Detection Defenses: An Empty Promise against Adversarial Patch Attacks on Optical Flow

Erik Scheurer Jenny Schmalfuss Alexander Lis Andrés Bruhn

A. Supplementary Material

In our evaluations, we consider the optical flow methods FlowNetC (FNC) [3], FlowNetCRobust (FNCR) [16], PWCNet (PWC) [18], SpyNet[1] [13], RAFT [19], GMA [6] and FlowFormer (FF) [5].

A.1. Defense hyperparameter evaluation

For the LGS [12] and ILP [1] defenses, we identify those hyperparameters that lead to the most effective defense against the vanilla patch attack [14] on FlowNetC [3]. The hyperparameters under consideration are the block size K, block overlap O and the block filtering threshold t, which are used in ILP and LGS. For LGS, we further consider the smoothing parameter b_{LGS}. For ILP, we consider the scaling s_{ILP}, inpainting radius r_{Talea} and the threshold t_{ILP}. Out of those, we directly set $r_{\mathrm{Talea}} = 5$ and $t_{\mathrm{ILP}} = 0.5$, which are the values from Anand *et al.* [1] that also produced good results for our experiments. For the other parameters, we perform a parameter study that jointly evaluates the parameter pairs K vs. O (for LGS and ILP), t vs. b_{LGS} (for LGS) and t vs. s_{ILP} (for ILP).

Per parameter combination, we evaluate the robustness of the defended FlowNetC against the vanilla attack via $\mathrm{EPE}(f, f_{\mathrm{D}}^{\mathrm{Van}})$ (Fig. A1 left, small values indicate good robustness) and also quantify how much the defense changes the flow prediction for unattacked frames via $\mathrm{EPE}(f, f_{\mathrm{D}})$ (Fig. A1 right, small values indicate that defense does not change the flow prediction on benign samples). Fig. A1 shows the plots for both metrics and all parameter pairs on LGS and ILP. To select the parameters, for each parameter pair we pick values that lead to small values in both metrics (dark colors in plots for $\mathrm{EPE}(f, f_{\mathrm{D}}^{\mathrm{Van}})$ and $\mathrm{EPE}(f, f_{\mathrm{D}})$), because then the defense protects against vanilla attacks but at the same time does not change the flow prediction on unattacked samples. Thus, we select the parameters $K = 16$, $O = 8$, $s_{\mathrm{ILP}} = 15$, $b_{\mathrm{LGS}} = 15$ and $t = 0.15$, which offer the best trade-off between the two metrics. Note that for the K vs. O plots (Fig. A1, Row 1 and 3), the dark area with low values in the upper left corner is unfeasible, because the block overlaps O can not be larger than the blocksize K. Overall, our optimized parameters differ slightly from the literature values: For LGS, the original publication [12] used $K = 15$, $O = 5$ and $b_{\mathrm{LGS}} = 2.3$ (for classification), while the original values for ILP from [1] are $t = 0.25$, and $s_{\mathrm{ILP}} = 10$ (for optical-flow-based action recognition).

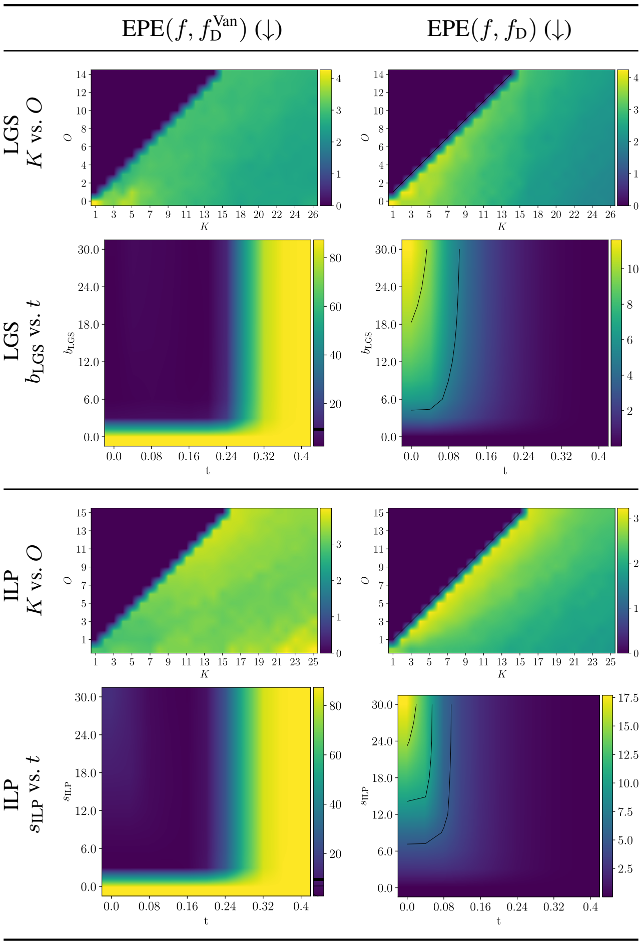


Figure A1. LGS and ILP hyperparameter study based on FlowNetC. Good parameters should balance the robustness against the vanilla attack $\mathrm{EPE}(f, f_{\mathrm{D}}^{\mathrm{Van}})$ (dark color = good robustness) and small flow perturbations through the defense on unattacked frames $\mathrm{EPE}(f, f_{\mathrm{D}})$ (dark color = small perturbation). We select $K = 16$, $O = 8$, $t = 0.15$, $b_{\mathrm{LGS}} = 15$ and $s_{\mathrm{ILP}} = 15$ as best parameters

A.2. Defense-aware attack setup and parameters

Next, we evaluate the best combination of optimizers, learning rates (LR), and box constraints to optimize defense-aware patches. As optimizers, we consider I-FGSM [8] and SGD, as learning rates 1, 0.1, 0.01 for I-FGSM and 10, 100 for SGD, and as box constraints either

[1] Implementation from github.com/sniklaus/pytorch-spynet.

Table A1. Robustness EPE(f, f^{Van}) [15] for undefended networks under vanilla patch attacks with different optimization parameter combinations. Non-evaluated settings are marked by "n.e.".

Optim.	LR	Box	FNC	FNCR	PWC	SpyNet	RAFT	GMA	FF
SGD	10.00	CoV	61.86	0.73	1.34	1.10	0.27	0.29	0.42
SGD	10.00	Clip	52.41	0.97	1.17	1.01	0.28	0.31	0.45
SGD	100.00	CoV	**76.28**	0.62	1.28	1.26	0.29	0.34	n.e.
SGD	100.00	Clip	63.74	0.44	1.17	1.26	0.27	0.28	0.49
IFGSM	0.01	CoV	58.56	1.28	**1.84**	1.30	0.29	0.61	n.e.
IFGSM	0.01	Clip	32.19	**1.58**	1.80	1.19	0.29	**0.55**	**0.54**
IFGSM	0.10	CoV	57.55	1.47	1.84	1.33	**0.34**	0.46	n.e.
IFGSM	0.10	Clip	55.92	0.50	1.03	1.15	0.24	0.30	0.49
IFGSM	1.00	CoV	60.62	1.23	1.60	**1.33**	0.34	0.41	n.e.
IFGSM	1.00	Clip	8.22	0.45	0.88	1.11	0.26	0.32	0.44

Table A2. Robustness EPE$(f_{\text{LGS}}, f_{\text{LGS}}^{\text{LGS}})$ for LGS-defended networks under LGS-aware patch attacks with different optimization parameter combinations. Non-evaluated settings are marked by "n.e.", while diverging optimization runs are marked as "div".

Optim.	LR	Box	FNC	FNCR	PWC	SpyNet	RAFT	GMA	FF
SGD	10.00	CoV	3.98	3.07	3.28	3.62	1.45	1.57	n.e.
SGD	10.00	Clip	3.02	2.98	3.03	3.50	1.31	1.49	1.59
SGD	100.00	CoV	3.44	3.17	3.29	3.74	**1.47**	div.	n.e.
SGD	100.00	Clip	3.27	3.17	3.15	3.59	1.33	div.	1.56
IFGSM	0.01	CoV	**22.64**	2.51	3.74	3.69	1.05	1.24	n.e.
IFGSM	0.01	Clip	19.03	2.71	**3.90**	3.67	1.17	1.34	1.31
IFGSM	0.10	CoV	20.70	2.89	3.68	**3.98**	1.33	1.49	n.e.
IFGSM	0.10	Clip	8.42	2.62	3.04	3.60	1.22	1.33	1.28
IFGSM	1.00	CoV	4.61	3.10	3.28	3.62	1.42	1.53	n.e.
IFGSM	1.00	Clip	3.48	**3.28**	3.37	3.86	1.45	**1.62**	**1.71**

clipping or a change of variables (CoV). Due to the algorithmic differences between I-FGSM and SGD, the considered learning rates have distinct magnitudes to achieve comparable results. For each defense (none, LGS and ILP) we evaluate the pipeline robustness of the defended method under four separately trained defense-aware patches (using four fixed random seeds). We report the averaged robustness values for our defense-aware patches in Tab. A1 (no defense), Tab. A2 (LGS defense) and Tab. A3 (ILP defense).

Each defense-aware patch is trained for 2500 steps. The patches are randomly placed on the image, randomly rotated in a range of $[-10, 10]$ degrees, and randomly scaled in a range of $[0.95, 1.05]$. Batch size is chosen as 1 as the effect of batch size on the patch training is negligible. Due to its size and the resulting computational cost to evaluate FlowFormer [5], we only train its patches for 1000 iterations and omit the change of variables to reduce the number of test runs, as it performed similarly to RAFT and GMA. We found patches to be sufficiently converged after the reduced number of iterations. Please note that across all methods, the choice of box constraint did not significantly influence the effectiveness of the adversarial patches. The patch optimization for GMA diverged with SGD and learning rate 100 for LGS- and ILP-aware patches.

Table A3. Robustness EPE$(f_{\text{ILP}}, f_{\text{ILP}}^{\text{ILP}})$ for ILP-defended networks under ILP-aware patch attacks with different optimization parameter combinations. Non-evaluated settings are marked by "n.e.", while diverging optimization runs are marked as "div".

Optim.	LR	Box	FNC	FNCR	PWC	SpyNet	RAFT	GMA	FF
SGD	10.00	CoV	11.55	1.53	2.21	1.73	1.40	1.46	n.e.
SGD	10.00	Clip	4.09	2.99	2.99	2.17	1.45	1.48	1.75
SGD	100.00	CoV	3.17	2.90	3.08	2.52	1.43	div.	n.e.
SGD	100.00	Clip	3.56	3.27	3.37	2.83	1.42	1.55	1.69
IFGSM	0.01	CoV	**57.46**	2.95	3.84	2.77	1.12	1.25	n.e.
IFGSM	0.01	Clip	42.87	2.91	**3.87**	2.72	1.08	1.24	1.15
IFGSM	0.10	CoV	54.74	**3.30**	3.87	**3.15**	1.36	1.46	n.e.
IFGSM	0.10	Clip	18.70	2.93	3.11	2.77	1.23	1.32	1.39
IFGSM	1.00	CoV	3.84	3.22	3.34	2.98	1.37	1.43	n.e.
IFGSM	1.00	Clip	3.58	3.28	3.42	2.78	**1.48**	**1.54**	**1.82**

Based on this extensive parameter evaluation, we select the best optimization parameters for all combinations of optical flow network and defense-aware attack in Tab. A4, which are boldfaced in the detailed evaluations in Tab. A1, Tab. A2 and Tab. A3. These parameters were used to produce the defense-aware patches for the experimental evaluation in the Main paper.

Additionally, we show the best (out of four) defense-aware patches for no defense, LGS-defense and ILP-defense in Fig. A2, Fig. A3 and Fig. A4, respectively. The best patch is selected based on the worst robustness score of the defended method after training.

A.3. Additional flow visualizations for vanilla attack

Here, we complement the limited selection of methods whose optical flow was visualized for unattacked and (vanilla) attacked frames in Main Fig. 4. Unattacked and vanilla-attacked flow visualizations on *all* tested optical flow methods for the previous KITTI scene are in Fig. A5 and for an additional KITTI sample in Fig. A6. For a lean representation, only a single frame of the attacked image pair is shown on the right.

In both figures, RAFT [19], GMA [6] and FlowFormer [5] are able to recognize the patch as a static object in the scene and therefore predict its output flow as zero. The less accurate methods SpyNet [13], PWCNet [18] and FlowNetCRobust [16] also recognize the zero flow, but their flow predictions are overall less precise and the patch bleeds into the surrounding area. The outlier is FlowNetC [3], where the entire flow prediction is deteriorated by the patch.

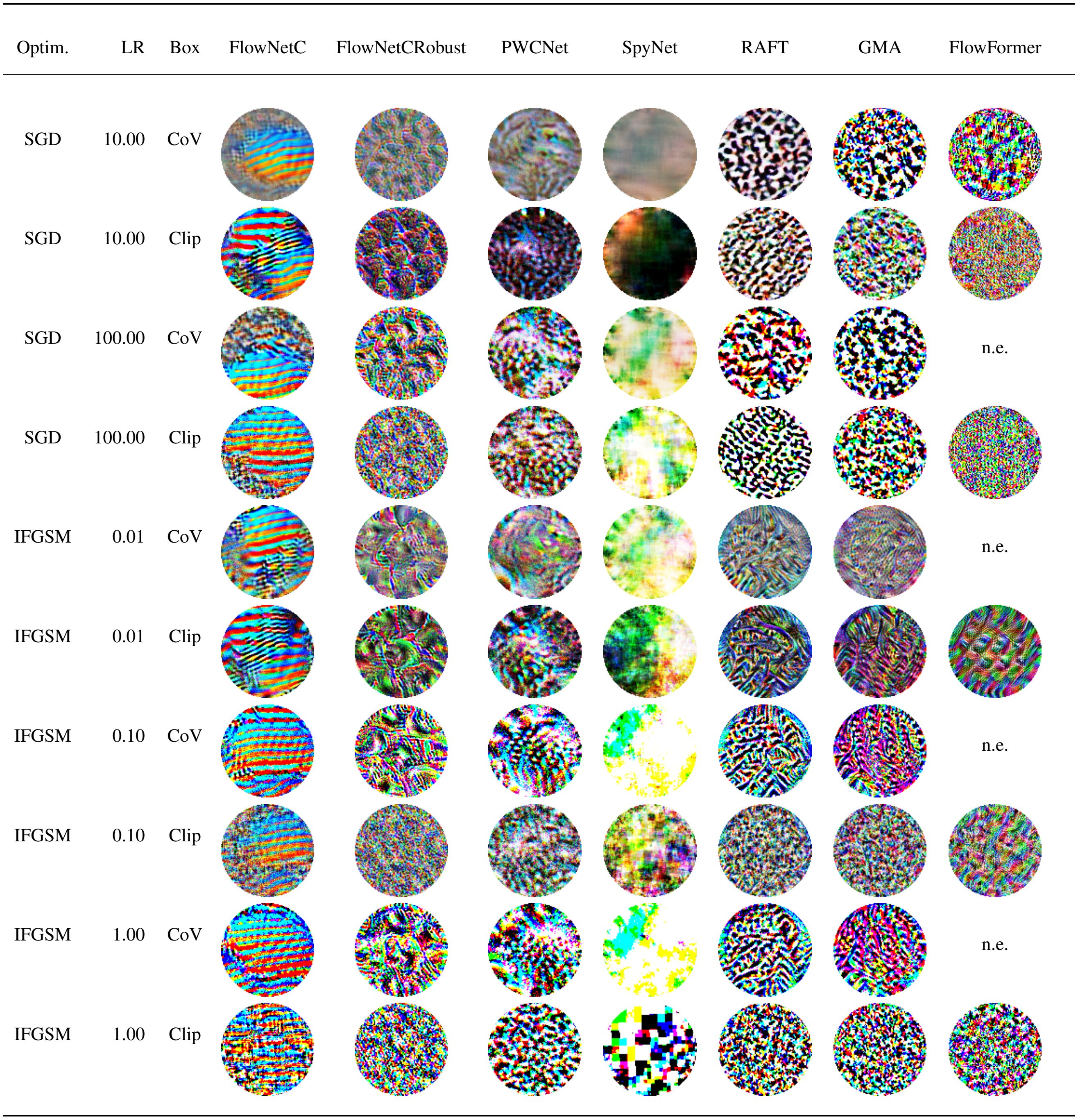


Figure A2. Best-performing vanilla patches for different networks and optimization parameter combinations. Non-evaluated settings are marked by “n.e.”. See Tab. A1 for the corresponding robustness values, averaged over four patches.

In both visualizations, almost all optical flow methods are hardly affected by the patch, as they correctly recognize it as an object and accurately predict its zero motion.

A.4. Manual patch attack: Defended quality

In the manual patch analysis in Sec. 6.4, the Main paper visually argued that our high-frequent, manual patch attacks qualitatively improve the optical flow predictions of LGS- and ILP-defended methods, as a result of the significant quality degradation of defenses on unattacked im-

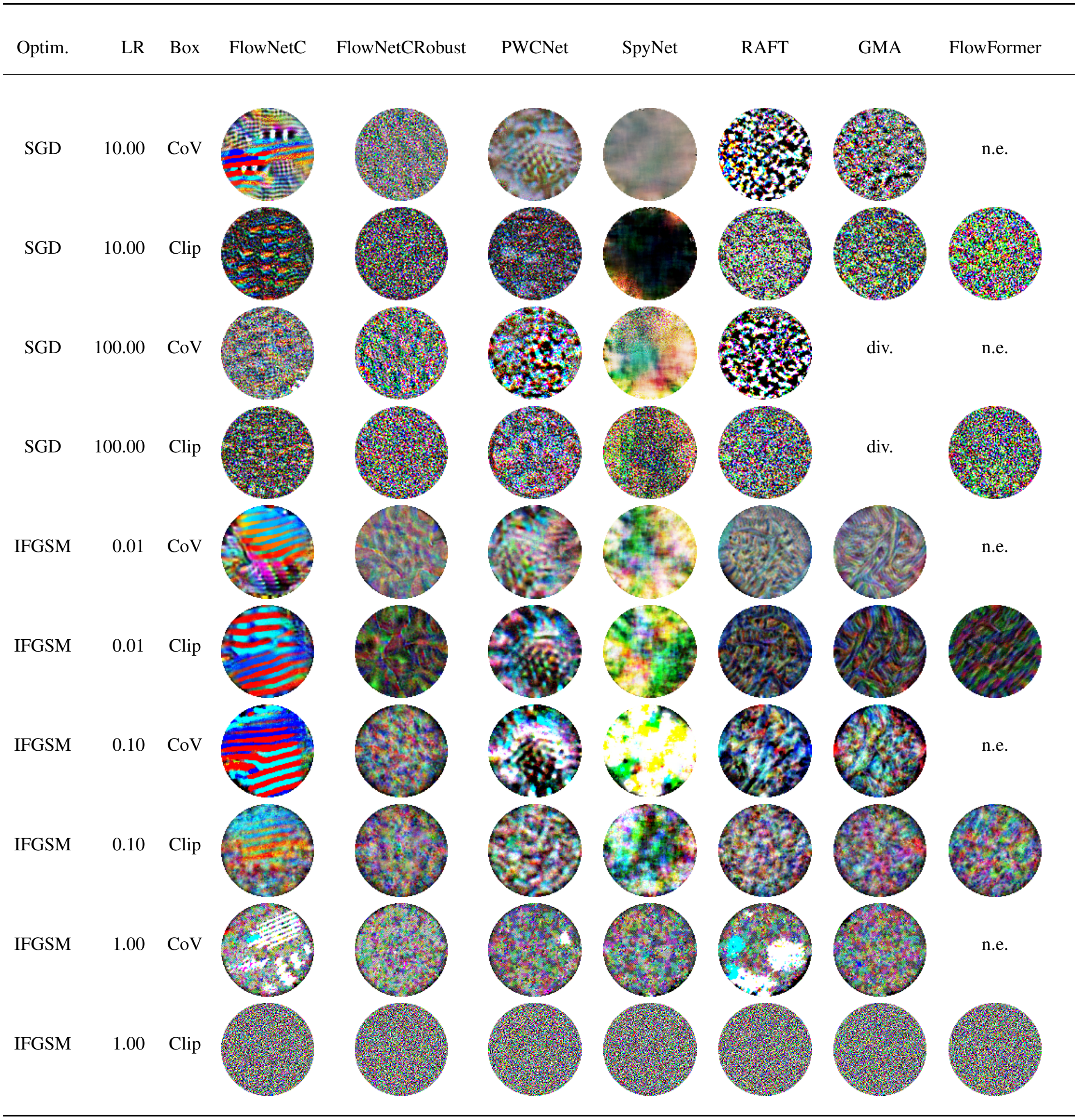


Figure A3. Best-performing LGS-aware patches for different networks and optimization parameter combinations. Non-evaluated settings are marked by "n.e.", while diverging optimization runs are marked as "div". See Tab. A2 for the corresponding robustness values, averaged over four patches.

ages. Here, we provide the corresponding quality scores over the whole set of KITTI frames. To this end, we quantify the quality $Q_{\mathrm{D}}^{\mathrm{A}} = \mathrm{EPE}(f^*, f_{\mathrm{D}}^{\mathrm{A}})$, *i.e.* the distance between ground truth flow f^* and optical flow predictions $f_{\mathrm{D}}^{\mathrm{A}}$ of methods that are defended with D and attacked with A.

Tab. A5 provides the quality scores for unattacked but defended networks (block 1, corresponds to values from Main Tab. 1), for full pipelines where the defended method is attacked with the corresponding defense-awareness (block 2) and for our manual attack on defended

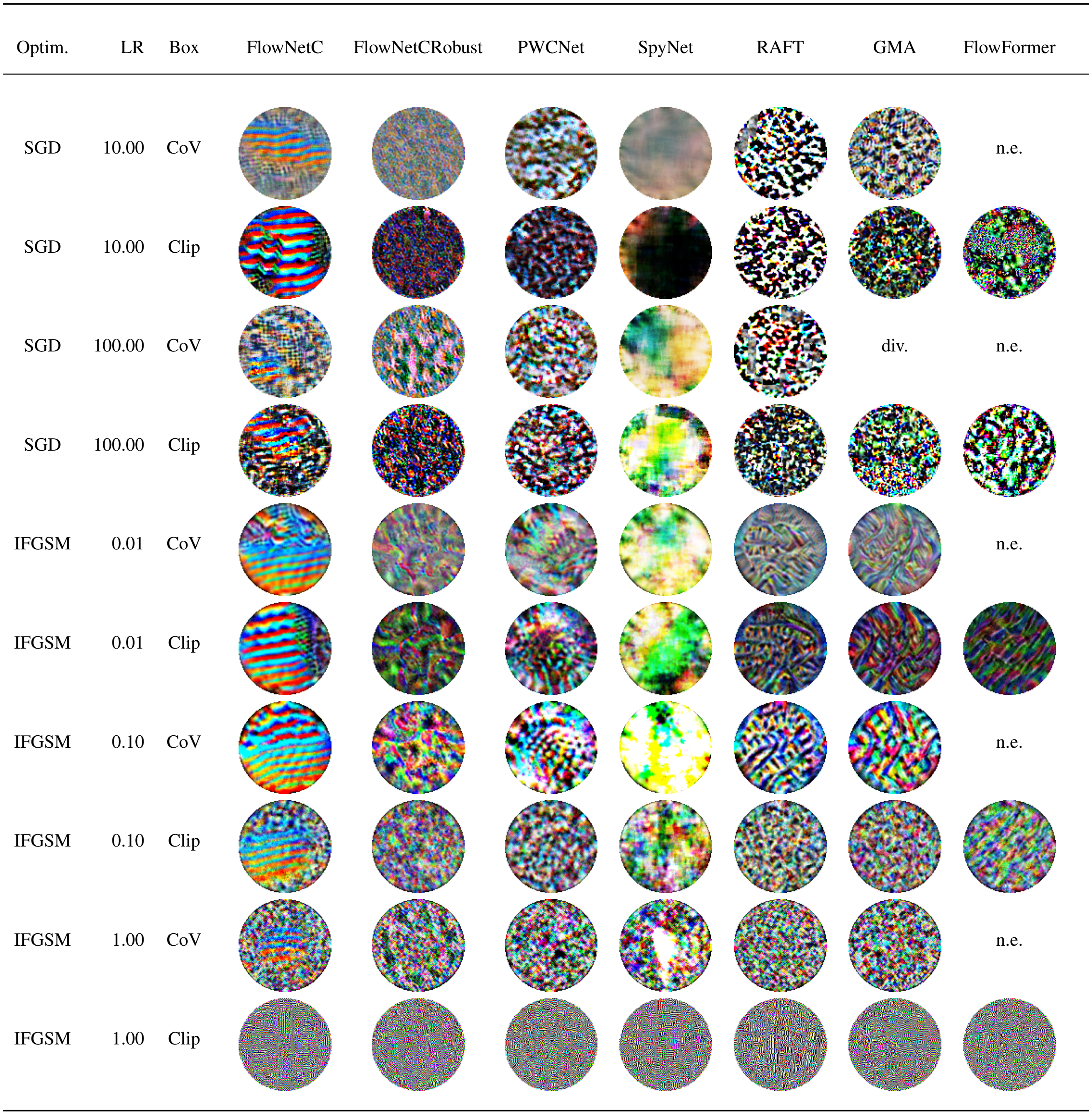


Figure A4. Best-performing vanilla patches for different networks and optimization parameter combinations. Non-evaluated settings are marked by “n.e.”, while diverging optimization runs are marked as “div”. See Tab. A3 for the corresponding robustness values, averaged over four patches.

networks (block 3). Compared to the original baseline $Q = \mathrm{EPE}(f^*, f)$ (block 1, marked in gray), all defenses decrease the quality. Fig. A7 and Fig. A8 visualize the flow output for unattacked models on KITTI samples when no defense, LGS or ILP is applied.

Then, we begin by comparing the quality for defended methods in the first two blocks, *i.e.* we exclude the manual patch attack in block 3, and mark the best quality **bold** in the table. Note that we exclude the gray rows, as they contain the quality for *undefended* methods. For optical flow methods that have a good undefended quality Q, *i.e.* FlowNetCRobust, RAFT, GMA and FlowFormer, we find

Table A4. Optimal parameter setups for defense-aware patch attacks on all optical flow methods with defenses. LR is the learning rate, and Box indicates whether a change of variables or clipping is used during optimization. The settings are a summary of the best results from Tab. A1, Tab. A2 and Tab. A3.

Attacked model	Defense	Attack	Optimizer	LR	Constraint
FlowNetC	None	Vanilla	SGD	100.00	CoV
	LGS	+LGS	IFGSM	0.01	CoV
	ILP	+ILP	IFGSM	0.01	CoV
FNCR	None	Vanilla	IFGSM	0.01	Clip
	LGS	+LGS	IFGSM	1.00	Clip
	ILP	+ILP	IFGSM	0.10	CoV
SpyNet	None	Vanilla	IFGSM	0.10	CoV
	LGS	+LGS	IFGSM	0.10	CoV
	ILP	+ILP	IFGSM	0.10	CoV
PWCNet	None	Vanilla	IFGSM	0.01	CoV
	LGS	+LGS	IFGSM	0.01	Clip
	ILP	+ILP	IFGSM	0.01	Clip
RAFT	None	Vanilla	IFGSM	1.00	CoV
	LGS	+LGS	SGD	100.00	CoV
	ILP	+ILP	IFGSM	1.00	Clip
GMA	None	Vanilla	IFGSM	0.01	CoV
	LGS	+LGS	IFGSM	1.00	Clip
	ILP	+ILP	IFGSM	1.00	Clip
FlowFormer	None	Vanilla	IFGSM	0.01	Clip
	LGS	+LGS	IFGSM	1.00	Clip
	ILP	+ILP	IFGSM	1.00	Clip

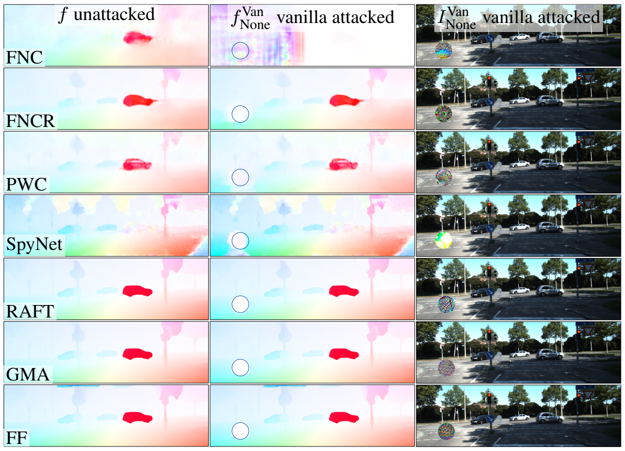

f unattacked
$f^{\mathrm{Van}}_{\mathrm{None}}$ vanilla attacked
$I^{\mathrm{Van}}_{\mathrm{None}}$ vanilla attacked
FNC
FNCR
PWC
SpyNet
RAFT
GMA
FF

Figure A5. Unattacked optical flow estimation (left) and corresponding vanilla-attacked optical flow (middle) for all tested methods on a KITTI sample (right). Complements Main Fig. 4, see Fig. A6 for more samples.

that a defense-aware attack on a defended model actually yields a better quality than the defended but unattacked model: $Q^{\mathrm{D}}_{\mathrm{D}} > Q_{\mathrm{D}}$. For these methods, a noisy patch was revealed to be the most effective. Hence, it is easier for an adaptive attack to exploit the changes introduced by the defense than to influence the flow estimation.

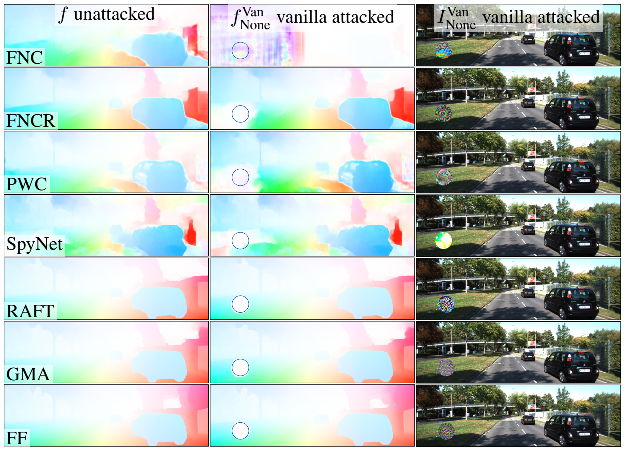

f unattacked
$f^{\mathrm{Van}}_{\mathrm{None}}$ vanilla attacked
$I^{\mathrm{Van}}_{\mathrm{None}}$ vanilla attacked
FNC
FNCR
PWC
SpyNet
RAFT
GMA
FF

Figure A6. Unattacked optical flow estimation (left) and corresponding vanilla-attacked optical flow (middle) for all tested methods on a KITTI sample (right). See Fig. A5 for more samples.

Now we also include the manual patch attack in the defense evaluation, again underlining the highest-quality flow per method over all three blocks in Tab. A5. Again we exclude the gray rows that contain the quality for *undefended* methods in order to compare the influence of the defenses. Now, for almost all methods the best defended quality is achieved for manual patch attacks. When we compare the underlined numbers to the baseline quality Q, we find that our manual patch attack almost restores the undefended and unattacked quality for our defended methods outside the patch area. While this underlines the finding from the Main paper that the low quality of defended but unattacked methods is the main reason for the low quality (and robustness) of defended methods, it also yields another point: If the defenses did not deteriorate the unattacked quality, they could be effective in terms of quality and robustness because they restore high-quality optical flow fields in the presence of adversarial-like patches.

A.5. Defense evaluation on additional datasets

We evaluate the defenses and their effectiveness on more datasets besides KITTI [11], and consider Sintel [2], Driving [9], HD1K [7] and Spring [10]. Because evaluating the defended quality requires ground truth optical flow data, we use validation splits of the respective test sets for all datasets. Dataset-specific patches are then trained on the remaining training data. For Sintel, we use the validation set from [20] which splits Sintel-test such that the flow magnitudes of the validation set match the flow-magnitude distribution of the full training set [17]. For HD1K and Spring, we are unaware of flow-magnitude matching validation splits in the literature, and create validation splits with matching flow-magnitude distributions as detailed in Tab. A6. For Driving, we use the scenes with focal length 15mm, forwards, fast speed and left camera as validation

Table A5. Quality $Q_{\text{D}}^{\text{A}} = \text{EPE}(f^*, f_{\text{D}}^{\text{A}})$, *i.e.* the distance between ground truth flow and optical flow predictions that are defended with D and attacked with A. The upper block shows the quality scores from Main Tab. 1 (for comparison), and the lower blocks contain the quality scores for full pipelines and manual patch attacks on networks with varying defenses. Per method, we mark the best defended quality for unattacked networks and full pipelines **bold** (includes the first two blocks, up to double line), and underline the best quality if the manual patch is also included – the undefended baselines that are marked in gray are excluded from both rankings.

u Attack type	Defense		FNC	FNCR	PWC	SpyNet	RAFT	GMA	FF
No Attack	None	Q	15.42	11.10	13.26	24.03	0.63	0.61	0.62
	LGS	Q_{LGS}	16.70	13.13	14.61	25.15	1.42	1.55	1.42
	ILP	Q_{ILP}	**16.46**	12.77	**14.52**	**24.74**	1.36	1.39	1.30
Vanilla	None	Q^{Van}	84.48	12.64	15.27	25.11	0.80	0.91	0.78
+LGS (LGS-aware)	LGS	$Q^{\text{LGS}}_{\text{LGS}}$	34.41	**11.68**	15.43	25.28	0.94	0.90	0.83
+ILP (ILP-aware)	ILP	$Q^{\text{ILP}}_{\text{ILP}}$	65.02	12.31	15.65	25.29	**<u>0.68</u>**	**0.70**	**<u>0.68</u>**
Manual	None	Q^{Man}	16.21	11.38	13.87	24.79	0.71	0.70	0.70
	LGS	$Q^{\text{Man}}_{\text{LGS}}$	16.66	11.69	14.24	24.87	0.92	0.91	0.88
	ILP	$Q^{\text{Man}}_{\text{ILP}}$	<u>16.16</u>	<u>11.52</u>	<u>13.97</u>	<u>24.69</u>	0.70	<u>0.68</u>	0.72

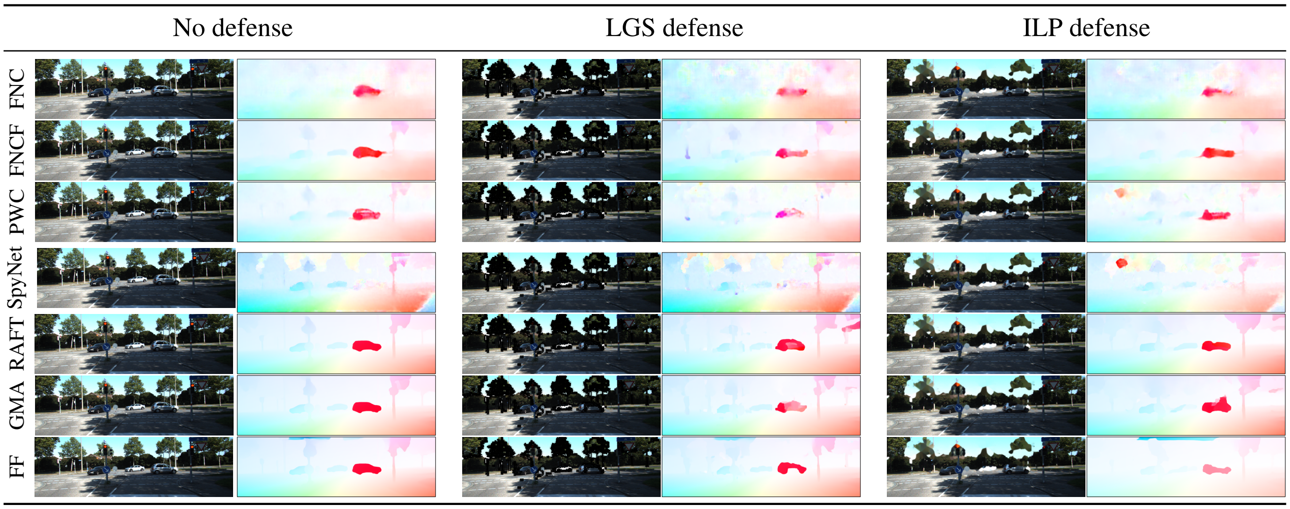


Figure A7. Optical flow prediction on an unattacked frame of the KITTI dataset for optical flow methods with different defenses. Defenses from left to right: None, LGS and ILP. See Fig. A8 for more samples.

split. Note that during our evaluations, we half the image resolution for HD1K and Spring, to keep the image sizes and hence results for patches with size 100 comparable across all datasets.

For the datasets HD1K, Spring, Sintel (clean and final) and Driving (clean and final), we show the numerical results of the defended quality analysis in Tab. A7, Tab. A8, Tab. A9 and Tab. A10, and the respective robustness analyses in Tab. A11, Tab. A12, Tab. A13 and Tab. A14. For a better overview, Fig. A9 shows the quality vs. robustness plots for all tested optical flow methods on all tested datasets, which can be compared to the results on KITTI in Main Fig. 5.

Focusing on the quality-robustness plots in Fig. A9, we observe that defenses worsen quality and robustness for all optical flow methods (except those of FlowNetC) on HD1K and Spring, *cf*. Fig. A9a and Fig. A9d. On Sintel and Driving, the results are more differentiated: For high-quality methods like RAFT, GMA and FlowFormer (red markers), defending them with ILP improves the robustness for the final versions of the datasets in Fig. A9e and Fig. A9f – on the clean dataset versions in Fig. A9b and Fig. A9c, however, both defenses deteriorate either quality, or robustness, or both. Defending the lower-quality methods SpyNet and PWCNet (blue markers) also deteriorates at least quality or robustness on both datasets, with the exception of PWCNet, where defending leads to minor robustness im-

[2] See Tab. A6 for details on the used validation splits

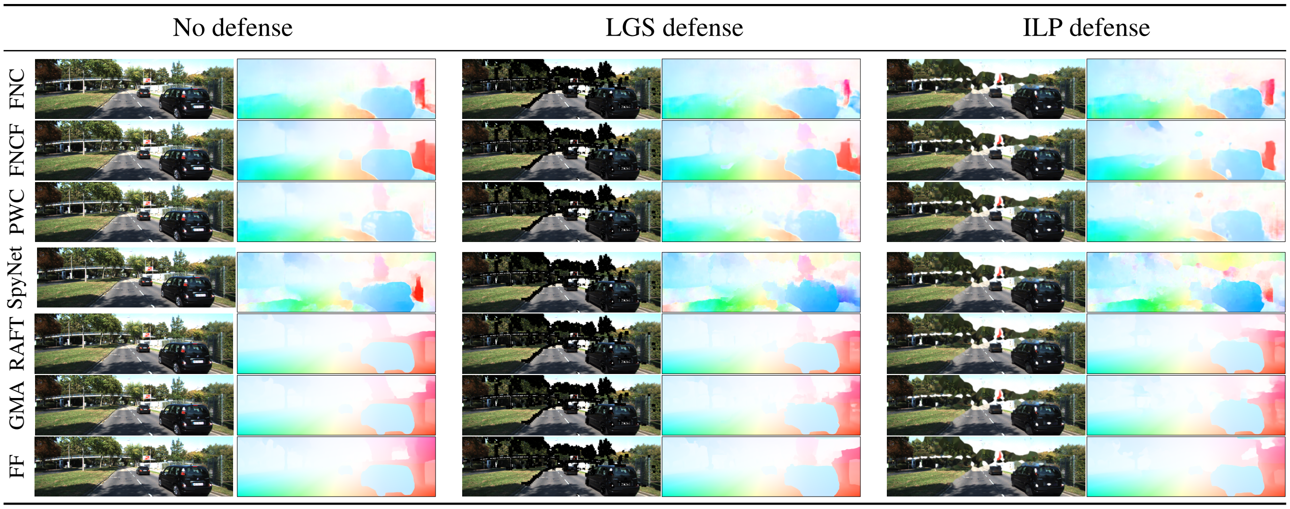
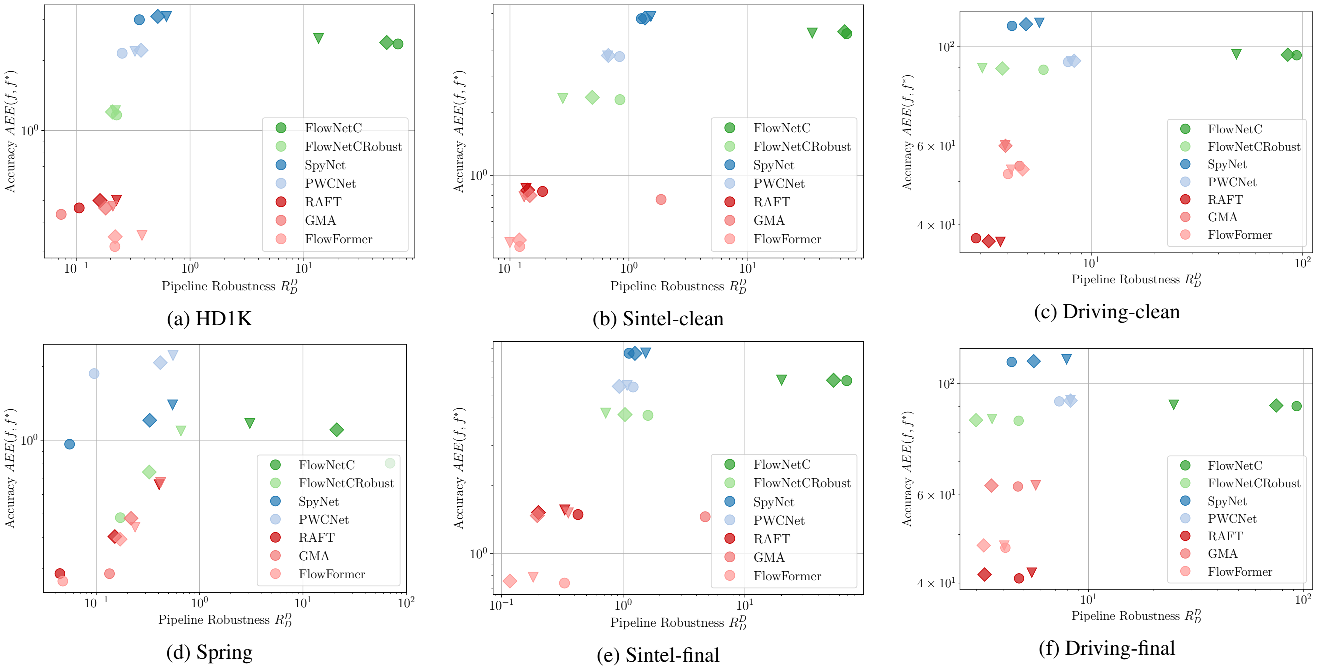


Figure A8. Optical flow prediction on an unattacked frame of the KITTI dataset for optical flow methods with different defenses. Defenses from left to right: None, LGS and ILP. See Fig. A7 for more samples.



(a) HD1K

(b) Sintel-clean

(c) Driving-clean

(d) Spring

(e) Sintel-final

(f) Driving-final

Figure A9. Quality vs. robustness of flow networks on different datasets in a double logarithmic plot. An ideal method would be in the origin. Undefended networks are circles ◯, networks defended with LGS are triangles ▽ and networks defended with ILP are diamonds ◇.

provements on Sintel. For FlowNetC and FlowNetCRobust (green markers), defenses do indeed improve the robustness on Sintel and Driving, but here it is LGS that leads to the best robustness scores. Overall, this clearly supports that defenses should not be used in a "plug'n'play" manner without extensive application-specific testing, as they either do not improve the optical flow methods at all, or – when they do improve the robustness – their effect is small and does not apply to more than a few selected optical flow methods. Hence, current detect-and-remove defenses cannot be recommended for general use.

To better understand the effectiveness differences of defenses on the tested datasets, we analyze the results in relation to the datasets in more detail. When we consider the datasets KITTI, HD1K and Spring and their quality-robustness plots in Main Fig. 5, Fig. A9a and Fig. A9d, we

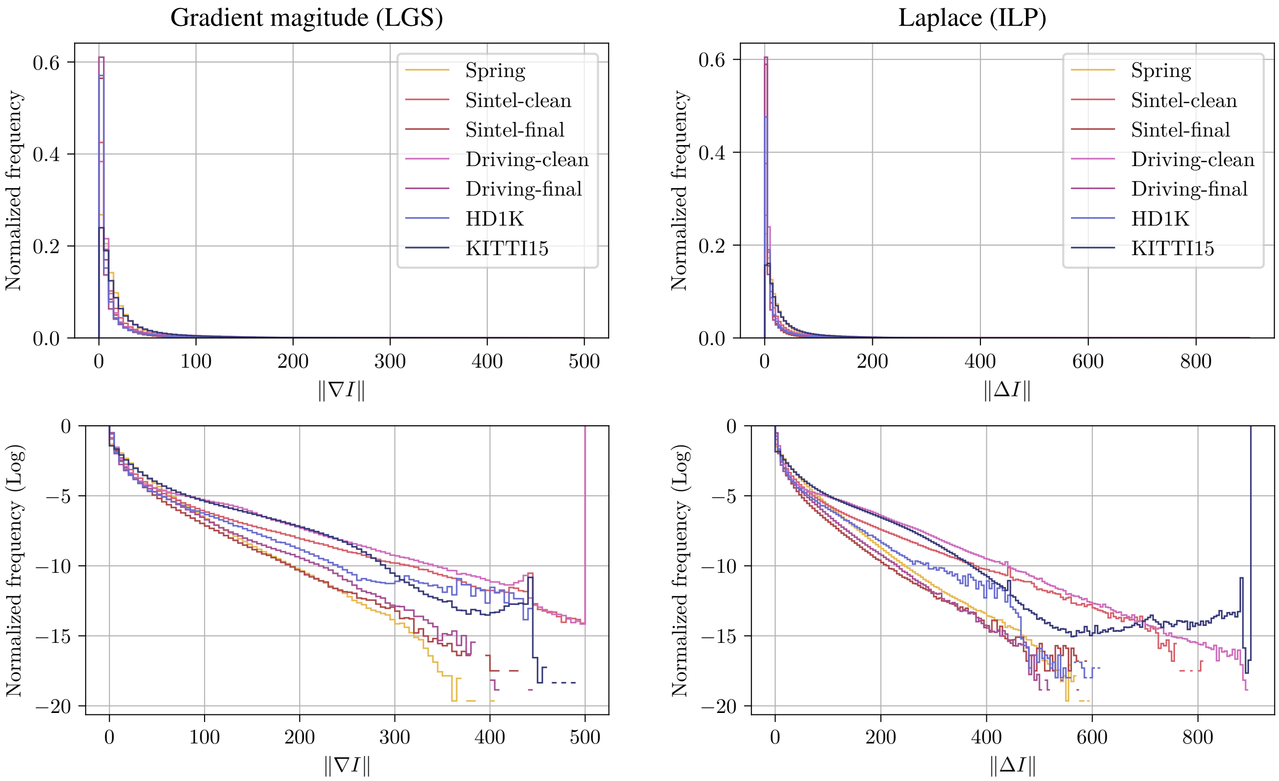


Figure A10. Image statistics for optical flow datasets. The plots show the histograms over the magnitude of first and second image derivatives for different optical flow datasets, where the LGS defense considers first (left) and ILP considers second (right) derivatives. The histograms are normalized by the number of pixels in the respective dataset. The top row shows the pure histograms, while the bottom row shows the log-transformed frequency for better visualization of statistics for large gradient magnitudes, which are filtered by the defenses.

find that applying defenses to optical flow methods worsens quality *and* robustness, which leads to a slanted line of markers per optical flow network. This indicates that for these datasets, the defenses affect the unattacked defended flow f_D as described in Sec. 6.4, Main paper, because worsening this flow enters into both, the quality calculation with $\mathrm{EPE}(f^*, f_\mathrm{D})$ and the robustness calculation with $\mathrm{EPE}_P(f_\mathrm{D}, f_\mathrm{D}^\mathrm{A})$. For the datasets Sintel and Driving in Fig. A9e, Fig. A9b, Fig. A9f and Fig. A9c, applying defenses almost exclusively changes the robustness, leading to a horizontal line of markers per optical flow network. This indicates that the defenses work "as intended", affecting only the attacked defended flow f_D^A and hence the robustness, but are on average not very effective under attack with defense-aware patches. In summary, defenses have the worst side effects on the image quality for natural or naturalistic data: KITTI and HD1K contain camera-captured real-world images and Spring is a recently rendered dataset that focuses on high-detail images. Even though defenses work partially on the synthetic datasets Sintel and Driving, which were rendered and created before 2016, they still fail to demonstrate consistent advantages over undefended networks on these datasets. These differences in the datasets are also visible in terms of the dataset image statistics that are considered by the LGS and ILP defenses. In Fig. A10 we show the histograms over first- and second-order image derivatives for all datasets. There, the synthetic Sintel and Driving datasets have a very "even" gradient magnitude decay for large gradients on the log scale (both for clean and final rendering passes), while the realistic KITTI, HD1K and Spring datasets do not show such a clear exponential gradient decay. All in all, defenses fail most severely on safety-critical real-world datasets, where reliable predictions are needed most.

References

[1] 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. 1

[2] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *Proc. European Conference on Computer Vision (ECCV)*, pages 611–625. Springer, 2012. 6, 10, 12

[3] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der

Table A6. Validation split details for the evaluation datasets. "Frames" denotes frame pairs (for the optical flow calculation) rather than single frames, if "Half" is checked the frame size is halved. If all scenes except the validation scenes make up the set for training patches, the training scenes are marked with "EV: except validation". "OFM-id" denotes "optical flow magnitude in-distribution", meaning the validation set is in-distribution w.r.t. to the optical flow magnitude distribution of the original training set. "I3" means that only every third frame pair of the validation scenes is added to the validation split.

Dataset	Val. frames	Val. scenes	Half	Train. scenes	Notes
KITTI [11]	200	KITTI-train [11]	–	Raw [4]	Split from [14]
Sintel [2]	89	ambush2, bamboo2, cave2, market2, shaman2, temple2	–	EV	Split from [20] OFM-id, I3
Driving [9]	299	15mm focal length, scene forwards, fast, into future, left	–	EV	
HD1K [7]	94	000009, 000013, 000018, 000019, 000032	✓	EV	OFM-id
Spring [10]	658	0002, 0010, 0018, 0026, 0032, 0045	✓	EV	OFM-id

Table A7. Quality $Q_D = \text{EPE}(f^*, f_D)$ for optical flow pipelines with defense D on the HD1K [7] validation split[2]; Best quality is **bold**. All defenses lead to a worse quality on unattacked frames.

Defense		FNC	FNCR	SpyNet	PWC	RAFT	GMA	FF
None	Q	**2.36**	**1.17**	**3.00**	**2.15**	**0.46**	**0.44**	**0.32**
LGS	Q_{LGS}	2.49	1.22	3.09	2.19	0.50	0.47	0.35
ILP	Q_{ILP}	2.39	1.20	3.10	2.21	0.50	0.46	0.35

Table A8. Quality $Q_D = \text{EPE}(f^*, f_D)$ for optical flow pipelines with defense D on the Spring [10] validation split[2]; Best quality is **bold**. All defenses lead to a worse quality on unattacked frames.

Defense		FNC	FNCR	SpyNet	PWC	RAFT	GMA	FF
None	Q	**0.81**	**0.48**	**0.96**	**1.87**	**0.29**	**0.29**	**0.27**
LGS	Q_{LGS}	1.17	1.09	1.39	2.21	0.66	0.67	0.44
ILP	Q_{ILP}	1.10	0.74	1.21	2.07	0.40	0.48	0.39

Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015. 1, 2

[4] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *International Journal in Robotics Research (IJRR)*, 32(11):1231–1237, 2013. 10

Table A9. Quality $Q_D = \text{EPE}(f^*, f_D)$ for optical flow pipelines with defense D on the Sintel [2] clean and final validation splits[2] [20]; Best quality is **bold**. All defenses lead to a worse quality on unattacked frames.

Defense		FNC	FNCR	SpyNet	PWC	RAFT	GMA	FF
		clean						
None	Q	**4.80**	**2.31**	**5.66**	**3.72**	**0.84**	**0.77**	**0.45**
LGS	Q_{LGS}	4.83	2.34	5.81	3.75	0.86	0.79	0.48
ILP	Q_{ILP}	4.90	2.37	5.70	3.77	0.85	0.80	0.49
		final						
None	Q	**5.79**	**4.07**	7.64	**5.42**	**1.49**	**1.45**	**0.74**
LGS	Q_{LGS}	5.82	4.16	7.66	5.51	1.56	1.51	0.79
ILP	Q_{ILP}	5.81	4.10	**7.63**	5.46	1.52	1.47	0.76

Table A10. Quality $Q_D = \text{EPE}(f^*, f_D)$ for optical flow pipelines with defense D on the Driving [9] clean and final validation splits[2]; Best quality is **bold**. All defenses lead to a worse quality on unattacked frames.

Defense		FNC	FNCR	SpyNet	PWC	RAFT	GMA	FF
		clean						
None	Q	**95.73**	**88.85**	**111.36**	92.49	37.32	**54.17**	**51.94**
LGS	Q_{LGS}	96.20	89.64	113.15	92.93	**36.63**	60.21	53.11
ILP	Q_{ILP}	95.99	89.43	112.32	**93.06**	36.75	60.06	53.21
		final						
None	Q	**90.21**	**84.34**	**110.52**	**92.19**	**40.87**	**62.35**	**47.05**
LGS	Q_{LGS}	90.78	85.07	111.84	92.70	41.96	62.68	47.52
ILP	Q_{ILP}	90.47	84.59	110.91	92.52	41.60	62.61	47.56

[5] 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. 1, 2

[6] 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. 1, 2

[7] Daniel Kondermann, Rahul Nair, Katrin Honauer, Karsten Krispin, Jonas Andrulis, Alexander Brock, Burkhard Gussefeld, Mohsen Rahimimoghaddam, Sabine Hofmann, Claus Brenner, and Bernd Jähne. 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. 6, 10, 11

[8] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *Proc. International*

Table A11. Robustness scores for all combinations of defended methods and defense-aware attacks on optical flow methods on the HD1K [7] validation split[2]. Per attack, the robustness values of the best defense are **bold**. Per defense, the robustness values for the attack it is most vulnerable to are underlined. Full pipelines are highlighted in gray, and provide the corresponding robustness values to the quality scores from Tab. A7.

Attack type	Defense		FNC	FNCR	SpyNet	PWC	RAFT	GMA	FF
Vanilla	None	R^{Van}	$\underline{67.24}$	$\underline{0.23}$	**0.36**	$\underline{0.25}$	$\underline{\mathbf{0.11}}$	$\underline{\mathbf{0.07}}$	$\underline{\mathbf{0.22}}$
	LGS	$R^{\text{Van}}_{\text{LGS}}$	**0.45**	0.21	0.59	0.25	$\underline{0.23}$	0.17	0.36
	ILP	$R^{\text{Van}}_{\text{ILP}}$	0.60	**0.17**	0.51	**0.24**	0.17	0.13	$\underline{0.22}$
+LGS (LGS-aware)	None	R^{LGS}	51.97	**0.07**	**0.35**	**0.22**	**0.09**	**0.06**	**0.17**
	LGS	$R^{\text{LGS}}_{\text{LGS}}$	$\underline{13.47}$	$\underline{0.22}$	$\underline{0.62}$	$\underline{0.33}$	0.23	0.21	0.38
	ILP	$R^{\text{LGS}}_{\text{ILP}}$	**10.02**	0.17	0.51	0.35	$\underline{0.18}$	$\underline{0.18}$	0.22
+ILP (ILP-aware)	None	R^{ILP}	60.52	**0.14**	$\underline{\mathbf{0.36}}$	**0.24**	**0.06**	**0.06**	**0.18**
	LGS	$R^{\text{ILP}}_{\text{LGS}}$	**2.89**	0.21	0.62	0.27	0.23	$\underline{0.21}$	$\underline{0.38}$
	ILP	$R^{\text{ILP}}_{\text{ILP}}$	$\underline{53.63}$	$\underline{0.21}$	$\underline{0.52}$	$\underline{0.37}$	0.16	0.18	0.22

Table A12. Robustness scores for all combinations of defended methods and defense-aware attacks on optical flow methods on the Spring [10] validation split[2]. Per attack, the robustness values of the best defense are **bold**. Per defense, the robustness values for the attack it is most vulnerable to are underlined. Full pipelines are highlighted in gray, and provide the corresponding robustness values to the quality scores from Tab. A8.

Attack type	Defense		FNC	FNCR	SpyNet	PWC	RAFT	GMA	FF
Vanilla	None	R^{Van}	$\underline{69.64}$	$\underline{\mathbf{0.17}}$	**0.06**	**0.10**	$\underline{\mathbf{0.04}}$	$\underline{\mathbf{0.13}}$	$\underline{\mathbf{0.05}}$
	LGS	$R^{\text{Van}}_{\text{LGS}}$	0.62	0.67	0.55	0.54	0.40	$\underline{0.42}$	$\underline{0.24}$
	ILP	$R^{\text{Van}}_{\text{ILP}}$	**0.58**	0.31	0.33	0.36	0.15	$\underline{0.22}$	$\underline{0.17}$
+LGS (LGS-aware)	None	R^{LGS}	33.27	**0.02**	$\underline{\mathbf{0.06}}$	**0.16**	**0.02**	**0.02**	**0.03**
	LGS	$R^{\text{LGS}}_{\text{LGS}}$	$\underline{\mathbf{3.06}}$	0.66	$\underline{0.55}$	$\underline{0.55}$	$\underline{0.41}$	0.42	0.24
	ILP	$R^{\text{LGS}}_{\text{ILP}}$	11.87	0.32	$\underline{0.33}$	$\underline{0.42}$	$\underline{0.15}$	0.22	0.17
+ILP (ILP-aware)	None	R^{ILP}	30.01	**0.10**	**0.06**	$\underline{\mathbf{0.19}}$	**0.01**	**0.02**	**0.03**
	LGS	$R^{\text{ILP}}_{\text{LGS}}$	**0.74**	$\underline{0.67}$	0.55	0.55	0.41	0.42	0.24
	ILP	$R^{\text{ILP}}_{\text{ILP}}$	$\underline{21.22}$	$\underline{0.33}$	0.33	0.42	0.15	0.22	0.17

Conference on Learning Representations (ICLR), 2017. 1

[9] 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. 6, 10, 12

[10] 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. 6, 10, 11

[11] 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. 6, 10

[12] Muzammal Naseer, Salman Khan, and Fatih Porikli. Local gradients smoothing: Defense against localized adversarial attacks. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1300–1307, 2019. 1

[13] Anurag Ranjan and Michael J. Black. Optical flow estimation using a spatial pyramid network. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[14] Anurag Ranjan, Joel Janai, Andreas Geiger, and Michael J. Black. Attacking optical flow. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 10

[15] 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

[16] 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. 1, 2

[17] Deqing Sun, Charles Herrmann, Fitsum Reda, Michael Rubinstein, David J. Fleet, and William T. Freeman. Disentangling architecture and training for optical flow. In *Proc. European Conference on Computer Vision (ECCV)*, pages 165–

Table A13. Robustness scores for all combinations of defended methods and defense-aware attacks on optical flow methods on the Sintel [2] final (f) and clean (c) validation splits[2] [20]. Per attack, the robustness values of the best defense are **bold**. Per defense, the robustness values for the attack it is most vulnerable to are underlined. Full pipelines are highlighted in gray, and provide the corresponding robustness values to the quality scores from Tab. A9.

Attack type	Defense		FNC		FNCR		SpyNet		PWC		RAFT		GMA		FF	
			f	c	f	c	f	c	f	c	f	c	f	c	f	c
Vanilla	None	R^{Van}	68.71	67.84	1.60	0.84	**1.12**	**1.27**	1.21	0.84	0.43	0.19	4.72	1.86	0.33	0.12
	LGS	$R^{\text{Van}}_{\text{LGS}}$	**1.06**	**0.90**	0.73	0.25	1.45	1.41	0.86	0.52	0.40	0.17	0.34	0.13	0.19	**0.09**
	ILP	$R^{\text{Van}}_{\text{ILP}}$	1.49	1.76	**0.63**	**0.29**	1.16	1.34	**0.74**	**0.51**	**0.19**	**0.14**	**0.24**	**0.13**	**0.13**	0.12
+LGS (LGS-aware)	None	R^{LGS}	57.21	57.11	0.44	**0.20**	**1.15**	**1.23**	0.90	**0.61**	0.23	**0.10**	0.20	**0.09**	0.16	**0.07**
	LGS	$R^{\text{LGS}}_{\text{LGS}}$	**20.04**	**34.78**	**0.72**	0.28	1.53	1.53	1.08	0.66	0.33	0.14	0.35	0.13	0.18	0.10
	ILP	$R^{\text{LGS}}_{\text{ILP}}$	31.42	40.56	0.63	0.35	1.23	1.33	**0.88**	0.64	**0.20**	0.13	**0.19**	0.15	**0.12**	0.12
+ILP (ILP-aware)	None	R^{ILP}	68.58	68.13	1.13	0.55	**1.17**	**1.26**	0.96	0.65	**0.20**	**0.09**	**0.19**	**0.09**	0.19	**0.08**
	LGS	$R^{\text{ILP}}_{\text{LGS}}$	**2.54**	**24.19**	**0.74**	**0.27**	1.53	1.53	0.98	**0.64**	0.35	0.16	0.35	0.13	0.19	0.10
	ILP	$R^{\text{ILP}}_{\text{ILP}}$	53.52	65.18	1.04	0.49	1.25	1.36	**0.93**	0.67	0.20	0.14	0.20	0.15	**0.12**	0.12

Table A14. Robustness scores for all combinations of defended methods and defense-aware attacks on optical flow methods on the Driving [9] final (f) and clean (c) validation splits[2]. Per attack, the robustness values of the best defense are **bold**. Per defense, the robustness values for the attack it is most vulnerable to are underlined. Full pipelines are highlighted in gray, and provide the corresponding robustness values to the quality scores from Tab. A10.

Attack type	Defense		FNC		FNCR		SpyNet		PWC		RAFT		GMA		FF	
			f	c	f	c	f	c	f	c	f	c	f	c	f	c
Vanilla	None	R^{Van}	93.15	3.64	4.72	5.94	**4.39**	**4.22**	7.30	7.77	4.74	2.85	4.69	4.58	4.10	4.04
	LGS	$R^{\text{Van}}_{\text{LGS}}$	**4.43**	**3.25**	3.39	**2.12**	7.72	5.33	6.83	**6.69**	5.49	3.41	5.18	3.40	3.91	3.44
	ILP	$R^{\text{Van}}_{\text{ILP}}$	4.88	6.05	**2.54**	2.23	5.54	4.90	**5.73**	7.33	**3.31**	**2.83**	**3.43**	**3.19**	**3.26**	**3.38**
+LGS (LGS-aware)	None	R^{LGS}	74.11	**4.11**	**1.30**	**1.64**	**4.25**	**4.07**	**6.63**	**7.00**	4.19	**2.90**	**3.14**	**3.01**	3.53	**3.71**
	LGS	$R^{\text{LGS}}_{\text{LGS}}$	24.93	8.61	3.55	3.05	7.90	5.68	8.24	8.02	5.44	3.72	5.67	3.91	4.05	4.18
	ILP	$R^{\text{LGS}}_{\text{ILP}}$	**24.21**	5.93	2.48	4.07	5.49	5.18	7.96	7.90	**3.98**	3.77	3.51	3.87	**3.16**	4.54
+ILP (ILP-aware)	None	R^{ILP}	84.98	**5.34**	**2.70**	3.49	**4.34**	**4.17**	**6.82**	**7.44**	**3.05**	**2.56**	**3.15**	**2.96**	3.74	**3.93**
	LGS	$R^{\text{ILP}}_{\text{LGS}}$	**7.92**	8.20	3.32	**2.38**	7.86	5.68	7.25	7.58	5.46	3.69	5.52	4.04	4.16	4.42
	ILP	$R^{\text{ILP}}_{\text{ILP}}$	74.86	5.00	2.99	3.80	5.55	4.91	8.24	8.30	3.28	3.27	3.53	3.93	**3.25**	4.73

182, 2022. 6

[18] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. 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. 1, 2

[19] 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. 1, 2

[20] Gengshan Yang and Deva Ramanan. Volumetric correspondence networks for optical flow. In *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc., 2019. 6, 10, 12