Consistent Semantic Attacks on Optical Flow

Tom Koren[2], Lior Talker[1], Michael Dinerstein[1], and Ran Vitek[1]

[1] Samsung Israel R&D Center, Tel Aviv, Israel
{lior.talker,m.dinerstein,ran.vitek}@samsung.com
[2] tomkore@gmail.com

Abstract. We present a novel approach for semantically targeted adversarial attacks on Optical Flow. In such attacks the goal is to corrupt the flow predictions of a specific object category or instance. Usually, an attacker seeks to hide the adversarial perturbations in the input. However, a quick scan of the output reveals the attack. In contrast, our method helps to hide the attacker's intent in the output flow as well. We achieve this thanks to a regularization term that encourages off-target consistency. We perform extensive tests on leading optical flow models to demonstrate the benefits of our approach in both white-box and black-box settings. Also, we demonstrate the effectiveness of our attack on subsequent tasks that depend on the optical flow.

Keywords: Adversarial attacks · Optical flow · Semantic attacks

1 Introduction

Optical Flow (OF) is a crucial subtask of many safety-critical pipelines. It is especially important for Advanced Driver Assistance Systems (ADAS) and autonomous vehicles, where unreliable optical flow can be hazardous and life-threatening. For example, *Time-To-Collision* (TTC) methods often rely on optical flow [1–4], and their errors can have dangerous consequences.

In this paper we consider malicious manipulations aiming to lead the OF predictions astray. These manipulations are represented by perturbations, sometimes subtle, that are introduced into the input pixels. In the literature such perturbations are referred to as Adversarial Attacks (AA) [5–7].

The attacker's goal is to damage a system's performance and remain unnoticed. The defender's goal is to design a system that operates reliably despite such attacks. To achieve this goal, the defender can, for example, use some AA detection method to discard suspicious inputs. One approach to detect AAs[9, 10], is to examine the input to the attacked model. Another approach, which we consider in this paper, is to examine the *output* of the attacked model. We show that a straightforward attack on OF may be fairly easy to detect in the output. We propose an AA method, which is more difficult to detect, but has a similar or stronger effect on OF.

We use the following observation: in the context of automotive applications some objects are more important than others. Obvious examples of such impor-

Fig. 1. An example of a targeted attack on HD3-PPAC [8] optical flow model. Top: the original predicted flow and the corresponding input image. Bottom left: the corrupted flow using a non-consistent attack. Bottom right: the corrupted flow using a consistent targeted attack on the vehicle in the scene.

tant objects are pedestrians and vehicles. For TTC systems, failing to estimate the correct flow for pedestrians and vehicles can lead to fatal accidents.

Assuming a semantic or instance segmentation of the observed scene is given, an attacker can specify a target to attack. That is, instead of targeting the entire image, only a subset, defined by its semantic segmentation, can be selected. Likewise, instead of perturbing all the pixels in the image, a subset of the pixels can be chosen at which the attacker introduces malicious perturbations.

Adversarial attacks targeting only a subset of pixels may still alter the predictions of other pixels in the image. To make the attack less detectable, it is beneficial to corrupt the target prediction without affecting the rest of the predictions. We refer to attacks that leave off-target predictions unaffected (or affected as little as possible) as "consistent attacks". In this paper we present a method to create targeted consistent adversarial attacks on optical flow. The chosen targets for the attacks are the "vehicle" and the "human" categories.

Figure 1 depicts an example of such an attack. The bottom row presents two attacked flows (encoded using Middlebury [11] color-wheel): consistent and non-consistent. Both attacks achieve their goal - the flow of the corresponding target, the vehicle, is heavily damaged. However, the difference between them is readily seen. A consistently attacked flow looks reasonable, while a non-consistent attack results in a flow which is chaotically cluttered. In this paper we show that the first attack is less detectable than the second.

To achieve the effect described above, we introduce a new optimization term. We refer to it as "consistency term". While being a relatively simple addition to the optimization loss, the consistency term leads to multiple improvements in the generated adversarial attacks when compared to the baseline non-consistent settings. First, as expected, the impact on the flow predictions of non-target scene objects is significantly reduced. Second, the effectiveness of attacks on targets

increases. Third, our experiments show that in the "black-box" setting, i.e. where the attacked model is inaccessible, we observe a much better transferability [12]. Finally, we demonstrate that consistent attacks are more effective on a TTC system while being less noticeable by detection methods.

We have conducted an extensive set of evaluations using five leading optical flow methods [13, 14, 8, 15, 16]. There are three main groups of experiments where we compare attacks obtained with and without our consistency term: global, local and cross-category attacks. The target of those attacks is always the same, but the subsets of pixels that are perturbed are different. In a global attack setting, the perturbation can be distributed over all pixels. In a local attack setting, only pixels of the target object can be perturbed. Finally, in a cross-category setting, in order to corrupt the flow of some target object, we perturb pixels of some other object. We have also evaluated the effect of these attacks on a downstream TTC task. We compare consistent and non-consistent attacks in terms of the tradeoff between their effect on TTC and a AA detection score.

To summarize our contribution, we are the first to study targeted attacks on optical flow models. We introduce a new term to the optimization loss, which we name the "consistency" term, to preserve the optical flow of non-target objects. This helps to hide the attacks in the output of the optical flow. We show that the resulting consistent attacks are more effective than the non-consistent attacks. We demonstrate that these attacks are more transferable, i.e. more efficient in a black-box setting, than the non-consistent attacks. Finally, we show that under three detection methods these attacks are more effective against a downstream TTC system.

1.1 Related Work

The history of optical flow methods goes back to the early 1980s, when the foundational studies of Lucas-Kanade [17] and Horn-Schunk [18] were published. Since then, hundreds of classical computer vision techniques were proposed. A substantive survey on the non-deep optical flow methods can be found in [19]. In the era of deep learning for computer vision much attention has been paid to optical flow. Deep neural network (DNN) based OF methods such as [20, 21, 13, 22, 14, 8, 15, 16] have left the classical approaches far behind in terms of performance, which is reflected in the results of the KITTI'15 benchmark [23], where the leading non-deep optical flow method [24] is scored about 150-th place.

DNN based optical flow models can be divided into groups according to their architecture characteristics: encoder-decoder [20, 21] and spatial pyramid [25, 13, 22, 16, 8] networks. Some models ([22, 16, 8, 15]) use a coarse-to-fine technique to refine their predictions. Others [14] operate with full resolution features at every stage of the model. In addition, a model can be equipped with a recurrent refinement mechanism, which is placed on top of an optical flow model, as in [26]. Finally, the RAFT model [14], which has demonstrated the *state-of-the-art* performance on KITTI'15 [23], consists of the encoder-decoder part followed by a simple recurrent module utilizing GRU [27] blocks.

Although adversarial pertrubations are possible in many ML models, the rise of deep neural networks has opened the door for massive research effort in adversarial perturbations. Historically, many of the early adversarial attacks were carried out in the context of image classification tasks [7, 6, 28, 29]. The attacker's goal was to force a model to misclassify the input image. Many attack schemes were developed and tested on such models. One of the most cited is the so-called Fast Gradient Sign Method (FGSM) [6]. In their original work, Goodfellow et al [6] suggest a fast method to create adversarial input to a classification model. Consider an input x to a classification model M, a hyper-parameter ϵ, a loss function l and y_{true} - the target associated with x. Assume the model predicts a label y for an input x, i.e., M(x)=y. In their work they show that an adversarial example x_{adv} could then be computed by $x_{adv} = x + \epsilon Sign(\nabla(l(M(x), y_{true}))$. Shortly after [30] introduces a straightforward way to extend this method. They named this new approach the Iterative Fast Gradient Sign Method (IFGSM) [30]. They suggest to iteratively use the same update step on the input. To do so, set $x^0_{adv} = x$ and iteratively update $x^{i+1}_{adv} = x^i_{adv} + \epsilon \cdot Sign(\nabla(l(M(x^i_{adv}), y_{true}))$. Since these attacks are thoroughly researched and well understood [31, 6, 32] we adopt them to our attack approach.

Later on, adversarial attack methods that target specific objects in the image were introduced against object detectors [33–35]. In [33] it is shown how to force a SOTA detection model to classify all detections of a semantic class as another class while leaving all other detections unchanged. Liao *et al* [35] proposed a local attack that only perturb a specific detection bounding box, achieving a stronger effect than a global perturbation for the same attack budget. In [34], an analysis of object detection from the viewpoint of multi-task learning leads to a method to (partially) defend object detectors against adversarial attacks.

Recently, adversarial attacks have expanded beyond image classification and object detection to include dense prediction tasks such as semantic segmentation, depth, and optical flow. Promising results are shown in each of these tasks [36–39]. Such attacks often demonstrate the ability to target specific subsets of pixels rather than the entire image. For semantic segmentation, it was shown that pixels belonging to specific instances of pedestrians can be labeled by the attacked model as a road [36]. In [39], depth prediction has been successfully manipulated in many ways, such as removing the target entirely and aligning its depth with the surrounding background.

In the past couple of years, there is a growing interest in adversarial attacks on optical flow models [38, 40–43]. Ranjan *et al* [38] demonstrated the possible benefits of a patch attack against leading models. First, they showed that this attack is very successful against encoder-decoder like architectures, but less effective for spatial pyramid types of models. They also showed it to be reproducible in "real life" conditions, with a hostile patch printed on a board and displayed in front of a camera. A follow-up work [40] conjectures that a principal cause for the success of adversarial attacks on OF is the small size of their receptive field. [41, 42] introduce methods to corrupt the prediction of action recognition systems by attacking the OF modules they rely on. Finally, [44] proposes to

rank OF method, in addition to their prediction accuracy, by their robustness to AAs. Specifically, to quantify the robustness, they propose a strong attack against OF models which is easily bounded so the comparison between methods is valid. Differently from the above, we consider the effectiveness of adversarial attacks on OF methods from the perspective of the ability to hide them in the output, in addition to their impact on performance.

Finally, one of the most important tools in risk assessment and collision avoidance for autonomous agents, e.g., robots and autonomous vehicles, is estimating the TTC [45, 46, 1–4]. A popular approach to estimate the TTC is using OF [1–4]. For example, [1] fuses 2D OF vectors, and per-pixel estimated scale change, to "upgrade" 2D OF to 3D, allowing the direct computation of the TTC. We use [1] to demonstrate the impact of our consistency term on the TTC, and the benefits it has over the non-consistent attack.

2 Method

The inputs for an optical flow network f_{flow} are two $H\times W\times 3$ RGB images $I_1(x,y), I_2(x,y)$ where RGB channels ranges between [0,1]. The output is an $H\times W\times 2$ optical flow vector map $V(x,y)$. The goal of an attacker is to find an additive perturbation to the input that would shift the attacked optical flow map V' away from the original prediction V, as in [38].

To calculate this perturbation, we use two binary masks (see examples in Figure 2). The first mask, M_{target}, selects target pixels with the aim to change their optical flow, where in Figure 2, M_{target} is the vehicle's instance mask. The second mask, $M_{perturb}$, specifies the pixels that may change due to the pertrubation, where in Figure 2 we allow only the pixels of the "nature" category to change.

Consider the first mask M_{target} with N non-zero entries specifying the pixels of the object (category or instance) we aim to attack. Our attack term, $l_{attack}(V', V)$, is then defined by the $L1$ norm between the attacked V' and original V flows, averaged over M_{target}, as given by

$$l_{attack}(V', V) = \frac{1}{N} \sum_{(x,y)\in M} |V'(x,y) - V(x,y)|_1, \tag{1}$$

where $(x,y) \in M$ iff $M_{target}(x,y) = 1$.

To encourage the flow on the remaining scene to stay unaffected by the attack, we add a consistency term, $l_{con}(V', V)$, which is the negative $L1$ norm of the difference between original and attacked flows, averaged over non-attacked pixels:

$$l_{con}(V', V) = -\frac{1}{HW - N} \sum_{(x,y)\notin M} |V'(x,y) - V(x,y)|_1, \tag{2}$$

where $(x,y) \notin M$ iff $M_{target}(x,y) = 0$, and since N pixels are attacked, we have $HW - N$ non-attacked pixels.

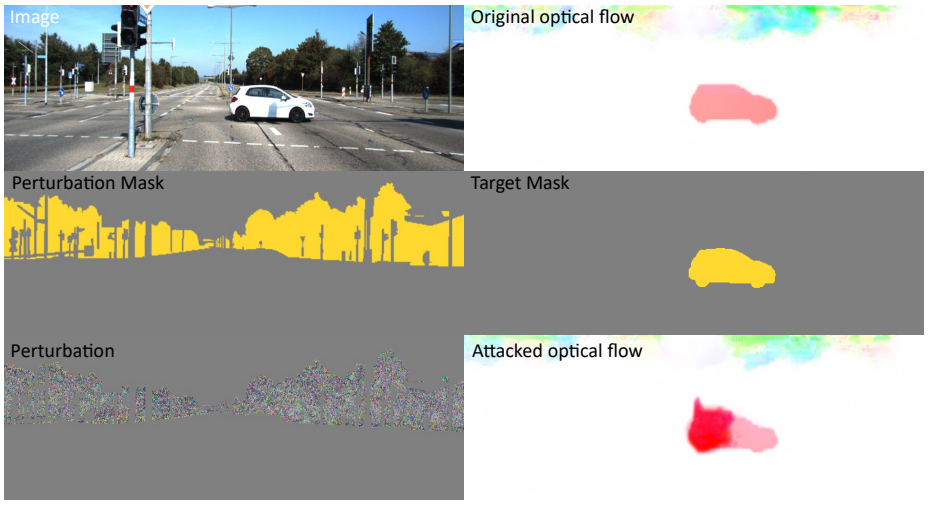

Fig. 2. Consistent attack method. In the top row, the original image (left) and OF (right) are presented. In the middle row, the perturbation mask (left) and target mask (right) are presented. In the bottom row, the perturbation on the "nature" category (left) and the attacked OF (right) are presented.

Our final loss is composed of these two terms, l_{attack} and l_{con}. The trade-off between the terms is controlled by the consistency coefficient α:

$$l_{total} = l_{attack} + \alpha l_{con}. \tag{3}$$

In order to attack semantic categories we require ground truth semantic labeling. This is only provided for the first image I_1 in the data we use for the attacks. Thus we have restricted our perturbation to the first image I_1. The second image I_2 is left unperturbed by our attack.

Let us denote the first image after the i-th perturbation as $I_1^{(i)}$, the i'th perturbation as $\delta I_1^{(i)}$ and the corresponding attacked flow $V^{(i)}$. Thus $V^{(0)} = V$ is the original flow, and $I_1^{(0)} = I_1$ is the unperturbed image. Since our first attack step is when $i = 1$ we have $\delta I_1^{(0)} = 0$.

Consider the second mask, $M_{perturb}$, with L non-zero entries, of the pixels we allow the attack to perturb. Given an attack strength coefficient ϵ, our i-th attack step follows the IFGSM [30] and given by

$$\begin{aligned} I_1^{(i)} &= I_1^{(i-1)} + \delta I_1^{(i)} \\ \delta I_1^{(i)} &= \epsilon \cdot M_{perturb} \cdot Sign\left(\nabla l_{total}(V^{(i-1)}, V)\right) \\ V^{(i)} &= f_{flow}(I_1^{(i)}, I_2), \end{aligned} \tag{4}$$

where *Sign* returns the negative or positive sign of its input.

In each attack step i we create a small perturbation to the first image $\delta I_1^{(i)}$. As shown in Figure 2 this perturbation is only applied in pixels (x, y) where $M_{perturb}(x, y) = 1$. It is equal to the sign of the loss function's gradient, weighted by the attack coefficient ϵ. After computing the i'th perturbation we add it to the image from the previous step $I_1^{(i-1)}$ to get the current perturbed input $I_1^{(i)}$. Inferring on this input with the optical flow network f_{flow} results in the i'th attacked optical flow map $V^{(i)}$. The loss between this flow $V^{(i)}$ and the original flow V will then be used to compute the perturbation for the following step. It is worth to note that for the first iteration ($i = 1$) we add a small amount of white noise to the original flow V so we would have non-zero gradients.

Let us define the target L1 norm of the perturbation as $||\Delta I||$. Given the number of perturbed pixels L and an estimated number of steps n for the attack, we set ϵ according to:

$$\epsilon = \frac{||\Delta I||}{n \cdot L} \tag{5}$$

We then iteratively update our input using Equation 4 until $||I_1^{(i)} - I_1||_1 \approx ||\Delta I||$ (up to 5%). We use $n = 2$ and $||\Delta I|| = 4 \cdot 10^{-3} (\approx 1/255)$ for most of our experiments, and will specifically state experiments with other values.

2.1 Implementation details

Throughout our experiments we use five optical flow models to evaluate the impact of adding our consistency term on targeted category-specific adversarial attacks - HD3 [16], PPAC [8], VCN [15], RAFT [14], LFN [13]. These models are some of the top performing methods on the KITTI"15 [23] dataset. We use the published, pre-trained models, given by the authors of each of the five chosen models. Since some models published multiple checkpoints, we always use the one fine-tuned on KITTI for our attack.

All of our experiments are performed and evaluated on the KITTI 12' [47] and KITTI 15' [23] datasets. These datasets contain a semantic segmentation labeling that we employ in our attacks. We could have used any semantic segmentation method [48–50] to label each scene. This would simulate a more realistic scenario where ground truth labeling is unavailable. However, it would also introduce another source of errors which we wish to avoid in order to focus our attention on consistent attacks.

We evaluate our attack using the average *end-point-error* (EPE) metric [11], which is the average $L2$ norm of the difference between attacked and original flows. The averaging is usually done over all image pixels, but since we are particularly interested in the effect of our attack on semantic classes, we compute the EPE averaged on pixels of specific classes. Using this metric the average shift in OF prediction due to the attack can be estimated for each class of interest.

In the subsequent section we will elaborate on the results from our main experiments. These experiments will encapsulate three different attack settings. These settings differ in the perturbed pixels mask ($M_{perturb}$, defined in Equation

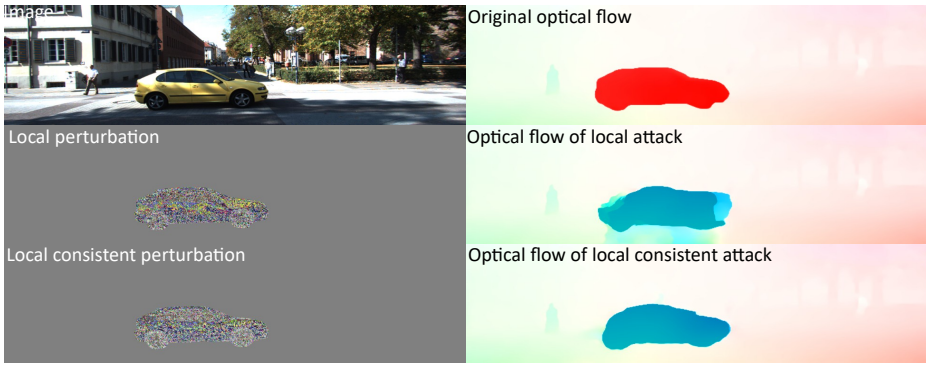

Fig. 3. A visualization of a local attack baseline and the impact of the consistency term on a vehicle instance using LFN [13] and $||\Delta I|| = 2 \cdot 10^{-2}$. Left: the original image and the perturbation optimized by each attack. Right: the corresponding optical flows. Adding a consistency term reduces the effect on non-vehicle pixels (as can be seen below the vehicle), while still significantly changing the vehicle's optical flow.

4) and the pixels we aim to attack (M_{target}, defined in Equation 1). In the first setting, a local attack, we perturb vehicle pixels and aim to attack the same subset of pixels. The second setting, a global attack, is where we perturb the entire image, but aim to attack vehicle pixels only. The third setting, a cross-category attack, is where we perturb the pixels of the "nature" catefory, and aim to attack vehicle category pixels (presented in the Supplementary material).

3 Experiments and Results

In this section, we present the experimental results obtained for the "vehicle" target category. The results for "human" target category, as well as the results obtained using the KITTI 12', are given in the supplementary.

3.1 Local attacks

Figure 3 visualizes an example local attack ($||\Delta I|| = 2 \cdot 10^{-2}$) experiment using the LFN model [13]. In this experiment a vehicle instance was attacked by only perturbing its pixels. Two attacks were conducted: a baseline, non -consistent, method with $\alpha = 0$ and a consistent attack with $\alpha = 10$.

Both attacks are successful in changing the car's optical flow and cause the previous right (red) moving vehicle to turn left (blue). However, the consistent attack preserves the non-targeted flow better, as can be seen by comparing the flow under the vehicle.

To quantify this effect, this experiment was expended to the entirety of the KITTI dataset. Here, for each image in the dataset we have attacked all of the

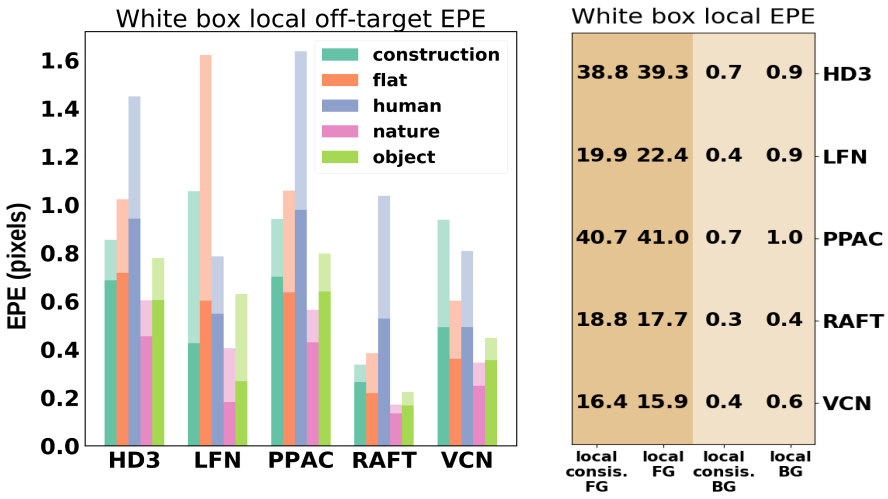

Fig. 4. Comparison between local attacks and local-consistent attacks on the KITTI dataset with $||\Delta I|| = 4 \cdot 10^{-3}$. Left - mean error caused by a local attack (transparent colors) and a consistent local attack (solid colors) over the corresponding category. Right - mean EPE for each model for the target (left) and off-target (right).

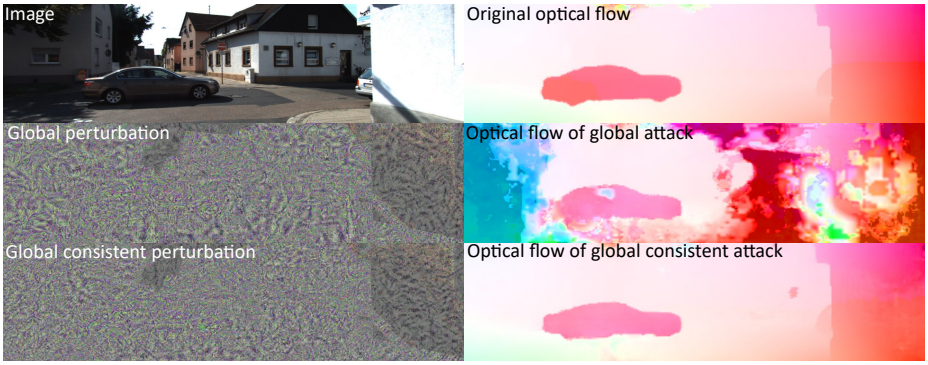

Fig. 5. A visualization of a global vehicle-targeting attack and the effect of adding a consistency term on HD3's [16] flow with $||\Delta I|| = 2 \cdot 10^{-2}$. Left: original image and perturbations. Right: the corresponding optical flows. Adding a consistency term reduces the effect on non-vehicle pixels while still significantly changing the vehicle optical flow

vehicles in that image (by perturbing vehicle pixels). We then evaluated the mean

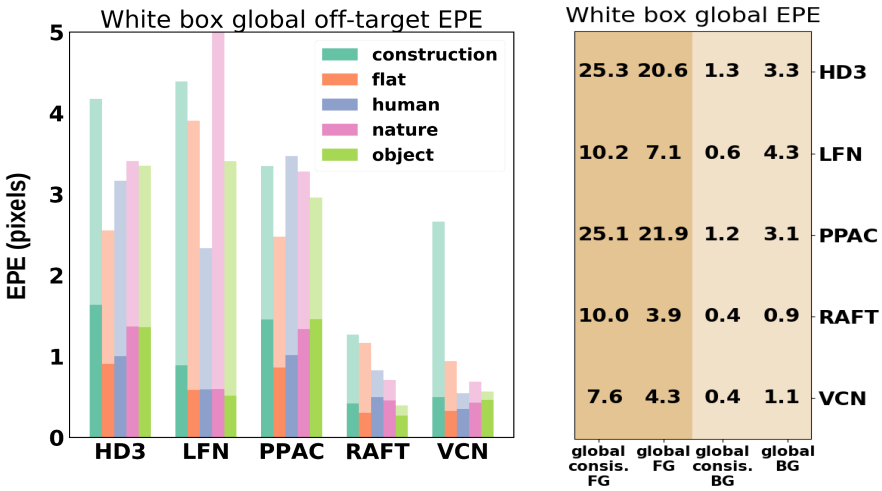

Fig. 6. Comparison between global attacks and global-consistent attacks on vehicles in the KITTI dataset with $||\Delta I|| = 4 \cdot 10^{-3}$. Left - mean error caused by a global attack (transparent colors) and a consistent global attack (solid colors) over the corresponding category. Right - mean EPE for each model for the target (left) and off-target (right).

EPE between original and attacked flow on selected categories: construction, flat, human, nature, object and vehicle.

Figure 4 presents the results on the KITTI dataset using the five selected models. The left sub-figure presents the (undesired) effect on non-targeted pixels per semantic category and the right sub-figure presents a summary for the (desired) effect on targeted pixels (left) and non-target pixels (right). We see that while the targeted vehicle category error does not vary much between attacks (right table) the non-targeted categories (left figure) suffer much less damage using a local consistent attack than our baseline non-consistent attack. The left side of the right sub-figure, that presents the targeted EPE, shows a small difference between attacks, while the left side of the right sub-figure shows a larger difference (in ratio). The effect on non-targeted categories is significantly reduced using our consistent attacks. In particular there is a 35% decrease on average (across methods) on the error induced on these categories.

3.2 Global attacks

One of the concerns with using a local attack is that since we perturb only a subset of the image pixels, we employ a high L_∞ norm to achieve the same L_1 norm as a *global* attack that perturbs the entire image. This, in turn, causes the local attack to be more perceptible compared to a global attack. Figure 5

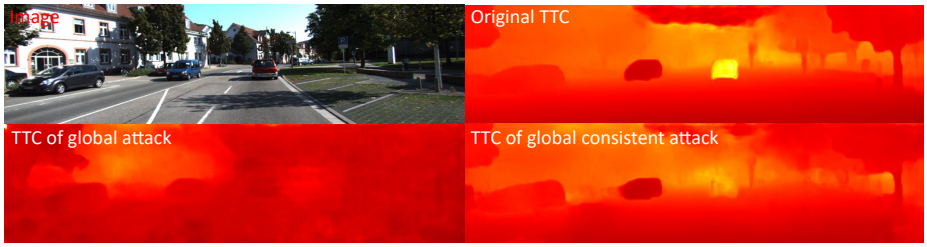

Fig. 7. Visualization of the TTC results for an AA on a vehicle instance using HD3 with $||\Delta I|| = 4 \cdot 10^{-3}$. Hot colors corresponds to shorter TTC than cold colors.

demonstrates this global attack in which we perturbed the entire image. The figure visually compares the results of the consistent and non-consistent attacks. The left column presents the original image and its perturbations. Here, unlike the local attack, the entire image is perturbed. The right column presents the effect both attacks have on HD3's optical flow. For the non-consistent attack we can notice multiple non-vehicle flow segments that changed drastically, turning the naturally smooth flow of the background into a rapidly varying flow. Repeating the methodology we used for the local settings, we expand this experiment by attacking all of the vehicle category in the KITTI dataset [23], and averaging the error over the selected classes.

Figure 6 presents the result of attacking all vehicles in a global setting over the KITTI dataset, for our five OF models with $||\Delta I|| = 4 \cdot 10^{-3}$. Similarly to the local case, the left and right side of the right sub-figures demonstrate the effect on the non-targeted pixels. The left side of the right sub-figure presents the effect of the targeted pixels. The resulting targeted vehicle category error is higher when using a consistent attack. Moreover, the non-targeted categories suffer significantly less damage using a global consistent attack than the baseline non-consistent attack. Thus, for example, using the consistent attack results in a 60% stronger effect on the targeted category (averaged across models), while removing 60% of the unwanted optical flow change on the remaining categories (averaged across models).

3.3 Time-To-Collision (TTC)

As discussed in Section 1, we emphasize the significance of adversarial attacks on OF models by their possible impact on TTC algorithms [1–4]. We chose the state-of-the-art TTC algorithm presented in [1], which uses OF to compute a per-pixel TTC. The model is supplied with the attacked OF instead of the original OF computed by the pre-trained VCN (without fine-tuning).

An attack on a vehicle instance, which is visualized in Figure 7, demonstrates the impact of the original flow, the global consistent and global non-consistent attacked flows, on the TTC. The TTC values are log-scaled and color-coded, where hot colors (redish) encode lower TTC than colder colors (yellowish-whitish). The

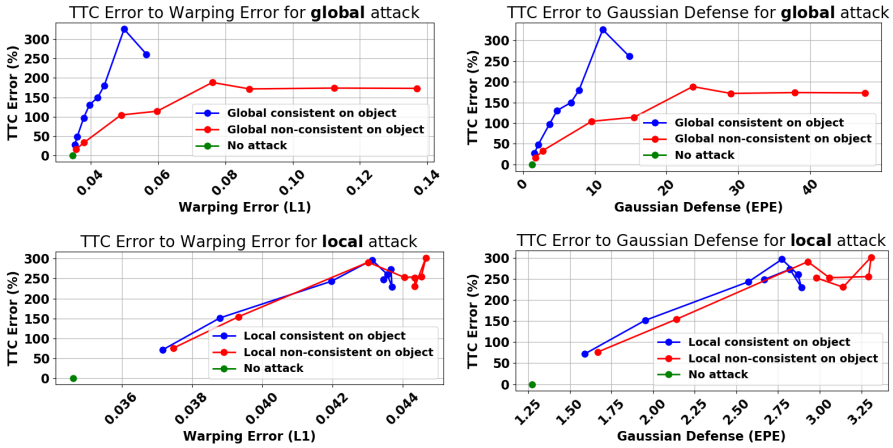

Fig. 8. TTC error to AA detection score. The top and bottom rows correspond to the global and local attacks, respectively. The Y-axis in all graphs corresponds to the TTC error, while the X-axis corresponds to the AA detection score using warping error and Gaussian defense for the left and right graphs, respectively. The graph is created using attacks with magnitude $||m \cdot 10^{-3}||$ for $m \in \{0.2, 0.4, 1.2, 2, 3.2, 4, 6, 8\}$.

attacked vehicle, which is yellow (high TTC) in the original flow, is significantly darker (low TTC) in both the consistent and non-consistent attacks. Importantly, the backgrounds of the original and consistent attack are quite similar, while the background of the not-consistent attack is very different.

As argued in Section 1, the effect of the off-target consistency loss term allows a better tradeoff between the impact on the TTC and a AA detection score, where by "detection score" we mean the output of an AA detection method. An example for this tradeoff would be that an attacked input with the same detection score will result in a higher average TTC impact. To quantify this tradeoff we've used three AA detection methods.

Warping error: The difference between I_1 and I_2 warped using the (attacked) OF V. That is, $||W_V(I_2) - I_1||_1$, where $W_V(I_2)$ warps I_2 using the OF V. The warping error is often used as an OF confidence measure [51]. Naturally, such confidence measure may be used to estimate an AA detection score.

The Gaussian/Median defenses [52]: The OF error (EPE) between the predicted flow and the flow from Gaussian/Median smoothed versions of the same images. That is, let $V'(I_1', I_2)$ be an attacked flow, and $V_K' = V'(K(I_1'), I_2)$ be an attacked flow (using the same attack) with I_1' smoothed using a 3×3 Gaussian/Median kernel K before the OF computation. $||V' - V_K'||_1$ is used to estimate the detection score. Such defense methods were used as AA detection methods in [52].

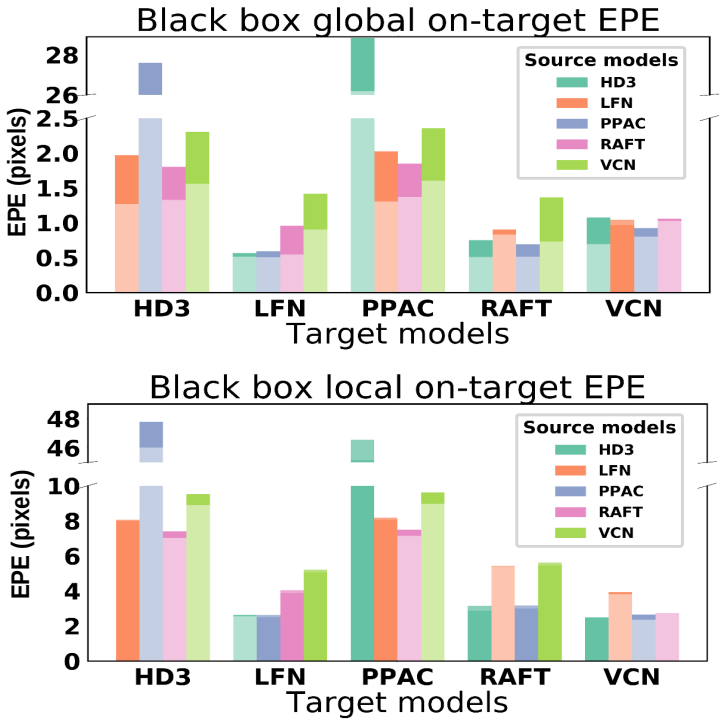

Fig. 9. "Black-box" setting results for the global attacks (top) and local attacks (bottom) on vehicles with $||\Delta I|| = 4 \cdot 10^{-3}$. Bar height signify the mean EPE over the vehicle category caused by an attack created using the source models (color-coded) on the target model (x-axis). Consistent and non-consistent attacks are marked by solid colors and transparent colors, respectively.

The graphs of the error in TTC as a function of the AA detection score are presented in Figure 8. (The median defense is presented in the supplementary material.) We measure the error in TTC as an average percentage of difference relative to the original TTC; that is, $|T_A - T_O|/T_O$, where T_A and T_O are the TTCs of the attacked and original flows, respectively. The graphs are created from 8 AA with different magnitudes, where all 5 OF models (in a white-box settings) are averaged per attack magnitude. In all 3 cases, the global consistent attack is superior to the global non-consistent attack in both the detection score (lower in X axis), and in impact on TTC (higher Y axis). In the local attack the trend is similar, however, the gap is much smaller. To conclude, the off-target consistency loss term is effective in terms of the TTC - detection score tradeoff.

3.4 Black-box attacks and transferability

Finally, we evaluated the transferability of the consistent attacks for the global and local attacks. To this end, we used each of the chosen models to attack the vehicle category in every image pair in the KITTI dataset. Each model was then evaluated on the adversarial datasets generated using the remaining models. The mean EPE over the vehicle targeted category for each attack is presented in Figure 9. We use transparent colors to visualize non-consistent attacks and solid colors to visualize for consistent attacks. We note that attacks created using HD3 seem to have a high impact on PPAC and vice versa, which could be related to HD3 and PPAC having most of their architecture shared.

Similar to the results presented in the white box settings, the local attack impact does not vary much with the addition of the consistency term. However, for the global case we observe a significant increase in the targeted impact transferred to other models. If we examine the results on RAFT, adding the consistency term resulted in a 44% increase in black-box attack strength, averaged across targeted models.

4 Discussion

To summarize, we presented a new methodology for targeted adversarial attacks against optical flow models. We introduced a new term to the attack, called 'consistency term', which is used to reduce the effect of the attack on the off-target pixels. In three different settings: local, global and cross category (supplementary), adding the consistency term to the loss reduces the impact on non-targeted object. Adding the term either preserves or increases the effect on the targeted category. Moreover, we have demonstrated that for some of the settings using a consistent attack results in a more transferrable attack. Finally, we have showed that for a TTC downstream task these attacks have a better detection - impact tradeoff, with an impact as high as 3x higher for the same detection score.

In our experiments we observe an obvious difference between the local and the global setting, where the effect on the non-targeted object is much more apparent in the global setting. In this setting, the danger of negatively impacting the rest of the scene is much greater since we directly change the non-target pixels. Adding the consistency term allows us to introduce global perturbations with a smaller effect on the resulting non-targeted optical flow. We also note that, in slight contrast to [38], our attacked models, which all have pyramid-like feature encoders are attacked successfully. A further analysis is left for future work.

An interesting follow-up for our work would be utilizing adversarial targeted attacks as a data augmentation technique for model training, which was demonstrated effective [53]. Other works [39] have demonstrated that some semantic classes are easier to attack than others. By leveraging consistent adversarial targeted attacks in its augmentation procedure, models might be able to learn a more robust representation of each semantic class. This, in turn, might decrease the probability of a successful attack against them [32], and increase the ability of a model to generalize its predictions for those classes [54].

References

1. Yang, G., Ramanan, D.: Upgrading optical flow to 3d scene flow through optical expansion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 1334–1343
2. Pedro, D., Matos-Carvalho, J.P., Fonseca, J.M., Mora, A.: Collision avoidance on unmanned aerial vehicles using neural network pipelines and flow clustering techniques. Remote Sensing **13** (2021) 2643
3. Blumenkamp, J.: End to end collision avoidance based on optical flow and neural networks. arXiv preprint arXiv:1911.08582 (2019)
4. Badki, A., Gallo, O., Kautz, J., Sen, P.: Binary ttc: A temporal geofence for autonomous navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 12946–12955
5. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial machine learning at scale. CoRR **abs/1611.01236** (2016)
6. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR. (2015)
7. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In Bengio, Y., LeCun, Y., eds.: ICLR (Poster). (2014)
8. Wannenwetsch, A.S., Roth, S.: Probabilistic pixel-adaptive refinement networks. In: CVPR, IEEE (2020) 11639–11648
9. Grosse, K., Manoharan, P., Papernot, N., Backes, M., McDaniel, P.: On the (statistical) detection of adversarial examples. arXiv preprint arXiv:1702.06280 (2017)
10. Tian, S., Yang, G., Cai, Y.: Detecting adversarial examples through image transformation. In: Thirty-Second AAAI Conference on Artificial Intelligence. (2018)
11. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. Int. J. Comput. Vis. **92** (2011) 1–31
12. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia conference on computer and communications security. (2017) 506–519
13. Hui, T., Tang, X., Loy, C.C.: Liteflownet: A lightweight convolutional neural network for optical flow estimation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 8981–8989
14. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: ECCV (2). Volume 12347 of Lecture Notes in Computer Science., Springer (2020) 402–419
15. Yang, G., Ramanan, D.: Volumetric correspondence networks for optical flow. In: NeurIPS. (2019) 793–803
16. Yin, Z., Darrell, T., Yu, F.: Hierarchical discrete distribution decomposition for match density estimation. In: CVPR, Computer Vision Foundation / IEEE (2019) 6044–6053
17. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2. IJCAI'81, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (1981) 674–679
18. Horn, B.K.P., Schunck, B.G.: Determining optical flow. ARTIFICAL INTELLIGENCE **17** (1981) 185–203
19. Sun, D., Roth, S., Black, M.: A quantitative analysis of current practices in optical flow estimation and the principles behind them. International Journal of Computer Vision **106** (2014) 115–137

20. Dosovitskiy, A., Fischer, P., Ilg, E., HÃďusser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: ICCV, IEEE Computer Society (2015) 2758–2766
21. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: CVPR, IEEE Computer Society (2017) 1647–1655
22. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: CVPR, IEEE Computer Society (2018) 8934–8943
23. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: CVPR, IEEE Computer Society (2015) 3061–3070
24. Hu, Y., Li, Y., Song, R.: Robust interpolation of correspondences for large displacement optical flow. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
25. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: CVPR, IEEE Computer Society (2017) 2720–2729
26. Hur, J., Roth, S.: Iterative residual refinement for joint optical flow and occlusion estimation. In: CVPR, Computer Vision Foundation / IEEE (2019) 5754–5763
27. Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. CoRR (2014)
28. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 427–436
29. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 2574–2582
30. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. In: ICLR (Workshop), OpenReview.net (2017)
31. TramÃĺr, F., Kurakin, A., Papernot, N., Goodfellow, I.J., Boneh, D., McDaniel, P.D.: Ensemble adversarial training: Attacks and defenses. In: ICLR (Poster), OpenReview.net (2018)
32. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: ICLR (Poster), OpenReview.net (2018)
33. Nezami, O.M., Chaturvedi, A., Dras, M., Garain, U.: Pick-object-attack: Type-specific adversarial attack for object detection. CoRR **abs/2006.03184** (2020)
34. Zhang, H., Wang, J.: Towards adversarially robust object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 421–430
35. Liao, Q., Wang, X., Kong, B., Lyu, S., Yin, Y., Song, Q., Wu, X.: Fast local attack: Generating local adversarial examples for object detectors. In: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE (2020) 1–8
36. Fischer, V., Kumar, M.C., Metzen, J.H., Brox, T.: Adversarial examples for semantic image segmentation. In: ICLR (Workshop), OpenReview.net (2017)
37. Arnab, A., Miksik, O., Torr, P.H.S.: On the robustness of semantic segmentation models to adversarial attacks. In: CVPR, IEEE Computer Society (2018) 888–897
38. Ranjan, A., Janai, J., Geiger, A., Black, M.J.: Attacking optical flow. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 2404–2413
39. Wong, A., Cicek, S., Soatto, S.: Targeted adversarial perturbations for monocular depth prediction. In: Advances in Neural Information Processing Systems. (2020)

40. Schrodi, S., Saikia, T., Brox, T.: Towards understanding adversarial robustness of optical flow networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2022) 8916–8924
41. Inkawhich, N., Inkawhich, M., Chen, Y., Li, H.: Adversarial attacks for optical flow-based action recognition classifiers. arXiv preprint arXiv:1811.11875 (2018)
42. Anand, A.P., Gokul, H., Srinivasan, H., Vijay, P., Vijayaraghavan, V.: Adversarial patch defense for optical flow networks in video action recognition. In: 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE (2020) 1289–1296
43. Yamanaka, K., Takahashi, K., Fujii, T., Matsumoto, R.: Simultaneous attack on cnn-based monocular depth estimation and optical flow estimation. IEICE Transactions on Information and Systems **104** (2021) 785–788
44. Schmalfuss, J., Scholze, P., Bruhn, A.: A perturbation constrained adversarial attack for evaluating the robustness of optical flow. arXiv preprint arXiv:2203.13214 (2022)
45. Manglik, A., Weng, X., Ohn-Bar, E., Kitani, K.M.: Future near-collision prediction from monocular video: Feasibility, dataset, and challenges. arXiv preprint arXiv:1903.09102 **1** (2019)
46. Mori, T., Scherer, S.: First results in detecting and avoiding frontal obstacles from a monocular camera for micro unmanned aerial vehicles. In: 2013 IEEE International Conference on Robotics and Automation, IEEE (2013) 1750–1757
47. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR). (2012)
48. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. CoRR **abs/1706.05587** (2017)
49. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40** (2017) 834–848
50. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). (2018) 801–818
51. Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C.: Learning to detect motion boundaries. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 2578–2586
52. Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155 (2017)
53. Liu, L., Zhang, J., He, R., Liu, Y., Wang, Y., Tai, Y., Luo, D., Wang, C., Li, J., Huang, F.: Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 6489–6498
54. Stutz, D., Hein, M., Schiele, B.: Disentangling adversarial robustness and generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 6976–6987