Sintel (right), more attacks in Supp. Fig. A6.

5.3 Evaluating Quality and Robustness for Optical Flow Methods

Finally, we jointly evaluate the optical flow quality and adversarial robustness. For quality, we take the official scores from KITTI and Sintel. For robustness, we apply PCFA's strongest configuration (Sec. 5.1, δ_t, δ_{t+1} with AAE and COV) and report the deviation from the initial flow on the respective test datasets for a zero-flow target, $\varepsilon_2 = 5{\cdot}10^{-3}$. Fig. 7 visualizes quality and adversarial robustness on different axes. On both datasets, we observe methods with good robustness (low adversarial robustness scores) to rank bad in terms of quality (high error) and vice versa. Further, we can identify methods with similar scores: Current networks like RAFT and GMA (recurrent) have good quality but little robustness, FlowNet2 (encoder-decoder) and PWCNet (feature pyramid) balance both, and SpyNet (image pyramid) leads in robustness but has the worst quality. These results indicate that flow networks are subject to a trade-off between accuracy and robustness [41], which also sheds new light on the development of high-accuracy methods that cannot sustain their top rank w.r.t. robustness.

6 Conclusions

This work describes the Perturbation-Constrained Flow Attack (PCFA), a novel global adversarial attack designed for a rigorous adversarial robustness assessment of optical flow networks. In contrast to previous flow attacks, PCFA finds more destructive adversarial perturbations and effectively limits their L_2 norm, which renders it particularly suitable for comparing the robustness of neural networks. Our experimental analysis clearly shows that high quality flow methods are not automatically robust. In fact, these methods seem to be particularly vulnerable to PCFA's perturbations. Therefore, we strongly encourage the research community to treat robustness with equal importance as quality and report both metrics for optical flow methods. With PCFA we not only provide a systematic tool to do so, but with our formal definition of adversarial robustness we also provide a general concept that allows to compare both methods and attacks.

Acknowledgments. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161 (B04). The International Max Planck Research School for Intelligent Systems supports J.S.

References

1. Anand, A.P., Gokul, H., Srinivasan, H., Vijay, P., Vijayaraghavan, V.: Adversarial patch defense for optical flow networks in video action recognition. In: Proc. IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 1289–1296 (2020)
2. 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 (IJCV) **92**(1), 1–31 (2011)
3. Barron, J.L., Fleet, D.J., Beauchemin, S.S.: Performance of optical flow techniques. International Journal of Computer Vision (IJCV) **12**, 43–77 (1994)
4. Black, M.J., Anandan, P.: A framework for the robust estimation of optical flow. In: Proc. IEEE International Conference on Computer Vision (ICCV). pp. 231–236 (1993)
5. Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial patch. In: arXiv preprint 1712.09665 (2018)
6. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: Proc. European Conference on Computer Vision (ECCV). pp. 25–36 (2004)
7. Brox, T., Malik, J.: Large displacement optical flow: Descriptor matching in variational motion estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **33**(3), 500–513 (2011)
8. Bruhn, A., Weickert, J., Schnörr, C.: Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods. International Journal of Computer Vision (IJCV) **61**(3), 211–231 (2005)
9. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: Proc. European Conference on Computer Vision (ECCV). pp. 611–625 (2012)
10. Capito, L., Ozguner, U., Redmill, K.: Optical flow based visual potential field for autonomous driving. In: IEEE Intelligent Vehicles Symposium (IV). pp. 885–891 (2020)
11. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: Proc. IEEE Symposium on Security and Privacy (SP). pp. 39–57 (2017)
12. Deng, Y., Karam, L.J.: Universal adversarial attack via enhanced projected gradient descent. In: Proc. IEEE International Conference on Image Processing (ICIP). pp. 1241–1245 (2020)
13. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
14. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
15. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: arXiv preprint 1412.6572 (2014)
16. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artificial Intelligence (AI) **17**(1-3), 185–203 (1981)
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: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

18. Janai, J., Güney, F., Wulff, J., Black, M., Geiger, A.: Slow Flow: exploiting high-speed cameras for accurate and diverse optical flow reference data. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1406–1416 (2017)
19. Jiang, S., Campbell, D., Lu, Y., Li, H., Hartley, R.: Learning to estimate hidden motions with global motion aggregation. In: Proc. IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9772–9781 (2021)
20. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. In: arXiv preprint 1611.01236 (2017)
21. Li, R., Tan, R.T., Cheong, L.F.: Robust optical flow in rainy scenes. In: Proc. European Conference on Computer Vision (ECCV) (2018)
22. Liu, C., Yuen, J., Torralba, A.: SIFT flow: Dense correspondence across scenes and its applications. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **33**(5), 978–994 (2010)
23. Menze, M., Heipke, C., Geiger, A.: Joint 3D estimation of vehicles and scene flow. In: Proc. ISPRS Workshop on Image Sequence Analysis (ISA) (2015)
24. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
25. Niklaus, S.: A reimplementation of SPyNet using PyTorch (2018), `https://github.com/sniklaus/pytorch-spynet`
26. Nocedal, J.: Updating quasi-Newton matrices with limited storage. Mathematics of Computation **35**(151), 773–782 (1980)
27. Nocedal, J., Wright, S.J.: Numerical optimization. Springer, 2nd edn. (2006)
28. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An imperative style, high-performance deep learning library. In: Proc. Conference on Neural Information Processing Systems (NeurIPS). pp. 8024–8035 (2019)
29. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
30. Ranjan, A., Janai, J., Geiger, A., Black, M.J.: Attacking optical flow. In: Proc. IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
31. Reda, F., Pottorff, R., Barker, J., Catanzaro, B.: flownet2-pytorch: Pytorch implementation of FlowNet 2.0: Evolution of optical flow estimation with deep networks (2017), `https://github.com/NVIDIA/flownet2-pytorch`
32. Schrodi, S., Saikia, T., Brox, T.: Towards understanding adversarial robustness of optical flow networks. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8916–8924 (2022)
33. Shafahi, A., Najibi, M., Xu, Z., Dickerson, J., Davis, L.S., Goldstein, T.: Universal adversarial training. Proc. AAAI Conference on Artificial Intelligence (AAAI) **34**(04), 5636–5643 (04 2020)
34. Stegmaier, T., Oellingrath, E., Himmel, M., Fraas, S.: Differences in epidemic spread patterns of norovirus and influenza seasons of Germany: an application of optical flow analysis in epidemiology. NatureResearch Scientific Reports **10**(1), 1–14 (2020)
35. Stein, F.: Efficient computation of optical flow using the census transform. In: Proc. German Conference on Pattern Recognition (DAGM). pp. 79–86 (2004)

36. Sun, D., Roth, S., Black, M.: Secrets of optical flow estimation and their principles. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2432–2499 (2010)
37. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
38. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: Proc. International Conference on Learning Representations (ICLR) (2014)
39. Teed, Z., Deng, J.: RAFT: Recurrent all-pairs field transforms for optical flow. In: Proc. European Conference on Computer Vision (ECCV). pp. 402–419 (2020)
40. Tehrani, A., Mirzae, M., Rivaz, H.: Semi-supervised training of optical flow convolutional neural networks in ultrasound elastography. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 504–513 (2020)
41. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., Madry, A.: Robustness may be at odds with accuracy. In: Proc. International Conference on Learning Representations (ICLR) (2019)
42. Ullah, A., Muhammad, K., Del Ser, J., Baik, S.W., de Albuquerque, V.H.C.: Activity recognition using temporal optical flow convolutional features and multilayer lstm. IEEE Transactions on Industrial Electronics (TIE) **66**(12), 9692–9702 (2019)
43. van de Weijer, J., Gevers, T.: Robust optical flow from photometric invariants. In: Proc. IEEE International Conference on Image Processing (ICIP). vol. 3, pp. 1835–1838 (2004)
44. Virmaux, A., Scaman, K.: Lipschitz regularity of deep neural networks: analysis and efficient estimation. In: Proc. Conference on Neural Information Processing Systems (NeurIPS) (2018)
45. Wang, H., Cai, P., Fan, R., Sun, Y., Liu, M.: End-to-end interactive prediction and planning with optical flow distillation for autonomous driving. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W). pp. 2229–2238 (2021)
46. Wang, L., Guo, Y., Liu, L., Lin, Z., Deng, X., An, W.: Deep video super-resolution using HR optical flow estimation. IEEE Transactions on Image Processing (TIP) **29**, 4323–4336 (2020)
47. Wong, A., Mundhra, M., Soatto, S.: Stereopagnosia: Fooling stereo networks with adversarial perturbations. Proc. AAAI Conference on Artificial Intelligence (AAAI) **35**(4), 2879–2888 (2021)
48. Xu, H., Ma, Y., Liu, H.C., Deb, D., Liu, H., Tang, J.L., Jain, A.K.: Adversarial attacks and defenses in images, graphs and text: A review. International Journal of Automation and Computing (IJAC) **17**(2), 151–178 (2020)
49. Yang, G., Ramanan, D.: Volumetric correspondence networks for optical flow. In: Proc. Conference on Neural Information Processing Systems (NeurIPS). pp. 794–805 (2019)
50. Yin, Z., Darrell, T., Yu, F.: Hierarchical discrete distribution decomposition for match density estimation. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6044–6053 (2019)
51. Yu, H., Chen, X., Shi, H., Chen, T., Huang, T.S., Sun, S.: Motion pyramid networks for accurate and efficient cardiac motion estimation. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 436–446 (2020)

52. Zhang, F., Woodford, O., Prisacariu, V., Torr, P.: Separable flow: Learning motion cost volumes for optical flow estimation. In: Proc. IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10807–10817 (2021)
53. Zhang, T., Zhang, H., Li, Y., Nakamura, Y., Zhang, L.: Flowfusion: Dynamic dense RGB-D SLAM based on optical flow. In: Proc, IEEE International Conference on Robotics and Automation (ICRA). pp. 7322–7328 (2020)