Attacking Optical Flow

Andreas Geiger

Autonomous Vision Group
University of Tübingen / MPI for Intelligent Systems Tübingen

August 23, 2020

University of Tübingen
MPI for Intelligent Systems
Autonomous Vision Group

Collaborators

Anurag Ranjan

Joel Janai

Aditya Prakash

Eshed Ohn-Bar

Kashyap Chitta

Aseem Behl

Michael Black

Andreas Geiger

Self-Driving must be Robust

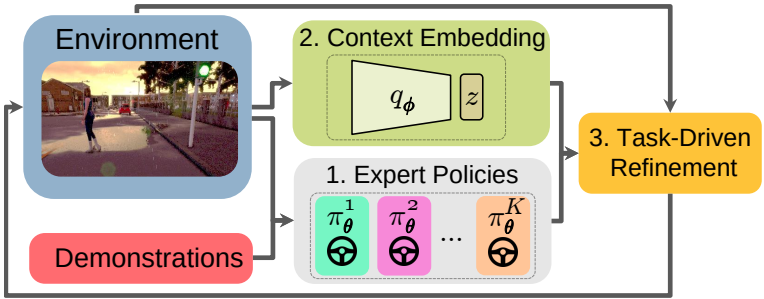
Self-Driving must be Robust

Self-Driving must be Robust

Robustness to environment

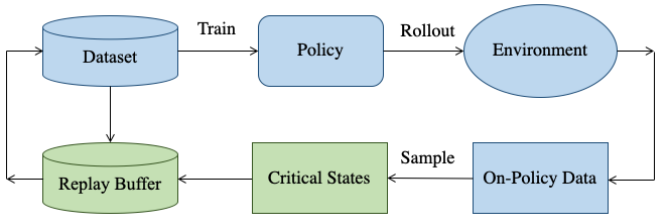
Robustness to perturbations

Situational Driving



- Learn **mixture of imitation experts** $\{\pi_\theta^k\}$
- **Task-driven** policy refinement

Data Aggregation



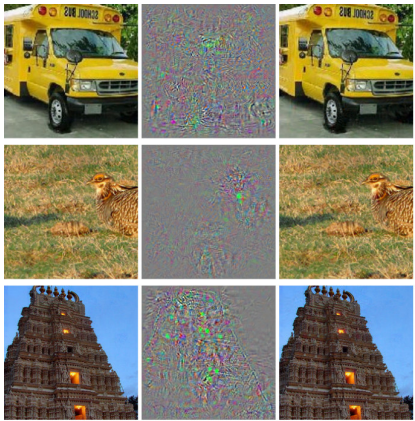
- **Data aggregation:** query expert for difficult situations

Self-Driving must be Robust

Robustness to environment

Robustness to perturbations

Adversarial Attacks on Image Classification



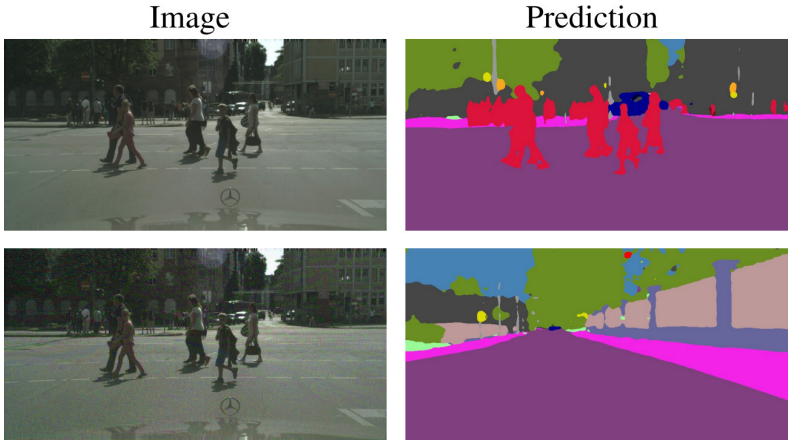
L-BFGS Attack:

- Given classifier $f : \mathbb{R}^m \to \{1, \ldots, L\}$
- Find adversarial example for image x:

$$x + \underset{\Delta x}{\operatorname{argmin}} \{\|\Delta x\|_2 : f(x + \Delta x) \neq f(x)\}$$

- All images classified as “ostrich” (right column)

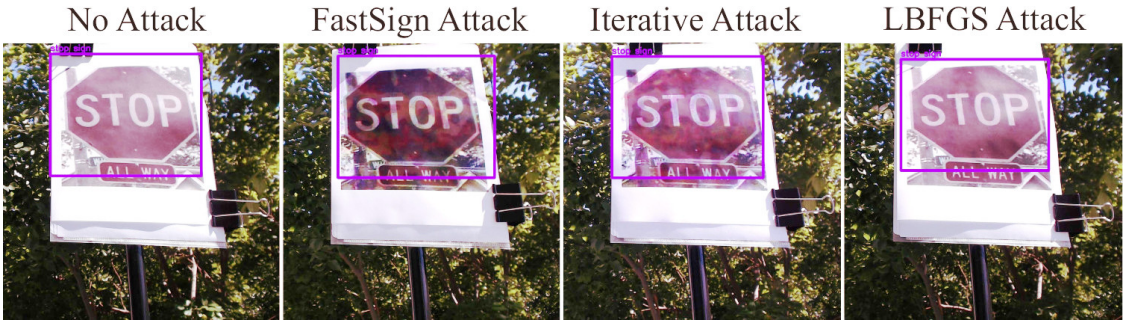
Adversarial Attacks on Semantic Segmentation



- Attack on **semantic segmentation** manipulates label map

Metzen, Kumar, Brox and Fischer: Universal Adversarial Perturbations Against Semantic Image Segmentation. ICCV, 2017.

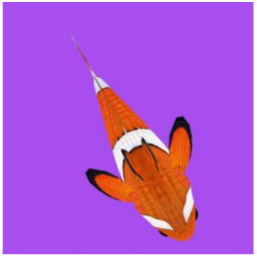
Physical Adversarial Attacks



"Adversarial perturbation methods applied to **stop sign detection** only work in carefully chosen situations, and our preliminary experiment shows that we might **not need to worry** about it in many real circumstances, specifically with autonomous vehicles."

Lu, Sibai, Fabry and Forsyth: NO Need to Worry about Adversarial Examples in Object Detection in Autonomous Vehicles. Arxiv, 2017.

Robust Adversarial Attacks

- Demonstrate existence of **robust adversarial examples** in the physical world
- Maximize **expectation over transformation** $\mathcal{T}$ (EOT):

$$\underset{x'}{\operatorname{argmax}} \; \mathbb{E}_{t\sim\mathcal{T}} \left[\log P(y|t(x')) - \lambda \|(t(x') - t(x)\|_2\right]$$

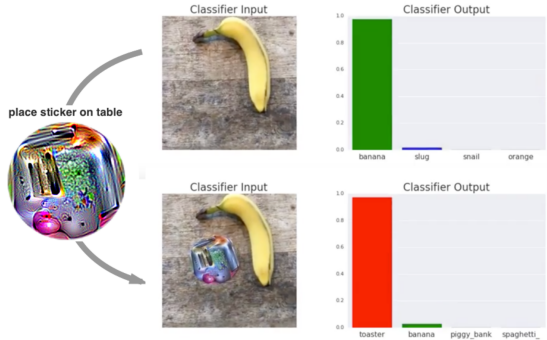
- Larger distributions require larger perturbations

Athalye, Engstrom, Ilyas and Kwok: Synthesizing Robust Adversarial Examples. ICML, 2018.

Robust Adversarial Attacks

- **Robust adversarial example** designed to mimic “graffiti”

Adversarial Patch Attacks



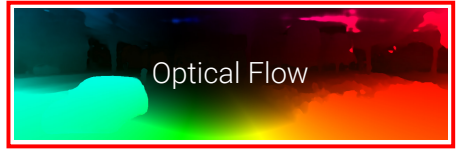
- **Patch attacks** use EOT idea, but also optimize across **many images**
- Easy to apply in real-world settings (attaching patch to an object)

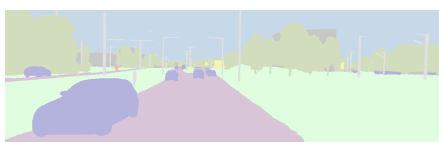
Brown, Mane, Roy, Abadi and Gilmer. Adversarial patch. Arxiv, 2017.

Optical Flow

Low-Level Perception

Low-Level Perception

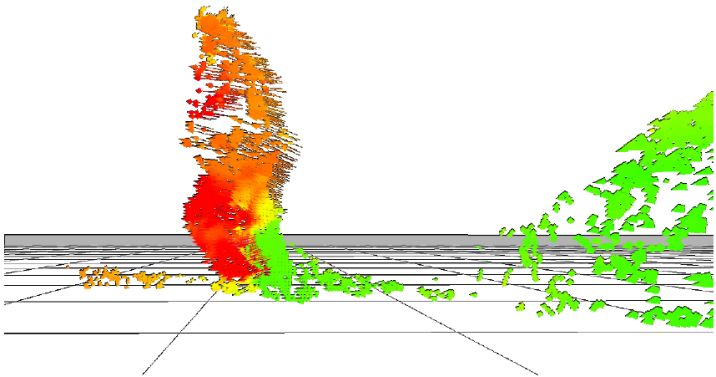




Motion Estimation

left image at time t

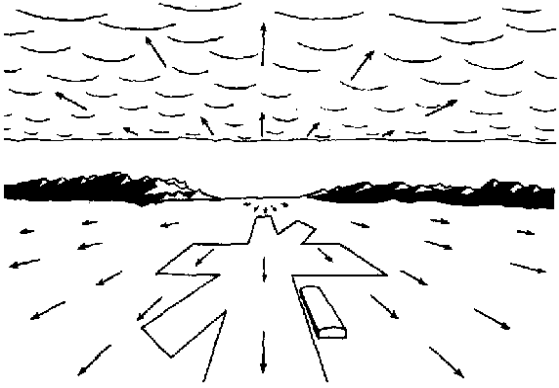
left image at time $t + 1$



- Accurate **motion estimation** is critical for self-driving
- Allows for making **predictions** about the future

Wedel, Brox, Vaudrey, Rabe, Franke and Cremers: Stereoscopic scene flow computation for 3D motion understanding. IJCV, 2011.

Optical Flow

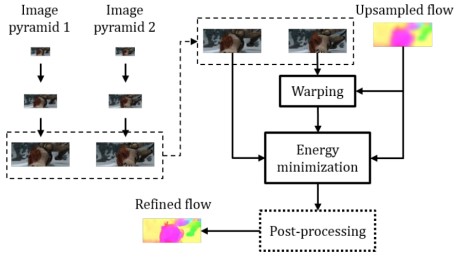


- Optical flow describes the **2D pixel motion** between two frames
- Optical flow contains information about **3D geometry** and **3D motion**

Gibson: The Perception of the Visual World. 1950.

Approaches to Optical Flow

Variational Optical Flow

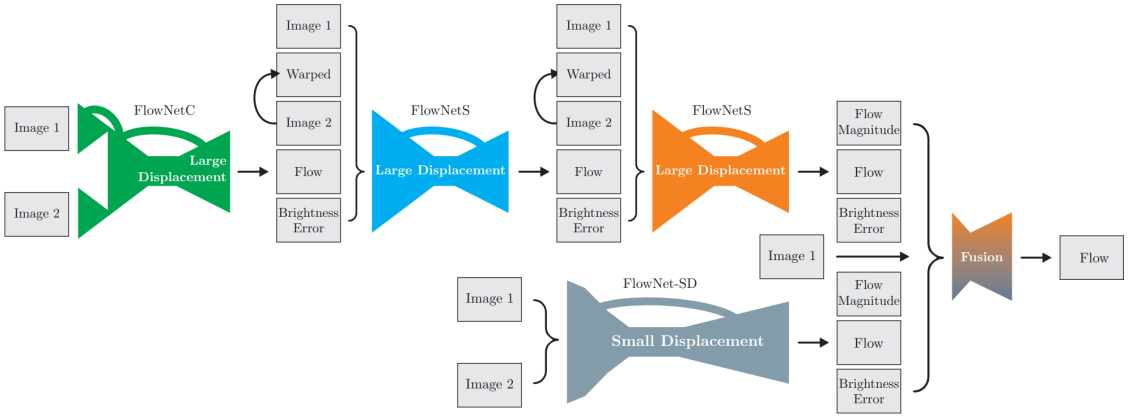


Variational Energy:

$$E(u,v) = \iint \underbrace{\left(I(x+u(x,y),y+v(x,y),t+1) - I(x,y,t)\right)^2}_{\text{penalize brightness change}} + \lambda \cdot \underbrace{\left(\|\nabla u(x,y)\|^2 + \|\nabla v(x,y)\|^2\right)}_{\text{penalize flow change}} dx\,dy$$

Horn and Schunck: Determining optical flow. Artificial Intelligence, 1981.

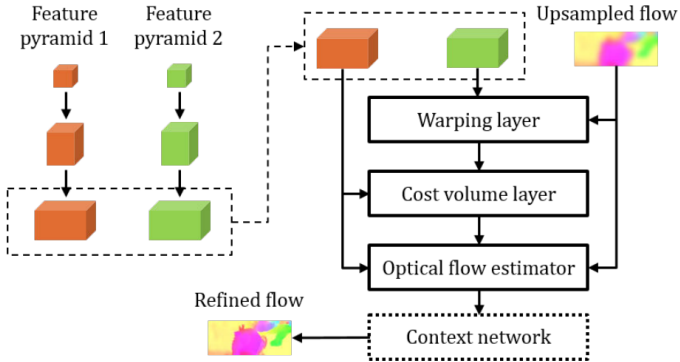
Encoder-Decoder Networks



FlowNet 2.0:

- Multiple stacked encoder-decoder networks

Spatial Pyramid Networks



PWC-Net:

- Coarse-to-fine optical flow estimation

Sun, Yang, Liu and Kautz: PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. CVPR, 2018.

Motivation

Optical flow on KITTI dataset computed via FlowNet2 [Ilg et al., CVPR 17]

- Until 2016: Classic **variational optical flow** methods state-of-the-art
- Since 2016: Leaderboards dominated by **deep learning** based methods
- But so far no investigation of **adversarial robustness** of optical flow approaches

Attacking Optical Flow

Attacking Optical Flow

Given an image pair from a video sequence ...

Attacking Optical Flow

Given an image pair from a video sequence ...

Attacking Optical Flow

... FlowNet2 predicts a smooth optical flow field.

Attacking Optical Flow

Can we obtain a small ($< 1\%$) attack patch $\hat{p}$ ...

Attacking Optical Flow

... that successfully attacks the optical flow network?

Attacking Optical Flow

Let $F(I, I')$ denote an optical flow network and $\mathcal{I}$ a dataset of frame pairs.
Let $A(I, p, t, l)$ denote image I with patch p transformed by t inserted at location l.
Let $\mathcal{T}$ denote a distribution over affine 2D transformations.
Let $\mathcal{L}$ denote a (uniform) distribution over the image domain.

Goal: Find a patch $\hat{p}$ such that

$$\hat{p} = \underset{p}{\operatorname{argmin}} \, \mathbb{E}_{(I,I') \sim \mathcal{I}, t \sim \mathcal{T}, l \sim \mathcal{L}} \left[\frac{(u, v) \cdot (\tilde{u}, \tilde{v})}{\|(u, v)\|_2 \cdot \|(\tilde{u}, \tilde{v})\|_2} \right]$$

with $\quad (u, v) = F(I, I')$

$$(\tilde{u}, \tilde{v}) = F(A(I, p, t, l), A(I', p, t, l))$$

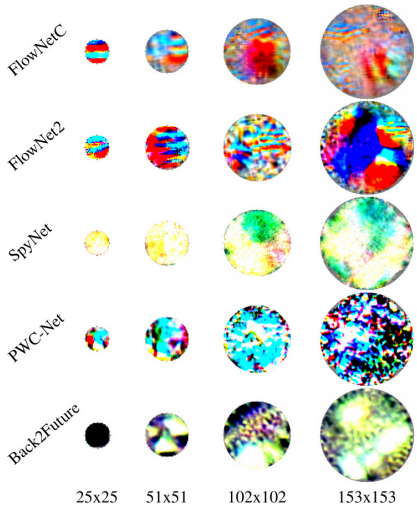
Intuition: Find patch $\hat{p}$ that reverses the direction of the optical flow.

Results

White-Box Attacks

Setting:

- Attack each network in separation
- Location sampled uniformly within image
- Scale: $\pm 5\%$
- Rotation: $\pm 10\%$
- Optimize on 32k unlabeled KITTI frames
- Flow predictions as pseudo ground truth
- Learn patches of 4 different sizes



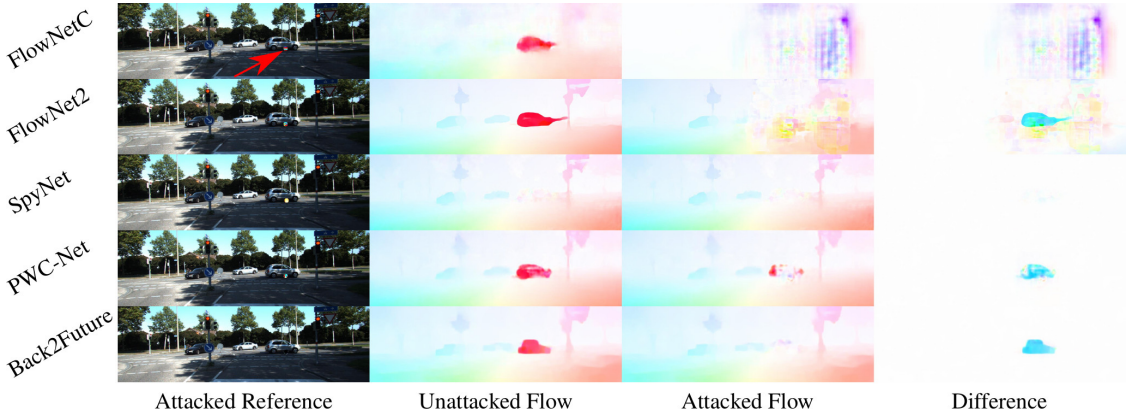
Ranjan, Janai, Geiger and Black: Attacking Optical Flow. ICCV, 2019.

White-Box Attacks

	Unattacked	25x25		153x153	
Network	**EPE**	**EPE**	**Rel**	**EPE**	**Rel**
FlowNetC [Fischer et al.]	14.56	29.07	+100 %	95.32	+555 %
FlowNet2 [Ilg et al.]	11.90	17.04	+43 %	59.58	+400 %
SpyNet [Ranjan and Black]	20.26	20.59	+2 %	21.00	+4 %
PWCNet [Sun et al.]	11.03	11.37	+3 %	12.52	+13 %
Back2Future [Janai et al.]	17.49	18.04	+3 %	18.43	+5 %

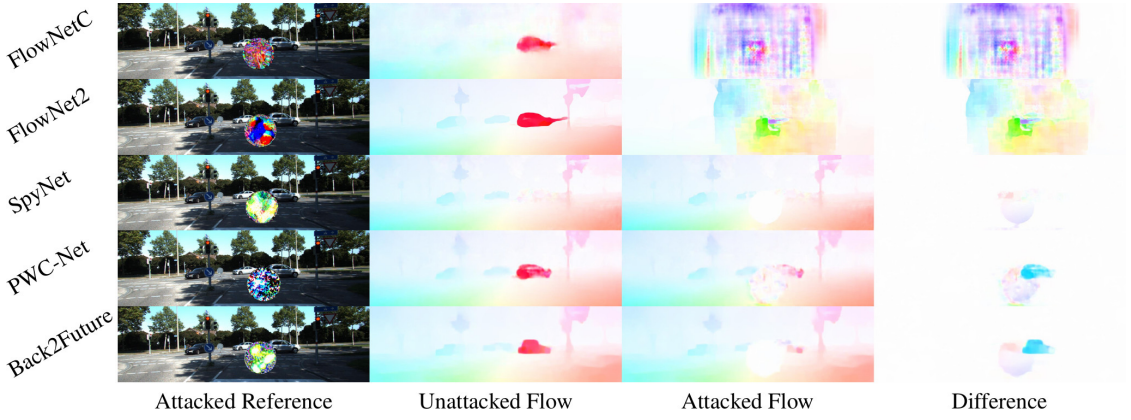
- FlowNetC and FlowNet2 use encoder-decoder architectures
- SpyNet, PWCNet and Back2Future use a spatial pyramid

White-Box Attacks



- Attacks extend beyond region of patch for encoder-decoder architectures

White-Box Attacks

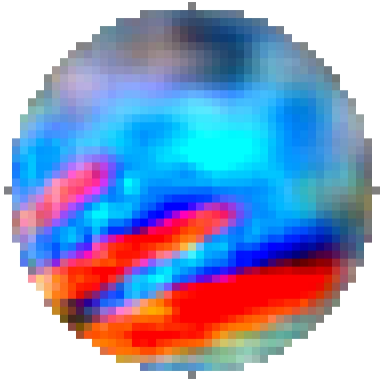


- Attacks extend beyond region of patch for encoder-decoder architectures

Black-Box Attacks

Setting:

- Patch is optimized over several networks (we use FlowNet2 and PWCNet)
- Patch is used to attack all networks
- Patch is moved as if it was part of the scene



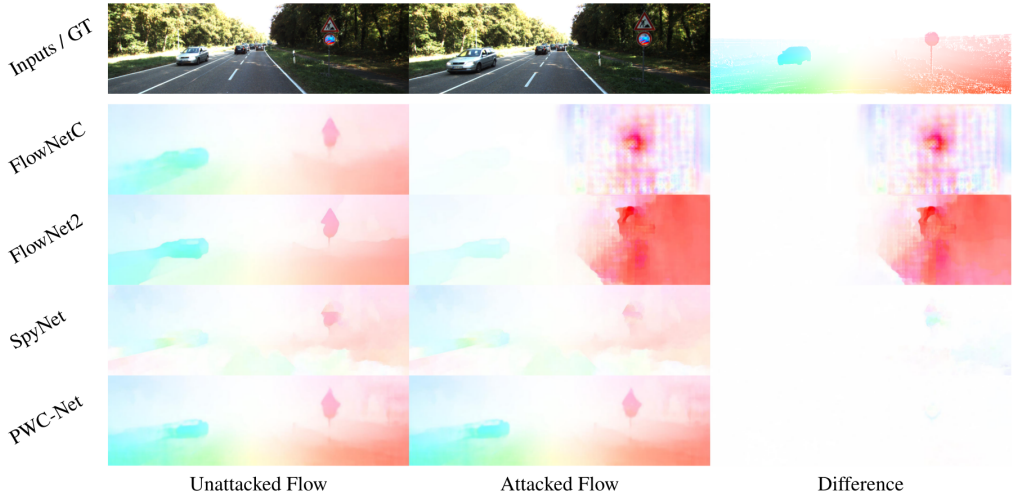
Ranjan, Janai, Geiger and Black: Attacking Optical Flow. ICCV, 2019.

Black-Box Attacks

	Unattacked EPE	Attacked EPE	Attacked Rel
FlowNet2 [Fischer et al.]	11.90	30.99	+160 %
PWCNet [Ilg et al.]	11.03	11.16	+1 %
FlowNetC [Ranjan and Black]	14.56	77.78	+434 %
SpyNet [Sun et al.]	20.26	20.65	+2 %
Back2Future [Janai et al.]	17.49	17.76	+2 %
Epic Flow [Revaud et al.]	4.52	4.57	+1 %
LDOF [Brox and Malik]	9.20	9.30	+1 %

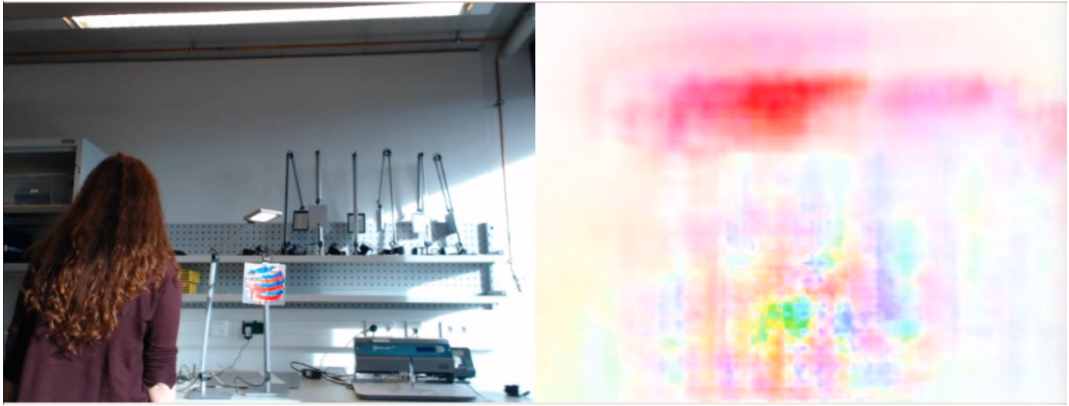
- FlowNetC and FlowNet2 use encoder-decoder architectures
- SpyNet, PWCNet and Back2Future use a spatial pyramid
- Epic Flow and LDOF are classical variational approaches

Black-Box Attacks



Ranjan, Janai, Geiger and Black: Attacking Optical Flow. ICCV, 2019.

Real-World Attack



- Printed patch attached to desk lamp

Insights into Attacks

Zero-Flow Test

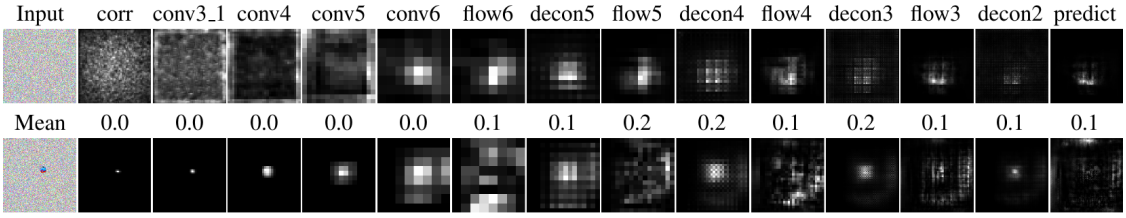
What happens when CNN sees

- Identical images?
- Identical images with identical attack patch?

Ideally

- The network *should* output a zero flow field
- The feature maps of the attacked and unattacked images should be similar

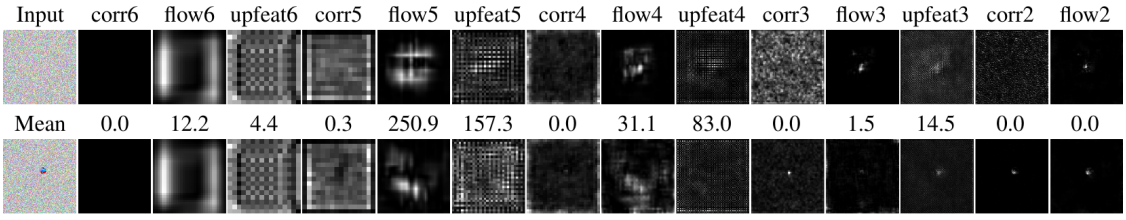
Zero-Flow Test



- Feature activations are not spatially invariant, even without an attack
- Deconvolution layers cause checkerboard artifacts
- Feature maps of encoder-decoder architectures are very different

Ranjan, Janai, Geiger and Black: Attacking Optical Flow. ICCV, 2019.

Zero-Flow Test



- Feature activations are not spatially invariant, even without an attack
- Deconvolution layers cause checkerboard artifacts
- The pyramid networks predict large motion in coarser flow levels

Ranjan, Janai, Geiger and Black: Attacking Optical Flow. ICCV, 2019.

Summary

Summary

- Placing a patch in the scene may lead to failure of optical flow networks
- Patch attacks are invariant to translation and small changes in scale and rotation
- Patch attacks work in the physical world (to some extent)
- Encoder-decoder architectures like FlowNetC and FlowNet2 are strongly affected
- Spatial pyramid methods like SPyNet and PWC-Net are quite robust
- Classical methods like LDOF and EpicFlow are not affected
- Zero-Flow test can provide insights into networks

Thank you!

http://autonomousvision.github.io