Consistent Semantic Attacks on Optical Flow - Supplementary Material

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

In this supplementary we provide additional material on consisted targeted attacks. First, we provide the details of the cross category attacks against the vehicle category conducted on KITTI 15'. Then, we analyze the effect of varying α on attack EPE. Following that, we provide our results on both the human category attack on KITTI 15' and the vehicle category attack on KITTI 12'. The results for the TTC as a function of the Median AA detection score is then presented. Finally, we provide further attack visualizations for the experiments conducted on KITTI 15'

1 Cross category attacks

In the cross-category setting, we perturb the nature category pixels in order to manipulate the optical flow of vehicle pixels. Figure 1 visualizes such an attack and shows the original input and flow, compared to the cross category and consistent cross category attacks. First, the second row of Figure 1 shows that our non-consistent attack results in a significant change to both the carâĂŹs flow and its environment. Both the sky, the nature pixels, and large portions of the road are affected by this attack. Then, once consistency is added (third row), the changes to the non-vehicle pixels are much less apparent, with the vehicleâĂŹs flow still significantly changed.

We've evaluated the cross category effect on the KITTI 15' dataset. Here, we've used $|\Delta I| = 4 \cdot 10^{-3}$ to perturb nature pixels. The target of the attack was vehicle pixels. Figure 2 presents the evaluation of this experiment. It shows the EPE comparing to the original flow on the off-target categories (left) and the on-target vehicle category (right). Two distinct feature of this attack settings are noticeable in the left figure. First, all of the categories that were neither attacked nor perturbed (human, flat, construction, object) are less effected from the consistent attack. Second, the perturbed category (nature) is highly effected in the non-consistent attack. This, since nature pixels are altered in this attack. Adding a consistency term greatly reduce the error we observe on the perturbed category.

2 Varying the consistency parameter

In the method section we've introduced the consistency parameter, α. This parameter controls a trade-off of the consistent attack. High values of α lead to an

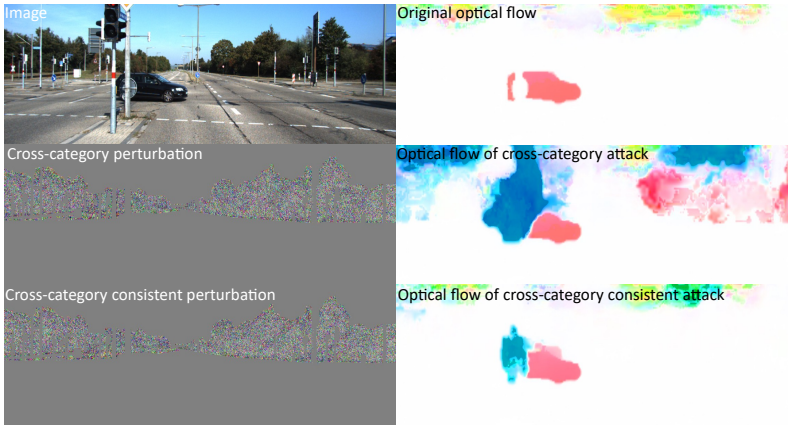

Fig. 1. A visualization of a cross-category attack baseline and the effect of adding a consistency term on a vehicle instance using HD3 with $|\Delta I| = 6.5 \cdot 10^{-2}$. While the original flow of the vehicle indicate a right-moving instance, the attacked flows also indicates a left moving instance. In the non-consistent case (second row) we see that the remaining scene is affected as well, and once a consistency term is being added (last row) the attack is much more focused on the attacked instance

attack that is focused on preserving non-target flow. Low values focus instead on damaging the optical flow of the target flow.

To demonstrate the trade-off α controls and the role of the consistency term as a regularization term we've conducted the following experiment. We've repeated the same global attack against vehicles described in our work with various values of α. These values range from $\alpha = 0.01$ to $\alpha = 100$. For each attack we've measured two metrics. First, the mean EPE (compared to original flow) on all target pixels. Second, the mean EPE on all non-targeted categories: construction, flat, nature, object, human.

The results of this experiment are presented in Figure 3. The left figure presents the effect on the off-target EPE. It shows that the more we increase α the less we modify the non-targeted flow. Once α is large enough its regularizing effect seem to stable and the off-target EPE plateaus. The right figure presents the effect on target EPE. For small enough values of α there is an increase in the attack efficiency on the target. This is a result of α's role as a regularization. For small values, it constrains the system to find a more efficient attack. For larger values of α we see a decrease in attack efficiency. This happens when the regularization term becomes very dominant. Instead of finding a more efficient perturbation we now focus the attack on preserving off-target flow.

3 Human attacks

While we focused our experiments on the vehicle category, other categories could be attacked as well. Here, we evaluate our attack against the human category

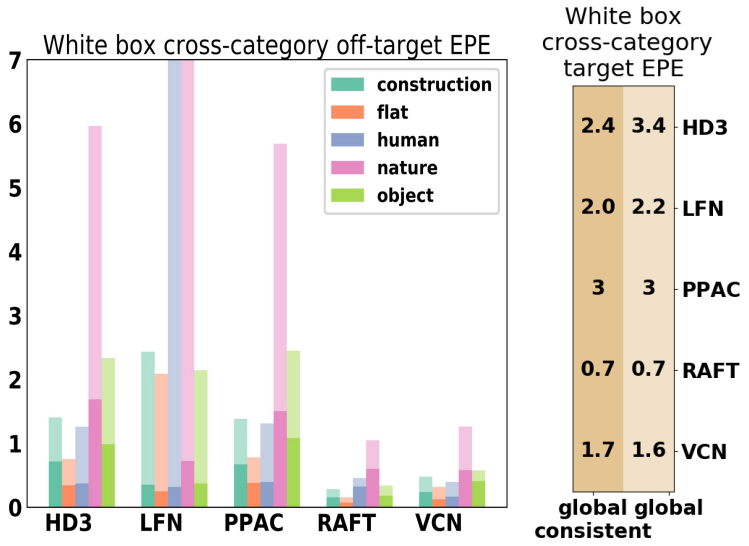

Fig. 2. Comparison between cross category attacks and consistent cross category on vehicles in the KITTI dataset with $|\Delta I| = 4 \cdot 10^{-3}$. For each model, we attacked vehicle pixels by perturbing nature pixels. We then evaluated the mean error caused by the non-consistent attack (clear colors) and the consistent attack (solid colors) over the corresponding category. Consistent attacks reduce off-target damage and keep attack efficiency similar. The effect on off-target is especially high on the perturbed nature category.

of the KITTI 15' dataset. We use two settings, global and cross-category, to attack this category and report our results. We did not use the third, local, setting presented earlier. This is since humans are usually composed by a small amount of pixels. Thus, the amount of perturbation we can introduce when only perturbing human pixels is highly limited.

We begin by reporting the global attack results on the human category. Here, we've perturbed the entire image in order to alter the optical flow of the human category pixels. Figure 4 presents the result of running this experiment on the entire KITTI 15' dataset. It shows the mean EPE (with respect to the original flow) averaged on the attacked, human, category (right). It also presents the same metric averaged on non-targeted categories (left). Here we see two effects previously demonstrated on the vehicle category. First, the consistent attack resulted in much less damage to the off-target categories. Thus for example, for HD3 the flat category error has decreased by 50%. Moreover, we see an increase in attack efficiency on the target. Thus, for example, using global consistent attacks increases the mean error on target pixels by 58% on average.

Next, we conducted a cross-category attack against the human category. In this setting we perturbed nature pixels to damage the optical flow of the human category. Figure 5 presents the result of this experiment on the KITTI

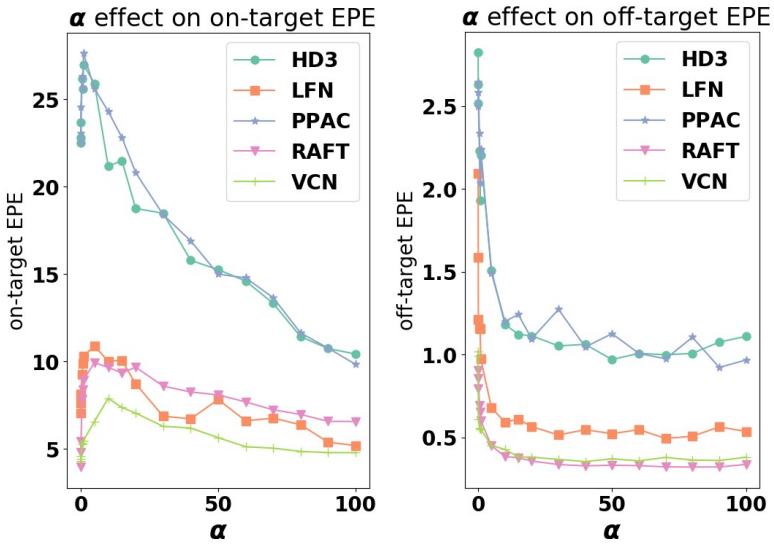

Fig. 3. The effect of α on attack metrics. Varying α effects both on-target EPE (left) and off-target EPE (right). Left figure shows that increasing α initially increase attack efficiency and then reduces it. Right figure shows that increasing α decreases our effect on non-targeted categories.

15' dataset. First, we see that the off-target categories error reduce using this attack. Thus, for example LFN reduced its error on the vehicle category by approximately 75%. The attack efficiency, however, remained similar for both attacks.

4 KITTI 12' dataset attacks

Throughout our work we have presented multiple experiments conducted using the KITTI 15' dataset. Here, we utilize an additional dataset, KITTI 12' [1], to evaluate consistent attacks against optical flow. The KITTI 12' optical flow dataset contains 193 image-pairs. Among those images, only the first 65 contain pixel-wise segmentation ground truth [1]. The rest of the images contain only vehicle labeling.

Since our experiments require pixel-wise segmentation, we restrict our evaluation on the KITTI 12' dataset to the first 65 images. We report our results on these images using the same evaluation protocol described in our method section. The consistent attack is evaluated under three settings: global, local, and cross-category attacks.

In the local setting we've perturbed vehicle pixels and evaluated our attack on vehicle pixels as well. Figure 6 presents the results of this experiment. We see that the off-target EPE was reduced in the consistency attack. Moreover, the on-target EPE varied only slightly by this addition.

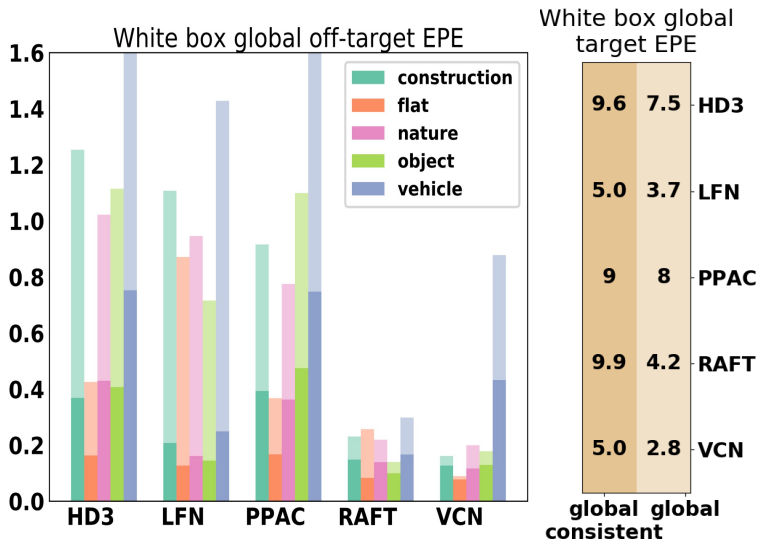

Fig. 4. Comparison between global attacks and global consistent attacks on humans in the KITTI dataset with $|\Delta I| = 4 \cdot 10^{-3}$. For each model, we attacked vehicle pixels by perturbing the entire image and evaluated the mean error caused by the global attacks (clear colors) and the consistent global attacks (solid colors) over the corresponding category. Consistent attacks reduce off-target damage and keep attack efficiency similar.

For the global setting we've perturbed the entire image pixels to attack vehicle pixels. Figure 7 visualizes the result of the global experiment. As for the KITTI 15' case, we see that adding consistency reduces off-target EPE and increases on-target EPE.

Finally, for the cross-category setting we've perturbed nature pixels to attack vehicle pixels. Figure 8 visualizes the results of this experiment. We notice three main aspects of this experiment. First, the off-target EPE on all category was reduced. Second, the nature category that was perturbed gained the most adding a consistency term to the attack. Last, we see that the on-target efficiency has remained similar once adding the consistency term.

Figure 9 visualizes the effect of adding consistency for the KITTI 12' dataset attacks. In both the attacked flows (bottom row) we note that the car flow has changed. In the non-consistent attack (bottom left) we see that the remaining scene flow has changed as well. For the consistent case (bottom right), the only visible change is that of the car.

5 TTC: The Median AA detection score

Similarly to the results of the TTC as a function of the AA detection scores presented in Section 3.3 of the paper, we present the Median AA detection score, as a function of the TTC error. The graphs for the TTC error as a function of

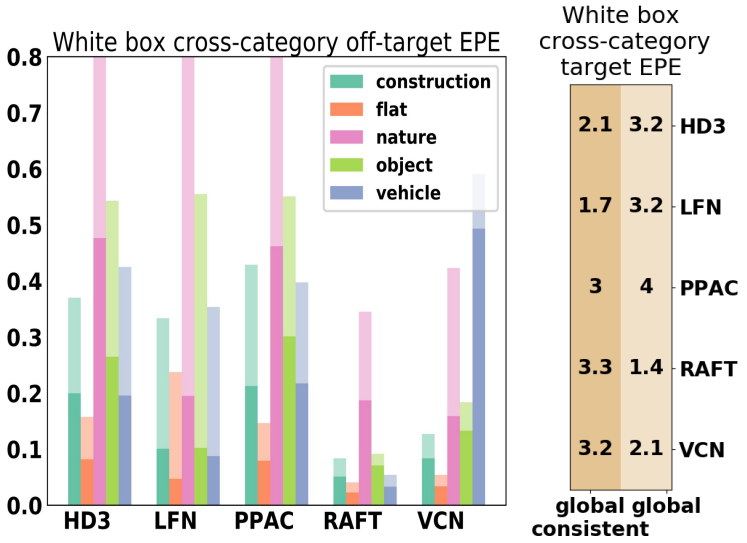

Fig. 5. Comparison between cross category attacks and consistent cross category on humans in the KITTI dataset with $|\Delta I| = 4{\cdot}10^{-3}$. For each model, we attacked humans pixels by perturbing nature pixels. We then evaluated the mean error caused by the non-consistent attack (clear colors) and the consistent attack (solid colors) over the corresponding category. Consistent attacks reduce off-target damage and keep attack efficiency similar. The effect on off-target is especially high on the perturbed nature class.

the Median detection score is presented in Figure 10. The results and trends for the TTC error as a function of the Median score are very similar to that of the warping error and the Gaussian AA detection scores.

6 KITTI 15' attack visualizations

In this section we provide further visualizations of the vehicle attack conducted on the KITTI 15' dataset. Example of the local and global attack results are visualized. Each example include the first image of the optical flow pair, I_1, the original flow, consistent flow, and the non-consistent flow.

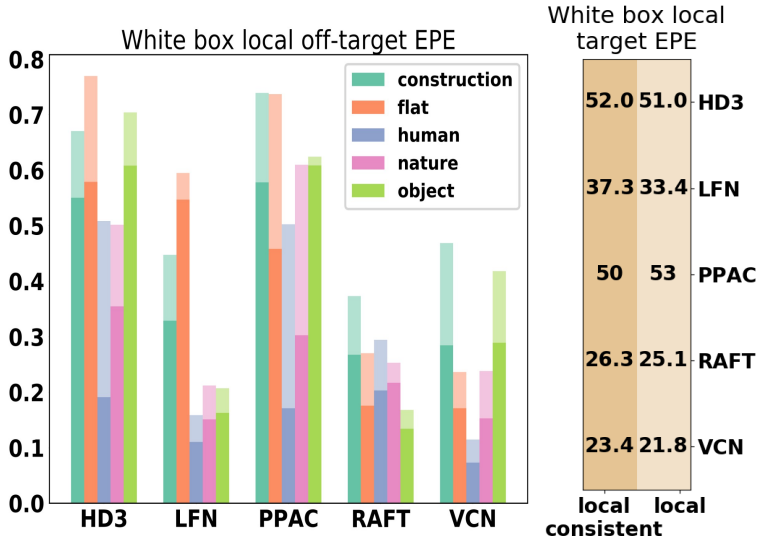

Fig. 6. Comparison between local attacks and local consistent attacks on vehicles in the KITTI 12' dataset with $|\Delta I| = 4 \cdot 10^{-3}$. For each model, we attacked vehicle pixels by perturbing vehicle pixels and evaluated the mean error caused by the global attacks (clear colors) and the consistent global attacks (solid colors) over the corresponding category.

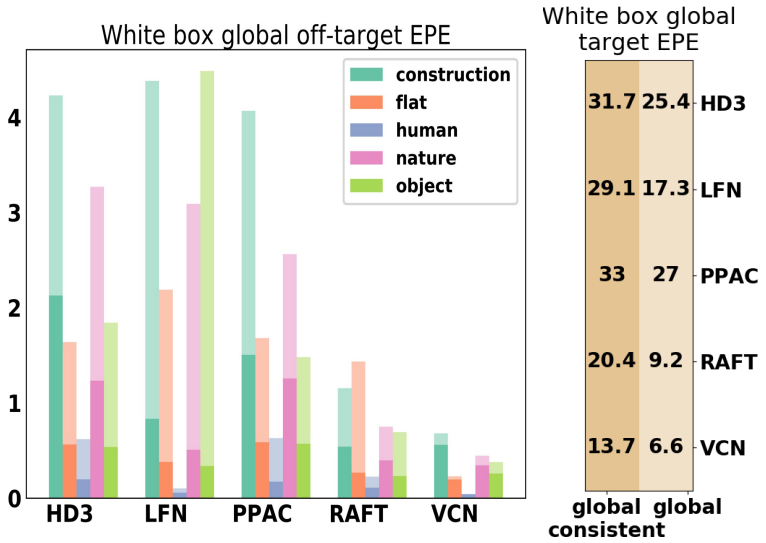

Fig. 7. Comparison between global attacks and global consistent attacks on vehicles in the KITTI 12' dataset with $|\Delta I| = 4 \cdot 10^{-3}$. For each model, we attacked vehicle pixels by perturbing the entire image and evaluated the mean error caused by the global attacks (clear colors) and the consistent global attacks (solid colors) over the corresponding category.

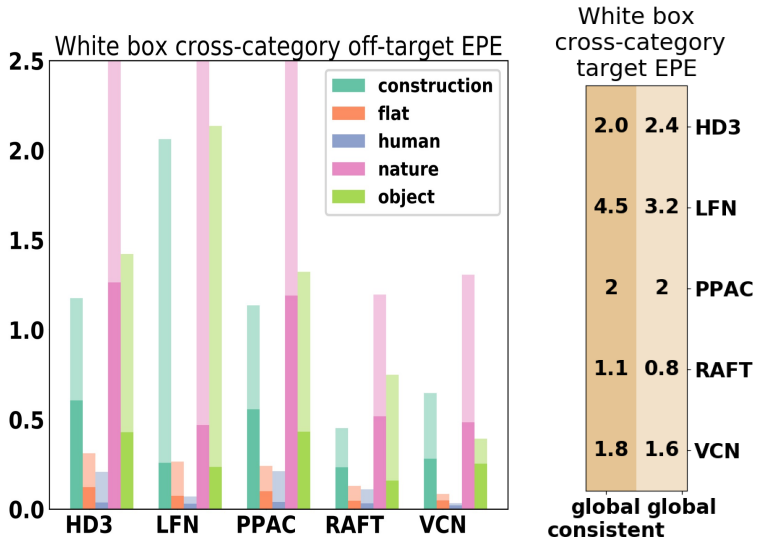

Fig. 8. cross category attacks on vehicles in the KITTI 12' dataset with $|\Delta I| = 2 \cdot 10^{-2}$. For each model, we attacked vehicle pixels by perturbing nature category pixels and evaluated the mean error caused by the global attacks (clear colors) and the consistent global attacks (solid colors) over the corresponding category.

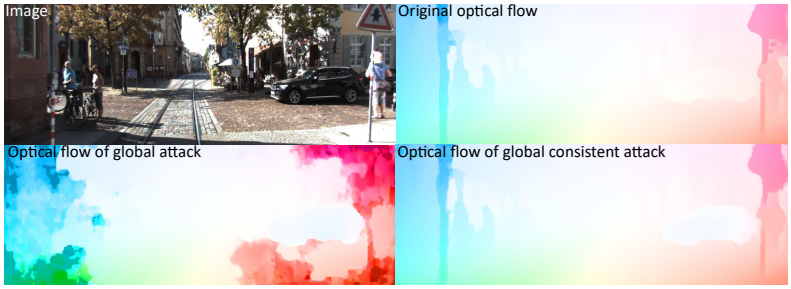

Fig. 9. A visualization of a global attack baseline and the effect of adding a consistency term on a vehicle instance using PPAC with $|\Delta I| = 4 \cdot 10^{-3}$. In the non-consistent case (second row) we see that the remaining scene is affected as well. Once a consistency term is being added (last row) the attack is much more focused on the attacked instance

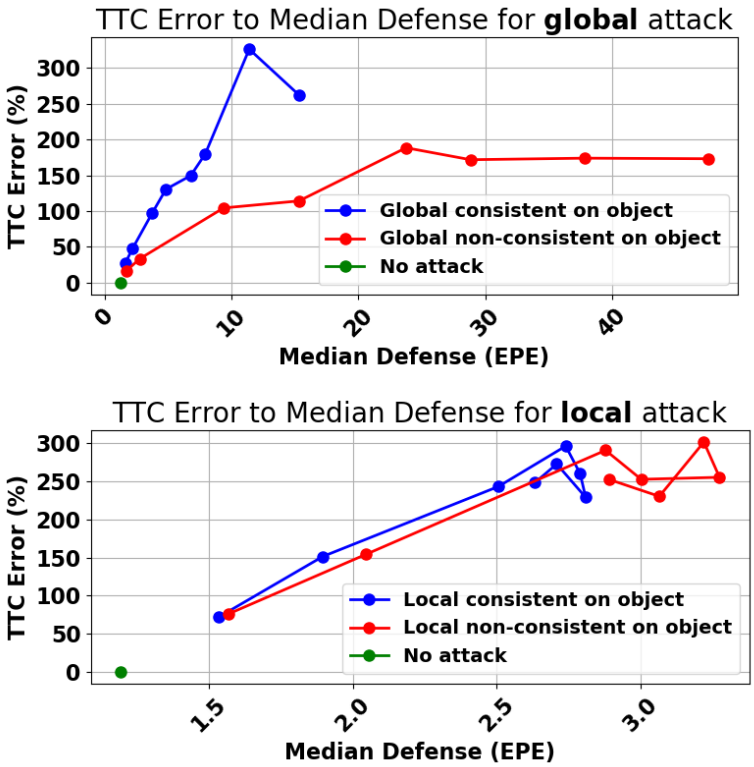

Fig. 10. TTC error to the Median AA detection score. The left and right sub-figures correspond to the global and local attacks, respectively. The Y-axis in all graphs corresponds to the TTC error, while the X-axis corresponds to the Median AA detection score. The graph is created using attacks with magnitude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

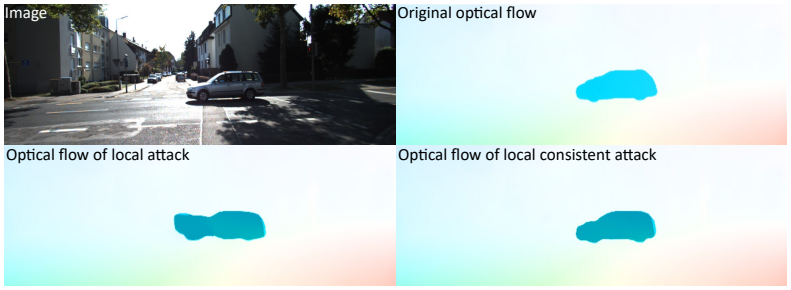

Fig. 11. A visualization of a local attack baseline and the effect of adding a consistency term on a vehicle instance using LFN with $|\Delta I| = 4 \cdot 10^{-3}$. Note that vehicle outline changes. In the non-consistent case (second row) we see that the remaining scene is affected as well, and once a consistency term is being added (last row) the attack is much more focused on the attacked instance

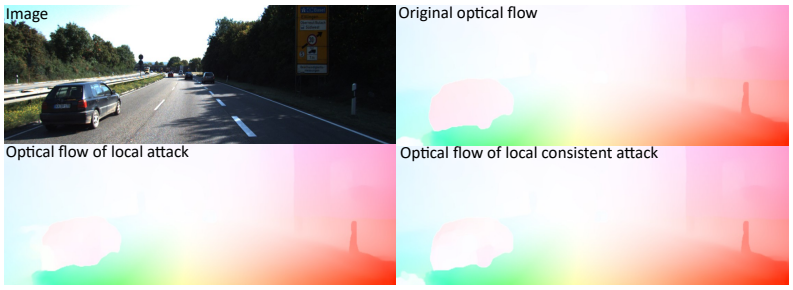

Fig. 12. A visualization of a local attack baseline and the effect of adding a consistency term on a vehicle instance using LFN with $|\Delta I| = 4 \cdot 10^{-3}$. The road behind the vehicle is effected from the attack. In the non-consistent case (second row) we see that the remaining scene is affected as well, and once a consistency term is being added (last row) the attack is much more focused on the attacked instance

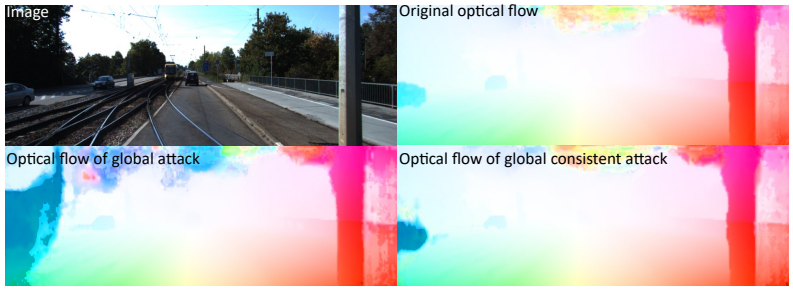

Fig. 13. A visualization of a global attack baseline and the effect of adding a consistency term on a vehicle instance using HD3 with $|\Delta I| = 4 \cdot 10^{-3}$. In the non-consistent case (second row) we see that the remaining scene is affected as well, and once a consistency term is being added (last row) the attack is much more focused on the attacked instance

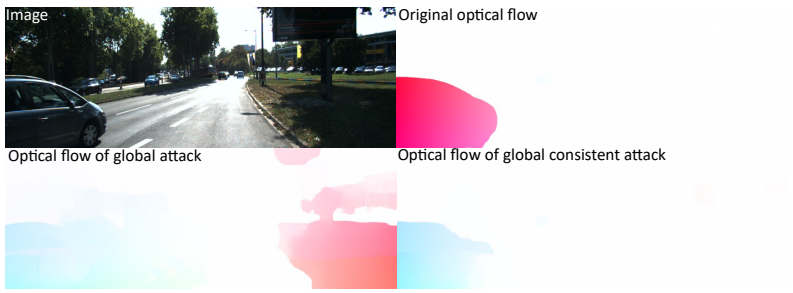

Fig. 14. A visualization of a global attack baseline and the effect of adding a consistency term on a vehicle instance using RAFT with $|\Delta I| = 4 \cdot 10^{-2}$. In the non-consistent case (second row) we see that the remaining scene is affected as well, and once a consistency term is being added (last row) the attack is much more focused on the attacked instance

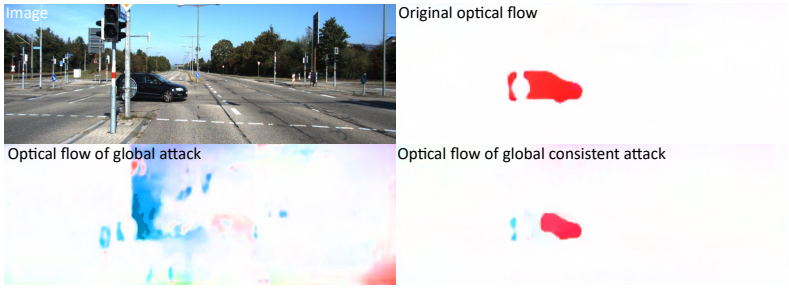

Fig. 15. A visualization of a global attack baseline and the effect of adding a consistency term on a vehicle instance using VCN with $|\Delta I| = 4 \cdot 10^{-2}$. In the non-consistent case (second row) we see that the remaining scene is affected as well, and once a consistency term is being added (last row) the attack is much more focused on the attacked instance

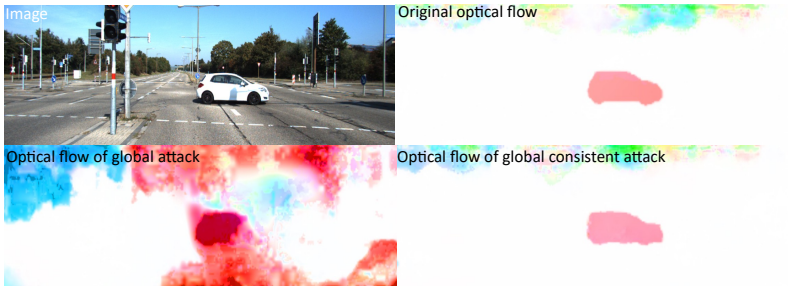

Fig. 16. A visualization of a global attack baseline and the effect of adding a consistency term on a vehicle instance using HD3 with $|\Delta I| = 2 \cdot 10^{-2}$. In the non-consistent case (second row) we see that the remaining scene is affected as well, and once a consistency term is being added (last row) the attack is much more focused on the attacked instance

References

1. 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)