Attacking Optical Flow

Anurag Ranjan[†]   Joel Janai[†‡]   Andreas Geiger[†‡]   Michael J. Black[†]

[†]Max Planck Institute for Intelligent Systems
[‡]University of Tübingen

{aranjan, jjanai, ageiger, black}@tue.mpg.de

Abstract

Deep neural nets achieve state-of-the-art performance on the problem of optical flow estimation. Since optical flow is used in several safety-critical applications like self-driving cars, it is important to gain insights into the robustness of those techniques. Recently, it has been shown that adversarial attacks easily fool deep neural networks to misclassify objects. The robustness of optical flow networks to adversarial attacks, however, has not been studied so far. In this paper, we extend adversarial patch attacks to optical flow networks and show that such attacks can compromise their performance. We show that corrupting a small patch of less than 1% of the image size can significantly affect optical flow estimates. Our attacks lead to noisy flow estimates that extend significantly beyond the region of the attack, in many cases even completely erasing the motion of objects in the scene. While networks using an encoder-decoder architecture are very sensitive to these attacks, we found that networks using a spatial pyramid architecture are less affected. We analyse the success and failure of attacking both architectures by visualizing their feature maps and comparing them to classical optical flow techniques which are robust to these attacks. We also demonstrate that such attacks are practical by placing a printed pattern into real scenes.

1. Introduction

Optical flow refers to the apparent 2D motion of each pixel in an image sequence. It is denoted by a vector field (u, v) that corresponds to the displacement of each pixel in the image plane. The classical formulation [12] seeks the optical flow (u, v) between two consecutive images $I(x, y, t)$ and $I(x + u, y + v, t + 1)$ in a sequence that minimizes the brightness constancy (i.e. photometric) error at each pixel, $\rho(I(x, y, t) - I(x + u, y + v, t + 1))$, for some robust function ρ, subject to spatial coherence constraints that regularize the solution [4].
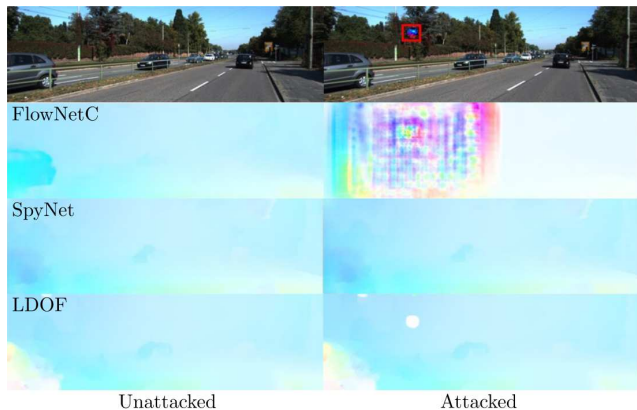
The estimation of flow from pairs of images has a long

Figure 1. **Overview**. The first column shows the optical flow results using an encoder-decoder architecture, FlowNetC [8], a spatial pyramid architecture, SpyNet [20], and a classical method, LDOF [6]. In the second column, a small circular patch is added to both frames at the same location and orientation (highlighted by the red box for illustration purposes). SpyNet and LDOF are barely affected by the patch. In contrast, FlowNetC is strongly affected, even in regions far away from the patch.

history [12] and is used in many applications spanning medicine, special effects, video analysis, action recognition, navigation, and robotics to name a few. A large number of methods [3] have approached the problem and steadily improved the results on current benchmarks [7, 10]. However, classical optical flow methods typically require extensive hand engineering [26] and computationally intensive optimization. Recent methods [8, 13, 20] have therefore approached optical flow estimation using deep neural networks. These methods typically run in real time and produce results that are competitive with, or even surpass, classical optimization-based approaches. Given the performance of recent optical flow networks, they could become an important component in applications such as self-driving cars. The safety issues surrounding many of these systems implies that the robustness of the algorithms must be well understood. To date, there has not been any study to mea-

sure the robustness of these networks to adversarial attacks.

Adversarial attacks on neural networks have been shown to work on image classification networks [18] and, consequently, it is reasonable to ask how such attacks can affect optical flow networks. Consider the optical flow (u, v) between two frames of an image pair (I_t, I_{t+1}) computed using a network F such that $(u, v) = F(I_t, I_{t+1})$. A successful adversarial attack would cause large changes in estimated optical flow for small, unnoticeable, perturbations in the images. Specifically, the adversary seeks a perturbed image pair $(\tilde{I}_t, \tilde{I}_{t+1})$ such that $\|I_t - \tilde{I}_t\|_0 + \|I_{t+1} - \tilde{I}_{t+1}\|_0 < \epsilon$, where ϵ is a small constant, and the estimated optical flow on the perturbed images is significantly worse, $\|F(I_t, I_{t+1}) - F(\tilde{I}_t, \tilde{I}_{t+1})\| > E$, with a large constant E. We are particularly interested in perturbations under the ℓ_0 norm, since the objective is to perturb a very small number of pixels in the original image.

In general, the perturbations on the image can be defined in several ways [11, 18, 25, 29]. Here, we focus on applying perturbations by pasting a small patch on the image motivated by Brown et al. [5], who use such an approach to carry out targeted adversarial attacks on image classification. The patch attack has real-world significance and we show that adversarial patches can compromise optical flow networks if an engineered patch is printed and placed in real-world scenes.

Optical flow networks can be classified into two types of architectures – the encoder-decoder [8, 13] and spatial pyramid networks [20, 28]. We contrast the robustness of adversarial attacks under these two architecture types. Similar to [5], we independently and jointly learn patches to attack all networks. The learned patches compromise the encoder-decoder architectures while spatial pyramid networks show more robust behaviour as shown in Figure 1.

In the automotive scenario, cameras for autonomous driving are often behind the windscreen. Patch attacks can potentially be accomplished by placing the patch on the windshield of the car or placing it in the scene (e.g. on a traffic sign or other vehicle). Note that when the patch has zero motion w.r.t. the camera, classical optical flow algorithms estimate zero optical flow over the patch (LDOF in Fig. 1). However, this engineered patch, even if it has no motion, can cause the optical flow predictions from encoder-decoder architectures to be wildly wrong (FlowNetC in Fig. 1). In such a scenario, the patch affects the estimated optical flow over large areas in the image that are far away from the patch. The patch in Figure 1 is 0.53% the size of the image and is barely noticeable, yet it affects the flow in about 50% of the image region.

For the encoder-decoder networks, the patches not only influence the optical flow estimates in the area of the patch, but their influence extends to remote areas in the image. In

spatial pyramid architectures, in contrast, the patch causes at most small degradations in the area of the patch. Classical approaches [6, 23] are even more robust to adversarial patches. We propose a *Zero-Flow* test (Section 5) to analyse the causes of patch attacks, where we visualize the feature maps of the networks while attacking a uniform random noise image without motion. Thereby, we identify three major problems of optical flow architectures. 1) Flow networks are not spatially invariant, leading to spatially varying feature maps even without any motion. 2) Spatial pyramid networks produce large errors at coarse resolutions but are able to recover. 3) Deconvolution layers lead to strong amplification of activations and checkerboard artifacts.

Our contributions are as follows. We extend adversarial patch attacks to optical flow neural networks. We learn adversarial patches and show that these attacks can considerably affect the performance of optical flow networks based on an encoder-decoder architecture. We show that spatial pyramid architectures, along with classical optical flow methods, are more robust to patch attacks. We show that such attacks are easy to implement in practice by simply printing the patch and placing it in the scene. We also analyse the feature maps of these networks under attack to provide insight into their behaviour under attack. Code is available at http://flowattack.is.tue.mpg.de/.

2. Related Work

Optical Flow. The classical version of the optical flow problem involves solving for a flow field that minimizes the brightness constancy loss. A survey of classical optical flow methods is available in [3, 27]. These classical methods solve a complex, non-convex, optimization problem and are often slow. Recent deep learning methods replace the optimization process, and instead directly predict optical flow using convolutional networks [8, 13, 20]. FlowNet [8] is the first work to regress optical flow by learning an end-to-end deep neural network. FlowNet is based on an encoder-decoder architecture with skip connections. Although FlowNet is much faster than classical methods [26], it is not as accurate. More recently, Ilg et al. [13] proposed FlowNet2, which achieves state-of-the-art performance on optical flow benchmarks. FlowNet2 is constructed by stacking multiple FlowNets together and fusing the output with a network specialized on small motions.

In contrast, motivated by the classical coarse-to-fine methods, SpyNet [20] splits the matching problem into simpler subproblems using coarse-to-fine estimation on an image pyramid. PWC-Net [28] extends this idea with a correlation layer to learn optical flow prediction; this gives state-of-the-art results on optical flow benchmarks.

Still other methods approach this problem from the unsupervised learning perspective [17, 31, 15, 21] by using a neural network to minimize the photometric error un-

der certain constraints. These methods use both encoder-decoder architectures [17, 31] and spatial pyramid architectures [15].

While deep networks show impressive results on benchmark datasets, their robustness is not yet well understood. Robustness, however, is critical for applications such as autonomous driving. Therefore, we investigate the robustness of representative approaches of each category in this paper. In particular, we test FlowNet [8], FlowNet2 [13], SpyNet [20], PWC-Net [28], and Back2Future [15] for their robustness against adversarial attacks.

Adversarial Attacks. Adversarial attacks seek small perturbations of the input causing large errors in the estimation by a deep neural network. Attacking neural networks using adversarial examples is a popular way to examine the reliability of networks for image classification [11, 18, 29]. The key to all such attacks is that the change to the image should be minor, yet have a large influence on the output. Adversarial examples typically involve small perturbations to the image that are not noticeable to the human eye. The adversaries are shown to work even when a single pixel is perturbed in the image [25].

Although these attacks reveal limitations of deep networks, they can not be easily replicated in real-world settings. For instance, it is rather difficult to change a scene such that one pixel captured by a camera is perturbed in a specific way to fool the network. However, recent work [16] demonstrates that adversarial examples can also work when printed out and shown to the network under different illumination conditions. Athalye et al. [2] show that adversarial examples can be 3D printed and are misclassified by networks at different scales and orientations. Sharif et al. [24] construct adversarial glasses to fool facial recognition systems. Etimov et al. [9] show that stop signs can be misclassified by placing various stickers on top of them.

Recently proposed patch attacks [5] place a small engineered patch in the scene to fool the network into making wrong decisions. Patch attacks are interesting because they can be easily replicated in the real world, and work at several scales and orientations of the patch. This makes deep networks vulnerable in real world applications. Therefore, we focus our investigation of the robustness of deep flow networks on these kinds of attacks.

To our knowledge, there has been no work on attacking optical flow networks. Consequently, we explore how such networks can be attacked using adversarial patches and analyse the potential causes of such vulnerabilities.

3. Approach

Adversarial attacks are carried out by optimizing for a perturbation that forces a network to output the wrong labels compared with ground truth labels. For example, if (u, v) represents the ground truth labels for inputs (I_t, I_{t+1}), a perturbed input $(\tilde{I}_t, \tilde{I}_{t+1})$ would produce incorrect labels $(\tilde{u}, \tilde{v})$. However, there are no optical flow datasets that have dense optical flow ground truth labels for natural images. Most of the optical flow datasets are synthetic [7, 8, 22]. SlowFlow [14] provides real world data but is limited in size. The KITTI dataset [10] has sparse ground truth labels and the annotations are limited to 200 training examples. To address the problem of limited ground truth labels, we use the predictions of the optical flow network as pseudo ground truth. We optimize for the perturbation that maximizes the angle between the predicted flow vectors obtained using the original images and the perturbed images respectively. Using the predictions instead of ground truth has the advantage that a patch can be optimized using any unlabelled video. This makes it easier to attack optical flow systems even in the absense of the ground truth.

Consider an optical flow network F that computes the optical flow between two-frames of an image sequence (I_t, I_{t+1}) of resolution $H \times W$. Consider a small patch p of resolution $h \times w$ that is pasted onto the image to perturb it. Let $\delta \in \mathcal{T}$ be a transformation that can be applied to the patch. These transformations in $\mathcal{T}$ can be a combination of rotations and scaling. We define the perturbation $A(I, p, \delta, l)$ on the image I, that applies the transformations δ to the patch p and pastes it at a location $l \in \mathcal{L}$ in the image. We apply the same perturbation to both frames in the sequence $\tilde{I}_t = A(I_t, p, \delta, l)$ and $\tilde{I}_{t+1} = A(I_{t+1}, p, \delta, l)$. This means that the optical flow between the perturbed frames is zero in the region containing the patch. In the real world, this would correspond to a patch being stationary with respect to the camera. In our experiments, we show that patches obtained with this assumption generalize well to the realistic scenario of moving patches where $\tilde{I}_t = A(I_t, p, \delta_t, l_t)$ and $\tilde{I}_{t+1} = A(I_{t+1}, p, \delta_{t+1}, l_{t+1})$ with $\delta_t \neq \delta_{t+1}$ and $l_t \neq l_{t+1}$.

Our goal is to learn a patch p that acts as an adversary to an optical flow network, F, and is invariant to location l or transformations δ of the patch. The resulting patch can be optimized using

$$\hat{p} = \underset{p}{\operatorname{argmin}} \mathbb{E}_{(I_t, I_{t+1}) \sim \mathcal{I}, l \sim \mathcal{L}, \delta \sim \mathcal{T}} \frac{(u, v) \cdot (\tilde{u}, \tilde{v})}{\|(u, v)\| \cdot \|(\tilde{u}, \tilde{v})\|} \quad (1)$$

with

$$(u, v) = F(I_t, I_{t+1}) \quad (2)$$
$$(\tilde{u}, \tilde{v}) = F(A(I_t, p, \delta, l), A(I_{t+1}, p, \delta, l)) \quad (3)$$

where l is sampled over all the locations $\mathcal{L}$ in the image, and $\mathcal{I}$ is a set of 2 frame sequences from a video.

Equation (1) computes the cosine of the angle between the optical flow estimated by the network for normal and perturbed images. Therefore, minimizing the loss is equivalent to finding those adversarial examples that reverse the direction of optical flow estimated by the network.
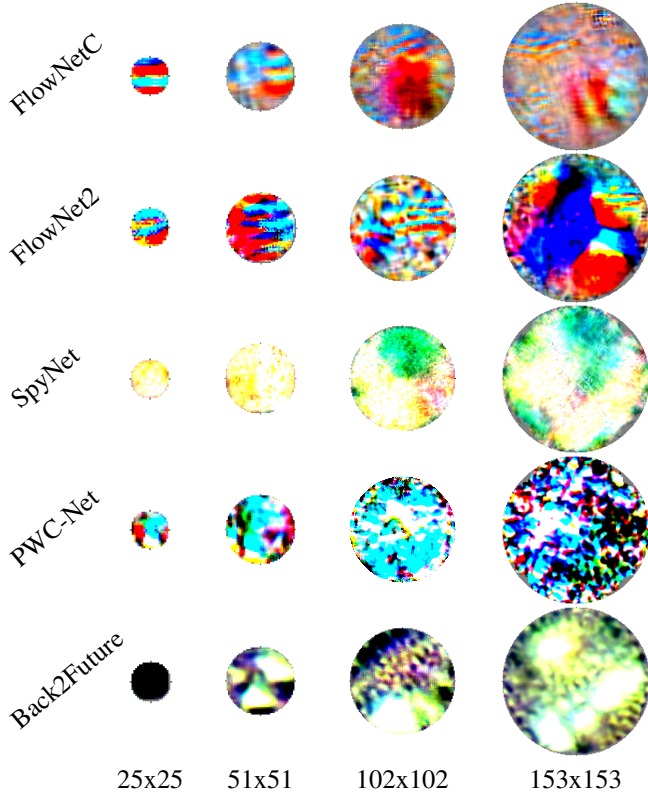
Figure 2. **Adversarial Patches**. Obtained for different optical flow networks. The size is enlarged for visualization purposes.

Table 1. **Optical Flow Methods.** Contrasting different optical flow methods with respect to the Type, Supervision (Super.), Network Architecture (Net.) - Encoder-Decoder (ED) vs. Spatial Pyramid (SP) - and Number of Parameters (#Params.). Number of parameters is denoted in Millions (M).

	Type	Super.	Net.	# Params
FlowNetC [8]	CNN	Yes	ED	39 M
FlowNet2 [13]	CNN	Yes	ED	162 M
SpyNet [20]	CNN	Yes	SP	1.2 M
PWC-Net [28]	CNN	Yes	SP	8.75 M
Back2Future [15]	CNN	No	SP	12.2 M
LDOF [6]	Classic	-	-	-
Epic Flow [23]	Classic	-	-	-

4. Experiments

We evaluate the robustness of five optical networks to adversarial patch attacks – FlowNetC [8], FlowNet2 [13], SpyNet [20], PWC-Net [28] and Back2Future [15]. We also evaluate the robustness of two very different classical methods – LDOF [6] and EpicFlow [23] which take image derivatives (LDOF) and sparse matches (EpicFlow) to compute optical flow. These methods cover deep networks vs. classical methods, supervised vs. unsupervised learning methods and encoder-decoder vs. spatial pyramid architectures, thus providing an overall picture of optical flow methods under adversarial patch attacks. We contrast these optical flow methods in Table 1. First, we evaluate the effect of *White-box* attacks, where we learn a patch of a specific size for each network by optimizing over that particular network. We then show *Black-box* attacks by creating a universal patch using two different networks and testing all the optical flow methods on this patch in a real world scenario.

4.1. White-box Attacks

We perform White-box attacks on each of the networks independently. We learn a circular patch that is invariant to its location in the image, scale and orientation. The scale augmentations of the patch are kept within $\pm 5\%$ of the original size, and rotations are varied within $\pm 10°$. For each network, we learn patches of four different sizes. The patch size is kept under 5% the size of the images being attacked. We optimize for each patch p using Eq. (1) on unlabeled frames from the raw KITTI [10]. We choose the KITTI dataset since it reflects a safety critical application and provides an annotated training set for evaluation. We use the optical flow predictions from the networks as pseudo ground truth labels to optimize for the patch that produces the highest angular error. This allows us to leverage approximately 32000 frames instead of 200 annotated frames from the training dataset. For each of the networks, we use the pre-trained model that gives the best performance *without* fine tuning on KITTI dataset. We use Pytorch [1] as our optimization framework and optimize for the patch using stochastic gradient descent. Figure 2 shows the patches we obtain by optimizing over different optical flow networks and different patch sizes.

Evaluation. To quantify the robustness of the networks to our adversarial patch attacks, we measure their performance on the training set of the KITTI 2015 optical flow benchmark using the original and perturbed images, respectively. We measure the vulnerability of the optical flow networks using average end point error (EPE) and relative degradation of EPE in the presence of the adversarial patch (Table 2). The errors are computed by placing the learned patch p at a random location for each image in the training set. All images are resized to 384×1280 and input to the networks. We evaluate optical flow at the full image resolution.

As shown in Table 2, the performance of encoder-decoder architectures – FlowNetC and FlowNet2 degrades significantly, even when using very small patches (0.1% of the image size). The increase in error is about $400 - 500\%$ under attack. For the spatial pyramid based methods – SpyNet, PWC-Net and Back2Future, the degradation is small. Back2Future, which is an unsupervised method and shares a similar architecture with PWC-Net, suffers less

Table 2. **White-box Attacks.** Effect of patch size (in pixel and percent of the image size) on different optical flow networks. We show average End Point Error (EPE) on KITTI 2015 for each network with and without an attack. For attacks, we also show relative degradation % in the EPE.

Network	Unattacked EPE	25x25 (0.1%) EPE	Rel	51x51 (0.5%) EPE	Rel	102x102 (2.1%) EPE	Rel	153x153 (4.8%) EPE	Rel
FlowNetC [8]	14.56	29.07	+100%	40.27	+177%	82.41	+466%	95.32	+555%
FlowNet2 [13]	11.90	17.04	+43%	24.42	+105%	38.57	+224%	59.58	+400%
SpyNet [20]	20.26	20.59	+2%	21.00	+4%	21.22	+5%	21.00	+4%
PWC-Net [28]	11.03	11.37	+3%	11.50	+4%	11.86	+7%	12.52	+13%
Back2Future [15]	17.49	18.04	+3%	18.24	+4%	18.73	+7%	18.43	+5%

Table 3. **Black-box Attacks.** Attacks on different optical flow methods using a universal patch. Methods below the line were not used for training the patch.

	Unattacked EPE	Attacked EPE	Rel
FlowNet2 [13]	11.90	36.13	+203 %
PWC-Net [28]	11.03	11.01	+0 %
FlowNetC [8]	14.56	86.12	+492 %
SpyNet [20]	20.26	20.39	+1 %
Back2Future [15]	17.49	17.44	+0 %
Epic Flow [23]	4.52	4.66	+3 %
LDOF [6]	9.20	9.17	+0 %

degradation than PWC-Net on the largest patch size.

In Figure 3, we illustrate that encoder-decoder architectures are extremely vulnerable to patch attacks irrespective of scene content, location and orientation of the patch. The patch size in Figure 3 is 25×25 pixels, about 0.12% of the size of the image being attacked. The figure shows that the attacks extend significantly beyond the region of the patch, causing a significant proportion of the optical flow vectors to be degraded. We also observe that SpyNet is least affected by the attack. For PWC-Net and Back2Future, the degradation is contained within the region of the patch. The strength of the attack increases with the size of the patch. Figure 4 shows attacks with a larger patch size of 153×153, which is about 4.76% of the image size. Other examples and patch sizes are shown in the supplementary material.

4.2. Black-box Attacks

In a real world scenario, the network used by a system like an autonomous car will most likely not be accessible to optimize the adversarial patch. Therefore, we also consider a Black-box attack, which learns a "universal" patch to attack a group of networks. Since there are two architecture types (Encoder-Decoder and Spatial Pyramid), we consider one network from each type to learn a "universal" patch to attack all networks. We pick the networks – FlowNet2 [13] and PWC-Net [28] and jointly optimize for a patch that attacks both networks using Eq. 1. In this way, we obtain a patch that is capable of attacking both encoder-decoder and spatial pyramid architectures. The resulting patch can be seen in Figure 7 and the Sup. Mat. While we focus on two networks for a proof of concept, the "universal" patch can easily be jointly optimized using more networks to obtain even more effective attacks.

The training of the universal patch is identical to the White-box case. However, we decided to make the evaluation of the Black-box attacks more realistic by using the camera motion and disparity ground truth provided by the KITTI Raw dataset to compute realistic motion of patches assuming they are part of the static scene. We project the random location of the patch from the first frame into the 3D scene and re-project it into the second frame considering the camera motion and disparity. We randomly pick a disparity between the maximal disparity of the whole scene and minimal disparity in region of the patch to obtain a location in the free space between the camera and other objects in the scene. Given the re-projection of points on the patch, we then estimate a homography to transform the patch and warp it using bilinear interpolation. For sequences without ground truth camera poses we use zero motion like in the White-box attacks. In this way, we simulate the situation where a patch is attached to some static object in the scene. In the supplementary material, we also show the effect of a universal patch that has zero motion w.r.t the camera as in the White-box attacks.

Evaluation. Table 3 shows the performance of all networks and classical optical flow methods from Table 1 in the presence of the Black-box adversary. Consistent with White-box attacks, we observe that encoder-decoder architectures suffer significantly in the presence of the adversarial patch. Both, spatial pyramid architectures and classical methods, are more robust to patch attacks.

We observe that the attacks on encoder-decoder networks are most effective when the patch is placed in the the center of the image. The patches influence image regions according to the receptive field of the convolutional networks. Although the patches are small ($\sim 0.5\%$) compared to the image size, they can affect up to 50% of the image area as can be observed in Figure 5. We show more qualitative results in the supplementary material.
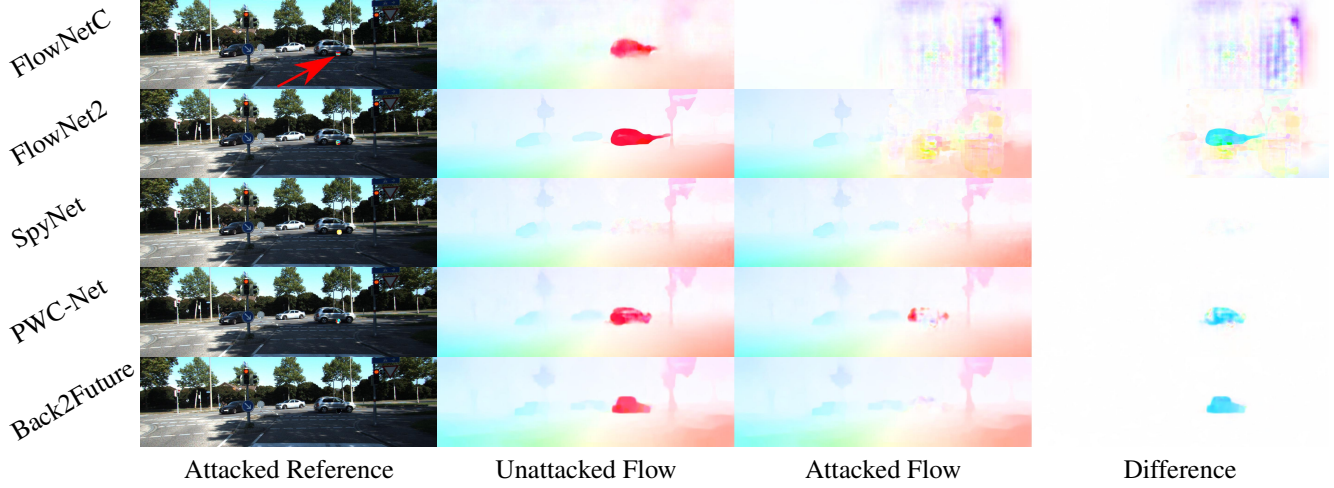
Figure 3. **White-box Attacks.** 25x25 patch attacks on all networks.
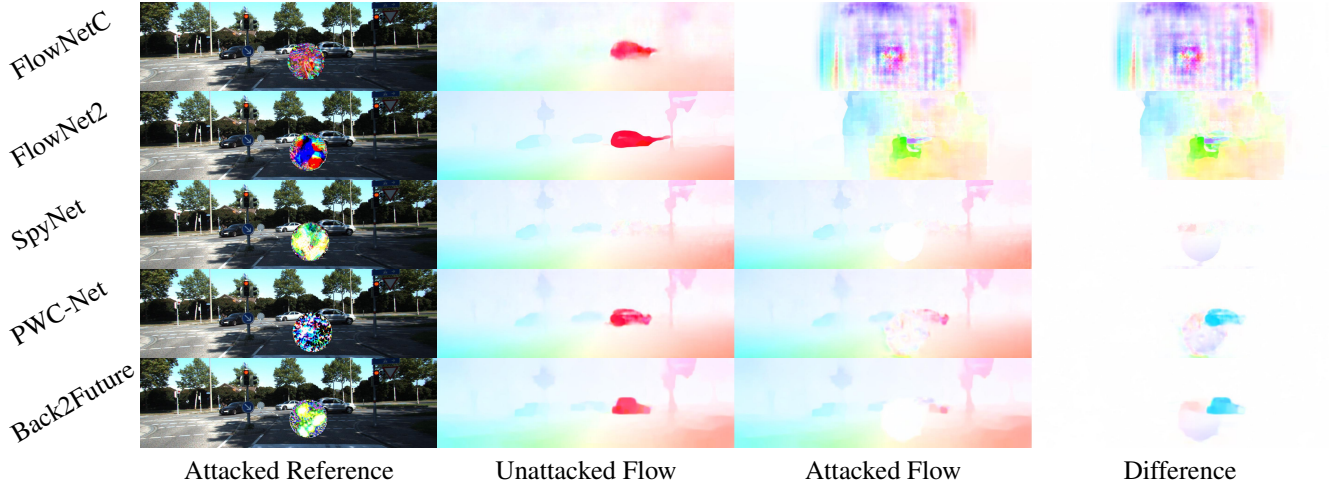

Figure 4. **White-box Attacks.** 153x153 patch attacks on all networks.

4.3. Real World Attacks

In order to evaluate the effectiveness of adversarial patches in the real world, we print the Black-box patch at 1200 dpi, scaled to 8 times the original size in the image. We then place the patch in the scene, record a video sequence, and observe the effects on two networks - FlowNetC and FlowNet2. Figure 7 shows the effect of the printed patch on a deployed optical flow system running FlowNetC. While the location of the patch in the scene was constant, we experiment with and without camera motion. We observe that the real world attacks on FlowNetC work equally well with and without camera motion, while attacks on FlowNet2 worked better with camera motion. For the video demo of this experiment, please follow this link – http://flowattack.is.tue.mpg.de/.

5. Zero-Flow Test

To better understand the behaviour of neural networks in the presence of an adversarial patch attack, we propose the *Zero-Flow* test. We generate an image I by independently sampling uniform random noise for each of the pixels. We then paste the universal patch p on I, resulting in a perturbed image $\tilde{I}$. We then replicate the image $\tilde{I}$ and input the two frames to the network. Since the image is replicated for both frames, the optical flow is zero over all the pixels.

Assuming that the attack was ineffective, $F(\tilde{I}, \tilde{I}) = 0$ must hold. Furthermore, if perturbation is not applied to the network inputs, $F(I, I) = 0$ must hold. For feature maps of the network, $F_k(\tilde{I}, \tilde{I}) = F_k(I, I)$ must hold. Here, F_k is the output from the k-th layer of the neural network. In Figure 6, we show feature map visualizations along with their average norm from two networks, FlowNetC and PWC-Net. For each network we show the feature maps with and without
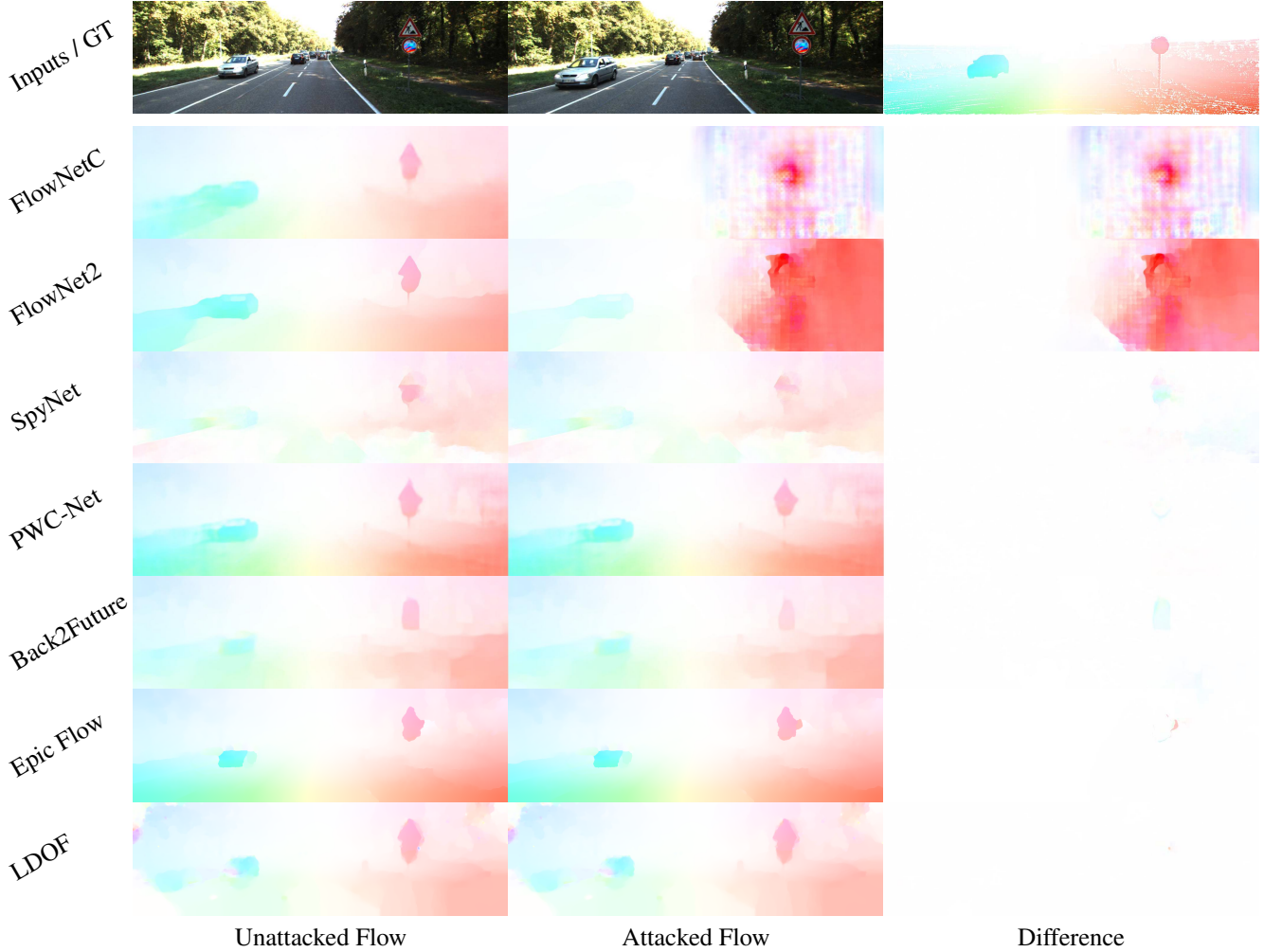
Figure 5. **Black-box Attacks.** Universal patch trained on FlowNet2 and PWC-Net used on all approaches pasted on a traffic sign with realistic motion as described in Section 4.2.

the patch. Each feature map is normalized independently for contrast and scaled to the same size.

Figure 6 illustrates that the encoder of FlowNetC (`conv<3_1,4,5,6>`) spatially propagates the patch information, as expected, and the feature activation encompasses the entire image towards the end of the encoder (`conv6`). While deconvolutions (`deconv<4,5,6>`) reduce the spatial propagation of activations, they also lead to a strong amplification. Finally, the flow prediction layers (`flow<5,4,3,2>`) combine both problems by fusing the encoder and decoder responses. This leads to large and widespread degradation of optical flow predictions for the Zero-Flow test. Without the adversarial patch, we observe well distributed activations in the first convolutions as well as low activations for all layers in general.

PWC-Net shows low activations in all correlation layers (`corr<6,5,4,3,2>`) irrespective of the adversarial patch. Interestingly, the flow decoder of lower levels

(`flow<6,5>`) predict large flow in both cases – with and without patch. However, the flow prediction at finer levels of the pyramid (`flow<3,2>`) seems to compensate for these errors. Similar to FlowNetC, the deconvolution layers (`upfeat<5,4,3>`) lead to an amplification of activations, with and without the patch. In addition, they seem to cause checkerboard artifacts that have also been observed in other contexts [19, 30].

The Zero-Flow test reveals several problems with the optical flow networks. First, the networks (FlowNetC, PWC-Net) are not spatially invariant. In the case of zero motion, the activation maps of the network vary spatially across the whole image even without the adversarial patch. Second, the pyramid networks predict large motion in coarser levels `flow<5,4>` irrespective of the presence or absence of the patch. Finally, the deconvolution layers lead to an amplification of the responses and checkerboard artifacts. The attacks exploit these weaknesses and amplify them, degrad-
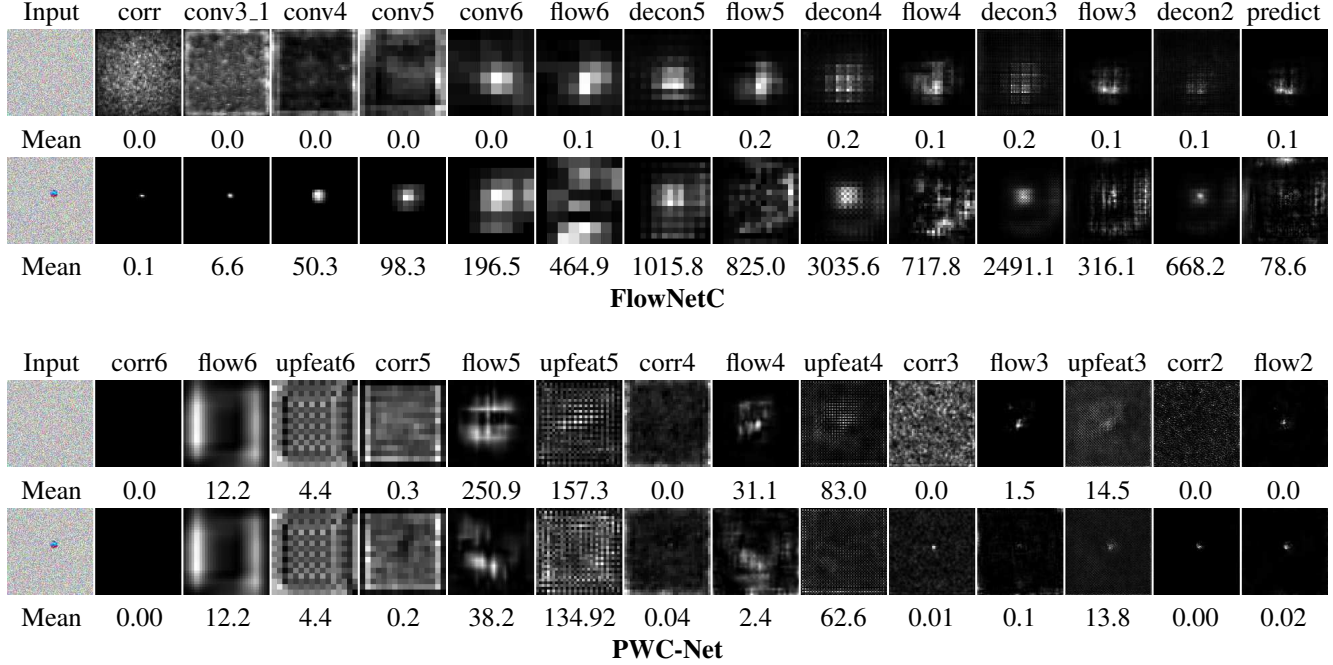
Figure 6. **Visualization of Feature Maps.** Average norm of feature maps as we move deeper through the network. Each image is normalized independently for contrast and scaled to the same size. Mean refers to average norm of the feature maps.

Figure 7. **Real World Attacks on FlowNetC.** Top: The presence of a printed adversarial patch visible in the centre of the image significantly degrades the optical flow predictions. Bottom: As the patch is covered by the subject, the effect of the patch vanishes.

ing predictions. In the Supplementary Material, we show additional visualizations of feature maps (from FlowNet2 and Back2Future).

6. Discussion

We have shown that patch attacks generalize to state-of-the-art optical flow networks and can considerably impact the performance of optical flow systems that use deep neural networks. The patch attacks are invariant to translation and small changes in scale and orientation of the patch. Learning patches for larger changes in scale and orientation is computationally costly. Nevertheless, it is possible to replicate the attacks with real printed patterns.

Small adversarial patches cause large errors even in remote regions with encoder-decoder networks (FlowNet, FlowNet2) while spatial pyramid networks (SpyNet, PWC-Net, Back2Future) are not strongly affected. Classical approaches are also robust to these attacks. This indicates that incorporating classical ideas such as image pyramids and correlations into architectures makes models more robust.

In our analysis, we observed the influence of such patches on feature maps in encoder-decoder as well as spatial pyramid networks. We also note that this results from inherent spatially variant properties of the optical flow networks. This indicates that convolutions alone are not enough to enforce spatial invariance in these networks.

Finally, independent of the attacks, the Zero-Flow test provides a novel approach to identify problems in deep flow networks. It reveals the internal network behaviour which may be useful to improve the robustness of flow networks.

References

[1] https://pytorch.org. 4

[2] Anish Athalye and Ilya Sutskever. Synthesizing robust adversarial examples. *arXiv.org*, 2017. 3

[3] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision (IJCV)*, 92(1):1–31, 2011. 1, 2

[4] Michael J Black and P Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding (CVIU)*, 63(1):75–104, 1996. 1

[5] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv.org*, 2017. 2, 3

[6] T. Brox and J. Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2011. 1, 2, 4, 5

[7] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *Proc. of the European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, Oct. 2012. 1, 3

[8] Alexey Dosovitskiy, Philipp Fischery, Eddy Ilg, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, Thomas Brox, et al. Flownet: Learning optical flow with convolutional networks. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, pages 2758–2766. IEEE, 2015. 1, 2, 3, 4, 5

[9] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on deep learning models. *arXiv.org*, 1, 2017. 3

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1, 3, 4

[11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv.org*, 2014. 2, 3

[12] Berthold K Horn and Brian G Schunck. Determining optical flow. *Artificial Intelligence (AI)*, 17(1–3):185–203, 1981. 1

[13] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017. 1, 2, 3, 4, 5

[14] Joel Janai, Fatma Guney, Jonas Wulff, Michael J Black, and Andreas Geiger. Slow flow: Exploiting high-speed cameras for accurate and diverse optical flow reference data. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3597–3607, 2017. 3

[15] Joel Janai, Fatma Güney, Anurag Ranjan, Michael Black, and Andreas Geiger. Unsupervised learning of multi-frame optical flow with occlusions. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2018. 2, 3, 4, 5

[16] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv.org*, 2016. 3

[17] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2, 3

[18] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436, 2015. 2, 3

[19] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016. 7

[20] Anurag Ranjan and Michael J. Black. Optical flow estimation using a spatial pyramid network. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 3, 4, 5

[21] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12240–12249, 2019. 2

[22] Anurag Ranjan, Javier Romero, and Michael J Black. Learning human optical flow. *Proc. of the British Machine Vision Conf. (BMVC)*, 2018. 3

[23] Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. EpicFlow: Edge-preserving interpolation of correspondences for optical flow. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 4, 5

[24] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proc. of the ACM SIGSAC Conference on Computer and Communications Security*, pages 1528–1540. ACM, 2016. 3

[25] Jiawei Su, Danilo Vasconcellos Vargas, and Sakurai Kouichi. One pixel attack for fooling deep neural networks. *arXiv.org*, 2017. 2, 3

[26] Deqing Sun, Stefan Roth, and Michael J Black. Secrets of optical flow estimation and their principles. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2439. IEEE, 2010. 1, 2

[27] Deqing Sun, Stefan Roth, and Michael J Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *International Journal of Computer Vision (IJCV)*, 106(2):115–137, 2014. 2

[28] 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 Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8934–8943, 2018. 2, 3, 4, 5

[29] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv.org*, 2013. 2, 3

[30] Zbigniew Wojna, Jasper R. R. Uijlings, Sergio Guadarrama, Nathan Silberman, Liang-Chieh Chen, Alireza Fathi, and

Vittorio Ferrari. The devil is in the decoder: Classification, regression and gans. In *Proc. of the British Machine Vision Conf. (BMVC)*, 2017. 7

[31] Jason J Yu, Adam W Harley, and Konstantinos G Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. *arXiv.org*, 2016. 2, 3