Towards Understanding Adversarial Robustness of Optical Flow Networks Supplementary Material

1. Implementation Details

We provide implementation details for different attacks and their evaluation in the following. In all our experiments, we used pre-trained models *without* fine-tuning on the KITTI dataset. We built upon the works of Ranjan *et al*. [10] for adversarial patch attacks, Wong *et al*. [15] for (global) adversarial perturbation attacks, and Teed *et al*. [13] for training of flow networks, All code is available at https://github.com/lmb-freiburg/understanding_flow_robustness.

Adversarial patch attacks. For adversarial patch attacks, we followed the attacking and white-box evaluation procedure of Ranjan *et al*. [10]. We optimized a circular patch by optimizing w.r.t. Equation 1 using the flow networks' predictions as pseudo ground truth from the raw KITTI 2012 dataset [3] for adversarial optimization and the annotated images as the validation set. We used scale augmentation within $\pm 5\%$, rotation augmentation within $\pm 10°$ and randomly pasted the patch at different image locations, but at the *same* location in both image frames.

For evaluation of patch-based experiments, we used the KITTI 2015 training set [8] and resized images to 384×1280. During the evaluation, we pasted the patch also at the *same* location in both image frames, if stated not otherwise. We always computed the unattacked and attacked End-Point-Error (EPE) and set the ground truth region occluded by the patch to zero motion. For the computation of the spatial location heat map, we moved patches in strides of 25 pixels in x- and y-direction to reduce the computational demands. For the t-SNE [14] plots, we extracted the feature maps from the flow networks, computed the mean over the spatial dimensions to reduce the dimensionality, and computed the t-SNE embeddings on them. For experiments with Robust FlowNetC and its variants, we optimized > 20 or 10 adversarial patches, respectively, across various learning rates for each patch size. We chose the three worst patches in terms of attacked EPE on the validation set, computed the spatial location heat map to get the worst-case attacked EPE and report the highest worst-case attacked EPE of the three. Other patches were not as effective as the three worst patches. Finally, we also tested moving the patch between image frames. For this, we randomly sampled translation within ± 50, full rotation (*i.e*. $\pm 180°$) and scale augmentation within $\pm 5\%$.

Network	Unattacked EPE	Attacked 102x102 (2.1%)	Attacked 153x153 (5.8%)
FlowNetC as [10]	14.52	94.51	197.00
FlowNetC as [5]	11.50	52.66	51.99
FlowNet2 as [10]	11.82	27.59	43.14
FlowNet2 as [5]	10.07	12.40	13.36

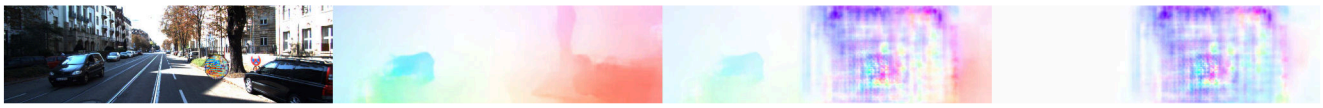
Table 1. **Adversarial patch attacks with different input data normalizations.** We show average unattacked and average worst-case attacked EPE on the KITTI 2015 training dataset.

Adversarial perturbation attacks. For adversarial perturbation attacks, we used the same procedure as Wong *et al*. [15], but pre-trained models were *not fine-tuned* on the KITTI dataset and minimized the l_2 loss, as it led to more severe flow performance deterioration compared to the l_1 or $cossim$ losses. We used the Iterative Fast Gradient Sign Method (I-FGSM) [6] for crafting adversarial perturbations. For brevity, we considered only our proposed Robust FlowNetC, PWC-Net, and RAFT. We used the KITTI 2015 training set [8] for evaluation and resized images to 256×640 due to computational limitations. We used the L_∞ norms $\epsilon = \{0.02, 0.01, 0.005, 0.002\}$ and momentum $\beta = 0.47$, but with the same learning rate $\alpha = 0.002$ each.

For targeted adversarial attacks, we used the same hyperparameters but minimized the l_2 loss. We used more steps (*i.e*., 100) for the target flow depicting the number 42. For universal adversarial attacks, we used the same procedure as Ranjan *et al*. [10] but optimized for a universal perturbation instead of a patch. We used I-FGSM with 5 steps, and all other hyperparameters remained unchanged.

2. Note on Input Data Normalization

In the typical deep learning setting, we normalize the input data in the preprocessing step, as this usually leads to faster convergence [7]. Since we use input data normaliza-



(a) Normalization scheme: Ranjan *et al*. [10]. Unattacked EPE: 22.71, Attacked EPE: 60.24.

(b) Normalization scheme: Ilg *et al*. [5]. Unattacked EPE: 16.18, Attacked EPE: 30.03.

Figure 1. **The use of different input data normalizations leads to different results for FlowNetC.** We optimized and evaluated a 102×102 patch for each input data normalization. Visualizations from left to right: the attacked first frame, the unattacked flow estimate, the attacked flow estimate, the difference between the attacked and unattacked optical flow estimates. Best viewed in color and with zoom.

tion during training, we must also use the *same* data normalization during inference, since the model learned based on these normalized inputs. Using the wrong input data normalization usually has a detrimental effect on performance.

We found that Ranjan *et al*. [10] normalized inputs of FlowNetC and FlowNet2 to the interval $[-1, 1]$, which is different from the input data normalization FlowNetC and FlowNet2 used during their training. More specifically, Ilg *et al*. [5] first normalize inputs to $[0, 1]$ and then subtract the mean of each RGB channel computed during the first 1000 iterations in training. As a result, FlowNetC's and FlowNet2's unattacked and attacked EPEs on the KITTI 2015 training dataset [8] drop significantly (Table 1 and Figure 1). However, despite this correction of the input data normalization, FlowNetC is still vulnerable, so the result of Ranjan *et al*. [10] is still valid.

3. Additional Examples for Handcrafted Patch Attacks

In Figure 2 we show additional results for our circular high-frequency black and white vertically striped patch for FlowNetC. We only show results for FlowNetC, since it is the most vulnerable flow network and thus shows the most severe effect in the optical flow estimates. Similar to optimized patches, our handcrafted patch severely deteriorates the optical flow estimates.

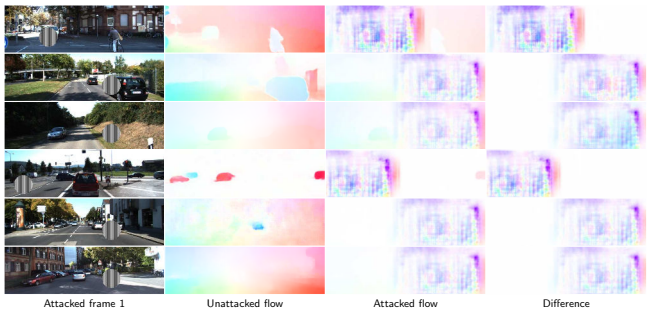
4. Ingredients for Handcrafted Patch Attacks

We conducted several ablations to identify the main ingredients for a successful handcrafted patch attacks besides its self-similar pattern. We chose FlowNetC as the flow network for the ablations because it is the most vulnerable w.r.t. patch-based attacks. To study the influence of the contrast between the stripes, we fixed the black or white color of our handcrafted patch and changed the respective other color, thereby changing the contrast between the stripes. Figure 3 shows that higher contrasts between the stripes cause more severe deteriorations of optical flow performance. Interestingly, we observe an exponential increase in worst-case attacked EPE with the increase in the contrast between the stripes. The handcrafted self-similar pattern also works when we use different color pairs (Figure 4). However, the effect of the handcrafted patch may be less severe for different color pairs. Note that regions with zero flow are more vulnerable w.r.t. patch-based attacks. Furthermore, our handcrafted patch requires high-frequency self-similar patterns to remain effective (Figure 5). The larger the strip thickness (*i.e*., the lower the frequency), the smaller the effect of the handcrafted patch attack. Finally, the handcrafted patch attack is more effective when self-similar patterns are oriented in axial directions (Figure 6).

5. Increasing the Receptive Field Size by Increasing the Dilation Rate

In the main paper, we showed that increasing the receptive field by increasing the network depth helps improve robustness. Alternatively, we also tried to increase the receptive field by increasing the dilation rate of the convolutional layers of FlowNetC's encoder. We used dilation rates $\{1, 2, 4, 8\}$, where a dilation rate of 1 corresponds to the original FlowNetC. Figure 7 shows that increasing the dilation rates also makes FlowNetC more robust w.r.t. patch-based attacks. The gap between unattacked and worst-case attacked EPE can be mainly attributed to occlusions causing optical flow performance to deteriorate. However, the flow performance for unattacked image pairs deteriorates significantly at larger dilation rates. More explicitly, the FlowNetC variant with a dilation rate of 8 has unattacked EPE 18.81 and worst-case attacked EPEs 23.4 and 22.66 for optimized adversarial patches with patch sizes 102×102 and 153×153, respectively. Note, however, that a uniform noise patch also has EPEs of 23.32 or 22.3 for patch sizes 102×102 and 153×153, respectively. Therefore, increasing the receptive field by adding more depth is preferable to make flow networks robust w.r.t. patch-based attacks.

(a) 102×102 handcrafted patch.



(b) 153×153 handcrafted patch.

Figure 2. **Additional examples for the handcrafted patch attack.** We show the handcrafted patch at the worst possible spatial location for FlowNetC [2]. Our handcrafted patch leads to severe deteriorations of the optical flow estimates. Best viewed in color and with zoom.

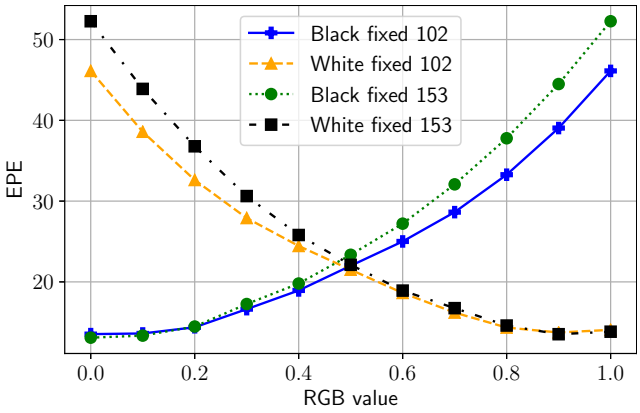


Figure 3. **Contrast between stripes of the handcrafted patch.** We fixed the black or white color of the black and white vertical striped patch fixed and moved the respective other color towards white and black, respectively. We attacked FlowNetC with these variants of our handcrafted patch. Interestingly, worst-case attacked EPE increases exponentially with contrast.

6. Additional Examples for Robust FlowNetC

Figure 8 shows that Robust FlowNetC is also robust against optimized patches. However, as with other flow networks, we would like to stress that some particular hard image frames can cause severe deterioration of flow performance.

7. Realistic Motion of Patches

We also tried to use realistic motion of patches by considering them as part of the static scene, as described by Ranjan *et al*. [10]. We found that it has a negligible effect w.r.t. the worst-case attacked EPE for Robust FlowNetC, *i.e*., 12.16 and 12.11 to 13.60 and 14.57 for 102×102 or 153×153 patches, respectively. We found the reason for higher worst-case attacked EPE is due to the placement of patches at boundary regions of the first image frame so that they disappeared in the second image frame.

8. Untargeted Adversarial Attacks

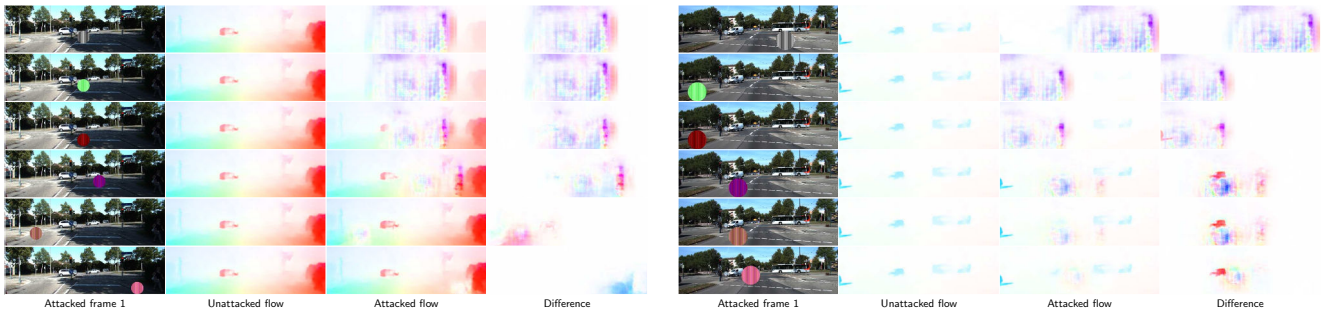
Wong *et al*. [15] showed that they could attack disparity estimation networks by adding an adversarial perturbation individually to each pixel. Figures 9 and 10 show that, as expected, the same is true for optical flow networks.

9. Additional Examples for Targeted Adversarial Attacks

Figure 11 shows additional examples for targeted adversarial attacks on optical flow networks. The flow estimates are closer to the adversarial target flow than the true flow. To quantify our results, also for different L_∞ norms (*i.e*., $\epsilon = \{0.002, 0.005, 0.01, 0.02\}$), we ran targeted adversarial attacks for different image pairs from the KITTI 2015 training dataset. Due to computational reasons, we randomly picked a subset of 10 image pairs. Figure 12 shows that the resulting flow is closer to the adversarial target flow than the true flow. Note that the resulting flow is closer to the adversarial target flow when the L_∞ norm is larger.

10. Examples for Adversarial Universal Attacks

Figure 13 shows universal perturbations for the L_∞ norm $\epsilon = 0.02$. Note that we can observe well-visible self-similar patterns for Robust FlowNetC and PWC-Net. Figure 14 shows examples for universal attacks with L_∞ norm $\epsilon = 0.02$. The flow networks are largely unaffected by the universal perturbations. However, there are some worst-case examples: if there are darker, homogeneous areas, *e.g*., shadows, in the image frames (and/or there is large ego-motion), the flow deteriorates more. However, this is to be expected because the lower contrast (and large ego-motion) make the estimation problem more difficult, leading to more ambiguities. An attacker could exploit this by overwriting the true flow with the help of adversarial ambiguities.



(a) 102×102 patches. (b) 153×153 patches.

Figure 4. **Different color pairs of the handcrafted patch.** We used the same self-similar pattern with different color pairs. From top to bottom for each subfigure: black-white, green-white, red-black, red-blue, green-violet, and violet-orange. We show each handcrafted patch at the worst possible spatial location for FlowNetC [2]. The handcrafted patch attack also works with different color pairs. However, the effect of the handcrafted patch may be less severe for some color pairs. Best viewed in color and with zoom.

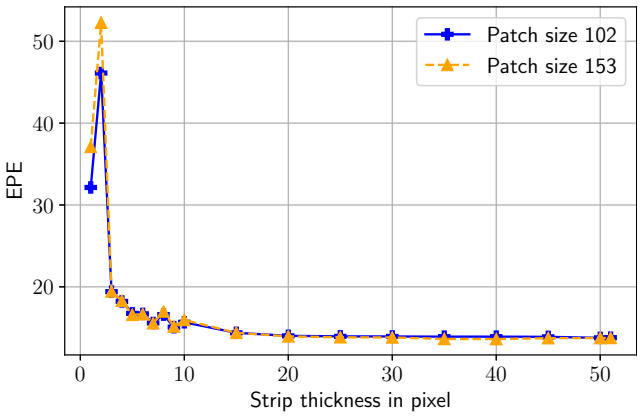


Figure 5. **Strip thickness of the handcrafted patch.** We altered the thickness of the stripes and attacked FlowNetC with these variants. The handcrafted patch requires high-frequency self-similar pattern to remain effective.

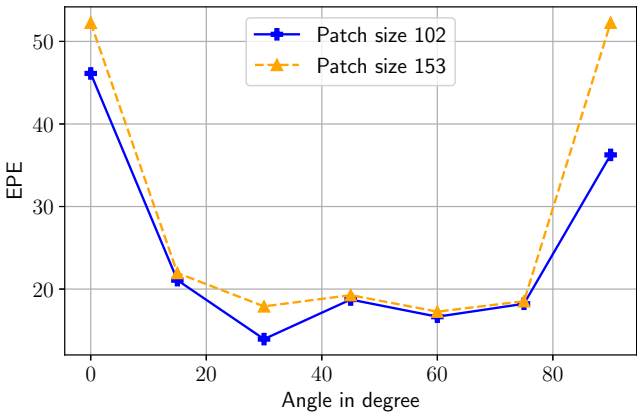


Figure 6. **Rotational orientation of the handcrafted patch.** We rotated our handcrafted patch and attacked FlowNetC. The rotational angle of 0° corresponds to vertical stripes and the rotational angle 90° corresponds to horizontal stripes. We observe a U-shaped form of worst-case attacked EPE: the stripes oriented in the axial directions cause more deterioration of flow performance.

11. Adversarial Data Augmentation

Wong *et al*. [15] showed that they could increase robustness with little negative effect on performance through adversarial data augmentation. To do this, they crafted adversarial examples using FGSM before the adversarial training and added them to the training set.

Different from Wong *et al*., we did not pre-compute adversarial examples but computed them during the training (as typically done in adversarial training for recognition networks). We crafted adversarial examples with the I-FGSM attack (with various L_∞ norms $\epsilon = \{0.002, 0.005, 0.01, 0.02\}$). We set hyperparameters for the untargeted adversarial attack, as described in Supplement Section 1. We used all 194 image pairs of the KITTI 2012 test dataset for adversarial training and resized images to 256×640. We chose learning rates 0.000125 for RAFT, and 0.00001 for Robust FlowNetC and PWC-Net, and weight decays 0.0001 for all flow networks. We fine-tuned the flow networks for 30000 steps with batch size 2 (*i.e*., the unattacked and attacked image frames). We did not apply any other data augmentation to the images. For evaluation, we crafted new, unseen (untargeted) adversarial examples on the KITTI 2015 training dataset, as described in Supplement Section 1. In addition, we attacked with the MI-FGSM attack [1] to evaluate the robustness of flow networks against a stronger, unseen adversarial attack.

Figure 15 shows that fine-tuning with adversarial data augmentation improves the robustness of all flow networks. Surprisingly, the adversarial trained flow networks are also robust against the stronger MI-FGSM attacks. Similar to Wong *et al*., we find that contrary to findings in classification [6], training with adversarial data augmentation has little negative effect on the performance of optical flow

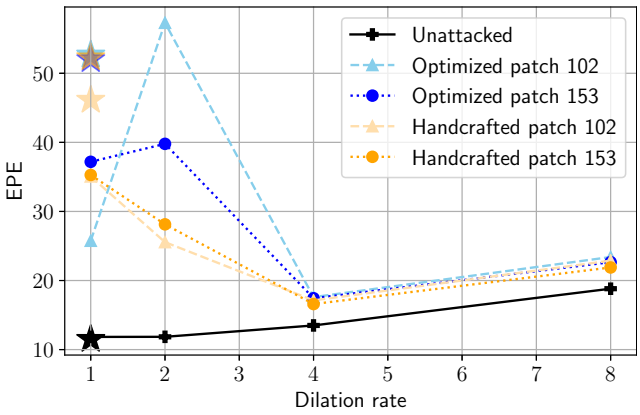


Figure 7. **Performance of FlowNetC variants with various dilation rates.** We show both unattacked and worst-case attacked EPE. Stars show results for the original FlowNetC. For optimized patches, we show results using the patch with the highest worst-case attacked EPE after optimization over ten runs. Increasing the dilation rate also improves the robustness against patch-based attacks, but overall flow performance deteriorates.

networks (except for Robust FlowNetC). For example, for RAFT, EPE only deteriorates from 5.47 to 5.96, 7.13, 8.96 and 10.15 for L_∞ norms 0.002, 0.005, 0.01 and 0.02 on unattacked image frames, respectively. In general, the smaller the norm the less drop in EPE on unattacked image frames. On the contrary, however, the larger the L_∞ norm, the higher the robustness against adversarial attacks. However, unlike Wong *et al*., we found that training on smaller L_∞ norms (*e.g*., $\epsilon = 0.002$) cannot (significantly) improve robustness on large L_∞ norms (*e.g*., $\epsilon = 0.02$). We leave further analysis for future work.

12. Common Image Corruptions

In this paper, we focused on adversarial attacks. However, (white-box) adversarial attacks are difficult to apply in the real world, and common image corruptions [4], *e.g*., snow, are more likely to occur. Thus, for completeness, we also studied the robustness of flow networks against common image corruptions across various severities [9]. Similar to our patch-based experiments, we also resized images to 384×1280. Figure 16 shows that all flow networks are robust against most common image corruptions. However, there are corruptions (*e.g*., Frost, Snow, Impulse noise, Gaussian noise, Shot noise) that can cause severe deterioration of flow estimates. We suspect that this deterioration is due to some part to the superposition of another flow, *e.g*., snowfall. We leave further analysis for future work.

References

[1] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018. 4

[2] 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 *ICCV*, 2015. 3, 4

[3] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1

[4] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ICLR*, 2019. 5

[5] 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 *CVPR*, 2017. 1, 2

[6] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *ICLR*, 2017. 1, 4

[7] Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*. 2012. 1

[8] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015. 1, 2

[9] 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*, 2019. 5

[10] Anurag Ranjan, Joel Janai, Andreas Geiger, and Michael J Black. Attacking optical flow. In *ICCV*, 2019. 1, 2, 3

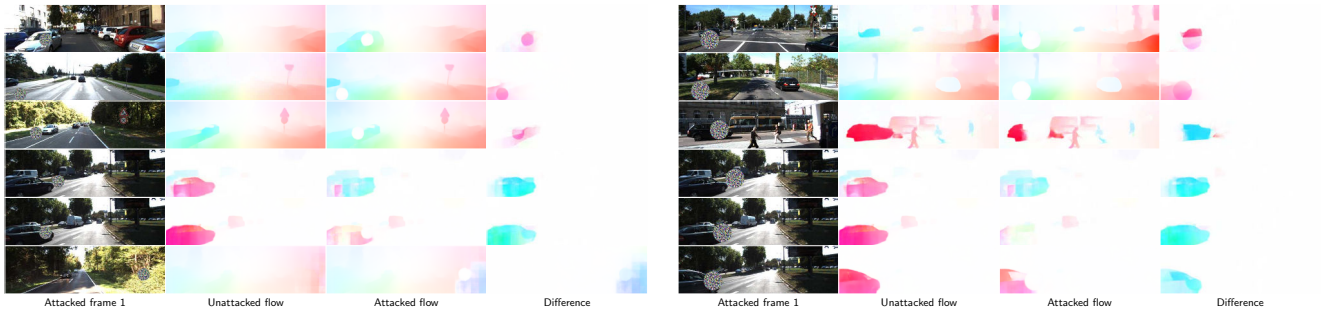
[11] Tonmoy Saikia, Cordelia Schmid, and Thomas Brox. Improving robustness against common corruptions with frequency biased models. In *ICCV*, 2021. 10

[12] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018. 6, 7, 8, 9, 10

[13] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 1, 6, 7, 8, 9, 10

[14] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008. 1

[15] Alex Wong, Mukund Mundhra, and Stefano Soatto. Stereopagnosia: Fooling stereo networks with adversarial perturbations. In *AAAI*, 2021. 1, 3, 4



(a) 102×102 optimized patch. (b) 153×153 optimized patch.

Figure 8. **Additional examples for Robust FlowNetC.** We show the best found optimized patch at the worst possible spatial location for Robust FlowNetC. The bottom three rows of each subfigure show the worst examples based on the greatest absolute degradation of worst-case attacked EPE in descending order. Best viewed in color and with zoom.

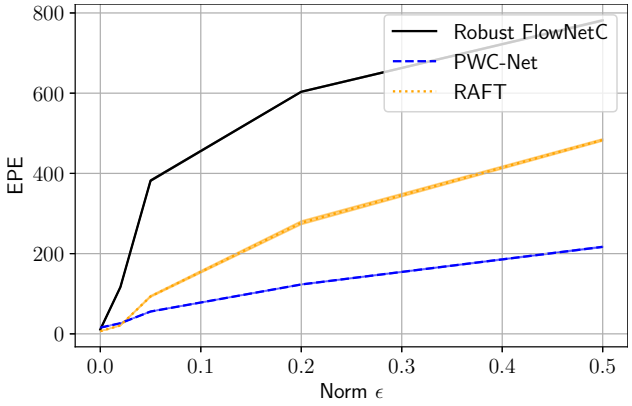
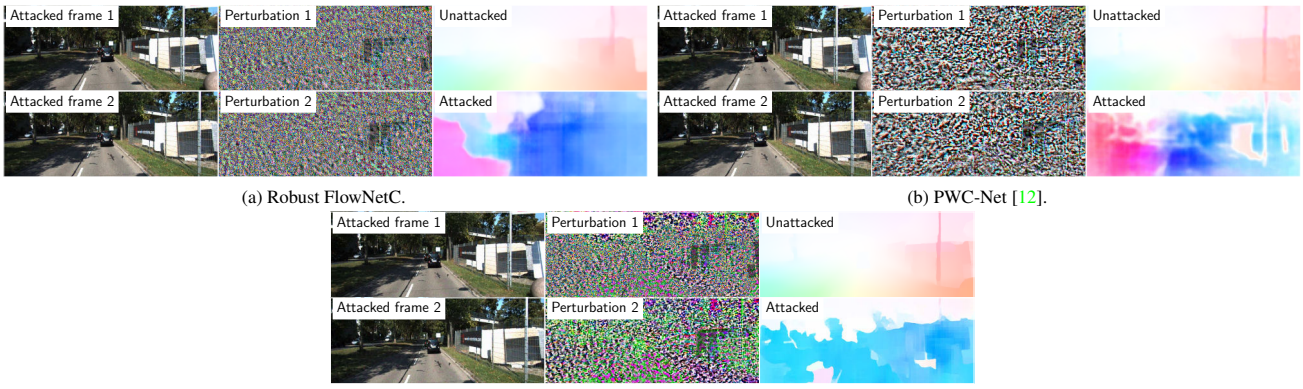


Figure 9. **Untargeted adversarial attacks.** We attacked flow networks with I-FGSM on the KITTI 2015 training dataset over various L_∞ norms $\epsilon = \{0.002, 0.005, 0.01, 0.02\}$. I-FGSM significantly deteriorates optical flow performance for all flow networks.



(a) Robust FlowNetC. (b) PWC-Net [12].

(c) RAFT [13].

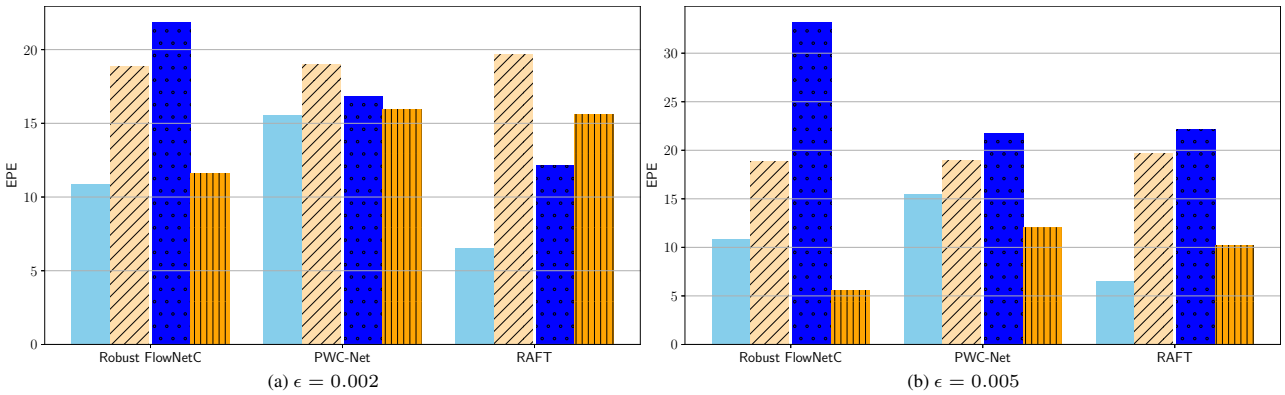
Figure 10. **Examples for untargeted adversarial attacks for different flow networks.** In each subfigure, we show in the first row the attacked first frame with perturbation, the perturbation of the first frame, and the unattacked flow estimate, and in the second row, the attacked second frame with perturbation, the perturbation of the second frame, and the attacked flow estimate. The crafted (imperceptible) adversarial perturbations completely distort the optical flow estimates. Best viewed in color and with zoom.

(a) Robust FlowNetC.

(b) PWC-Net [12].

(c) RAFT [13].

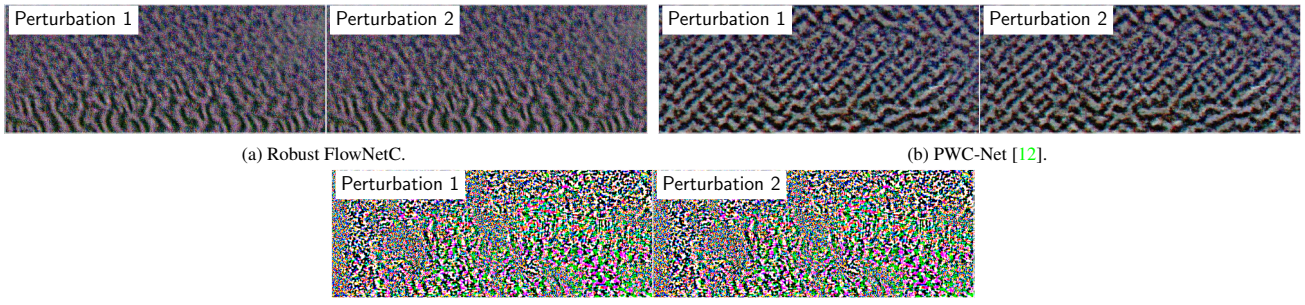
Figure 11. **Additional examples for targeted adversarial attacks.** Targeted adversarial attacks with I-FGSM with L_∞ norm $\epsilon = 0.02$. Each row in each subfigure, the ground truth of the left or right block, is the adversarial target flow of the right or left block. For each block in each subfigure, we show both attacked image frames with perturbations in the first column, the perturbations for both image frames in the second column, and the unattacked and attacked flow estimate in the third column. Note that the flow estimates are closer to the target flows than the actual flows.



(a) $\epsilon = 0.002$ (b) $\epsilon = 0.005$

(c) $\epsilon = 0.01$ (d) $\epsilon = 0.02$

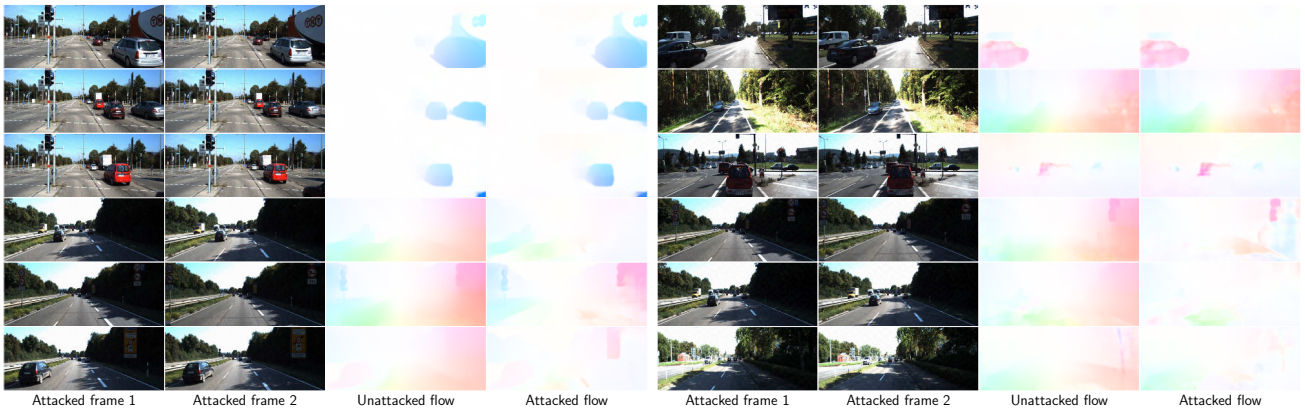
Figure 12. **Targeted adversarial attacks for different flow networks and L_∞ norms.** We show the effectiveness of targeted adversarial attacks across different flow networks and L_∞ norms (*i.e.*, $\epsilon = \{0.002, 0.005, 0.01, 0.02\}$). We computed EPE for each of the unattacked and attacked image frames for both the true ground truth (true GT) and adversarial target ground truth (target GT). Across all L_∞ norms the flow estimates get closer to the adversarial target flow. The larger the L_∞ norm (*e.g.*, $\epsilon = 0.02$), the larger the shift.



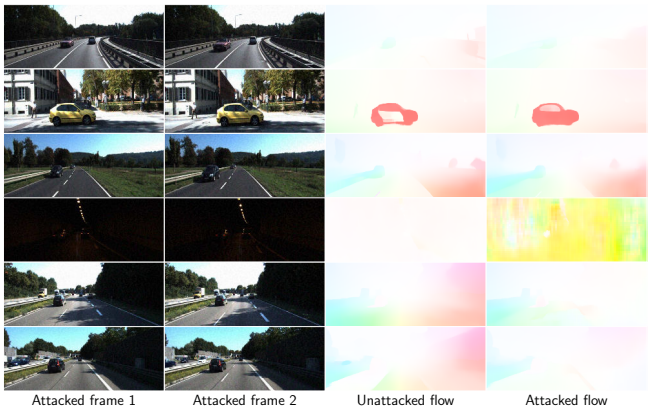
(a) Robust FlowNetC. (b) PWC-Net [12].

(c) RAFT [13].

Figure 13. **Adversarial universal perturbations.** We show the best found adversarial universal perturbations with L_∞ norm $\epsilon = 0.02$ for each flow network for the first and second image frame left or right, respectively. Note that for both Robust FlowNetC and PWC-Net the adversarial universal perturbations contain well-visible self-similar patterns. Best viewed in color and with zoom.

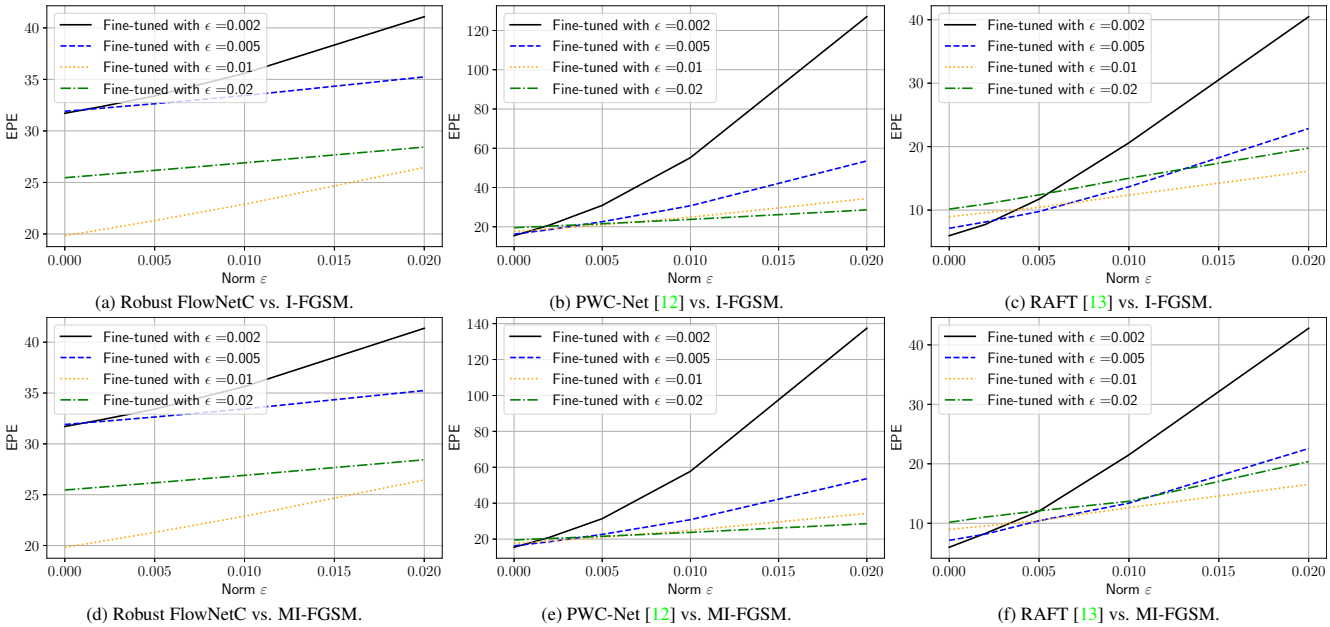


(a) Robust FlowNetC. (b) PWC-Net [12].



(c) RAFT [13].

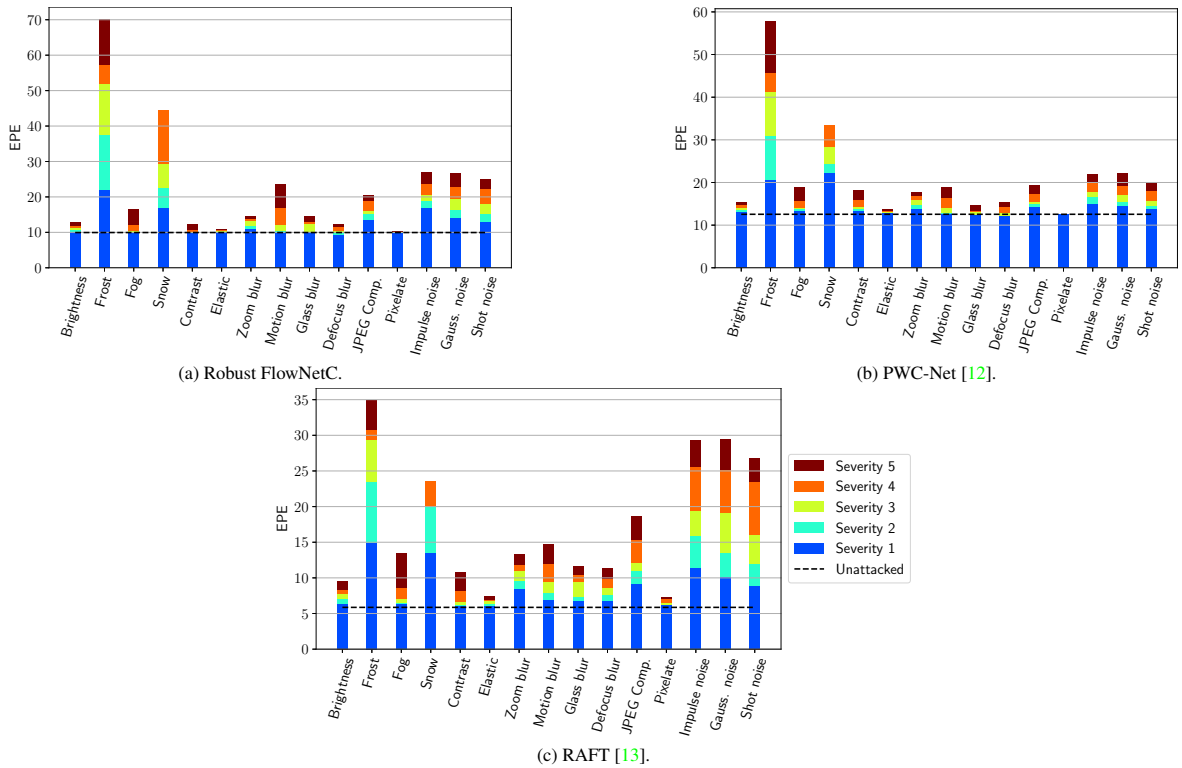
Figure 14. **Adversarial universal perturbation attack examples.** Adversarial universal perturbation attacks with I-FGSM and L_∞ norm $\epsilon = 0.02$. The bottom three rows of each subfigure show the worst examples based on the absolute degradation of the flow estimate in descending order. Note that the deterioration of flow estimates is to be expected for these image pairs since there are more ambiguities due to the lower contrast. Best viewed in color and with zoom.



(a) Robust FlowNetC vs. I-FGSM. (b) PWC-Net [12] vs. I-FGSM. (c) RAFT [13] vs. I-FGSM.

(d) Robust FlowNetC vs. MI-FGSM. (e) PWC-Net [12] vs. MI-FGSM. (f) RAFT [13] vs. MI-FGSM.

Figure 15. **Adversarial data augmentation.** Fine-tuning makes the flow networks more robust w.r.t. (untargeted) adversarial attacks. We show fine-tuned versions of Robust FlowNetC, PWC-Net and RAFT using various L_∞ norms (*i.e.*, $\epsilon = \{0.002, 0.005, 0.01, 0.02\}$). For PWC-Net and RAFT, fine-tuning significantly improves robustness w.r.t. adversarial attacks, while having only a minor negative effect on flow performance for unattacked images. For Robust FlowNetC, flow performance deteriorates significantly for unattacked images. Surprisingly, all flow networks are also robust against the stronger MI-FGSM attacks, although they were not fine-tuned for them.



(a) Robust FlowNetC. (b) PWC-Net [12].

(c) RAFT [13].

Figure 16. **Common image corruptions.** X-axis: corruption types ordered from low to high frequency (based on Saikia *et al.* [11]). Y-axis: EPE for a given corruption type across various severities. All corruption types, except for Frost, Snow, Impulse noise, Gaussian noise, and Shot noise, have little negative effect on flow performance.