Physical Passive Patch Adversarial Attacks on Visual Odometry Systems

Yaniv Nemcovsky*[1], Matan Yaakoby*[1], Alex M. Bronstein [1], Chaim Baskin [1]
[1] Technion – Israel Institute of Technology

Abstract

Deep neural networks are known to be susceptible to adversarial perturbations – small perturbations that alter the output of the network and exist under strict norm limitations. While such perturbations are usually discussed as tailored to a specific input, a universal perturbation can be constructed to alter the model's output on a set of inputs. In this work, we study physical passive patch adversarial attacks on visual odometry-based autonomous navigation systems. To the best of our knowledge, we show for the first time that the error margin of a visual odometry model can be significantly increased by deploying patch adversarial attacks in the scene. We provide evaluation on synthetic closed-loop drone navigation data and demonstrate that a comparable vulnerability exists in real data. A reference implementation of the proposed method and the reported experiments is provided at `https://github.com/patchadversarialattacks/patchadversarialattacks`.

1. Introduction

Deep neural networks (DNNs) were the first family of models discovered to be susceptible to adversarial perturbations – small bounded-norm perturbations of the input that significantly alter the output of the model [3, 10] (methods for producing such perturbations are referred to as adversarial attacks). Such perturbations are usually discussed as tailored to a specific model and input, however, universal adversarial attacks are another setting where the aim is to produce an adversarial perturbation for a set of inputs [4, 7, 13]. Universal perturbations present a more realistic case of adversarial attacks, as awareness of the model's exact input is not required.

Monocular visual odometry (VO) models aim to infer the relative camera motion (position and orientation) between two corresponding viewpoints. In the present work, we investigate the susceptibility of VO models to universal adversarial perturbations, aiming to mislead a corresponding navigation system by disrupting its ability to spatially position itself in the scene. Previous works that discuss adversarial attacks on regression models mostly discuss standard adversarial attacks where the perturbation is inserted directly into a single image [2, 6, 8, 12]. In contrast, we take into consideration a time evolving process where a physical passive patch adversarial attack is inserted into the scene and is perceived differently from multiple viewpoints. This is a highly realistic settings, as we test the effect of a moving camera in a perturbed scene, and do not require direct access to the model's input. Below, we outline our main contributions.

Firstly, we produce physical patch adversarial perturbations for VO systems on both synthetic and real data. Our experiments show that while VO systems are robust to random perturbations, they are susceptible to such adversarial perturbations. For a given trajectory containing multiple frames, our attacks are aimed to maximize the generated deviation in the physical translation between the accumulated trajectory motion estimated by the VO and the ground truth. We show that inserting a physical passive adversarial patch into the scene substantially increases the generated deviation.

Secondly, we continue to produce universal physical patch adversarial attacks, which are aimed at perturbing unseen data. We optimize a single adversarial patch on multiple trajectories and test the attack on out-of-sample unseen data. Our experiments show that when used on out-of-sample data, our universal attacks generalize and again cause significant deviations in trajectory estimates produced by the VO system.

Lastly, we further test the robustness of VO systems to our previously produced universal adversarial attacks in a closed-loop scheme with a simple navigation scheme, on synthetic data. Our experiments show that in this case as well, the universal attacks force the VO system to deviate from the ground truth trajectory. To the best of our knowledge, ours is the first time the vulnerability of visual navigation systems to adversarial attacks is demonstrated, and, possibly, the first instance of adversarial attacks on closed-loop control systems.

The rest of the paper is organized as follows: Sec. 2 describes our proposed method, Sec. 3 provides our experi-

[*] Equal contribution.

mental results, and Sec. 4 concludes the paper.

2. Method

2.1. Patch adversarial attack setting

Let $\mathcal{I} = [0, 1]^{3 \times w \times h}$ be a normalized RGB image space, for some width w and height h. For an image $I \in \mathcal{I}$, inserting a patch image $P \in \mathcal{I}$ onto a given plane in I would then be a perturbation $A : (\mathcal{I} \times \mathcal{I}) \to \mathcal{I}$. Let $I^0, I^1 \in \mathcal{I}$ be the black and white albedo images of the patch P as viewed from viewpoint I, and let $H : \mathcal{I} \to \mathcal{I}$ be the linear homography transformation of P to viewpoint I, then:

$$I^P := A(I, P) = H(P) * (I^1 - I^0) + I^0 \qquad (1)$$

where $*$ denotes element-wise multiplication.

Let $VO : (\mathcal{I} \times \mathcal{I}) \to (\mathbb{R}^3 \times so(3))$ be a monocular VO model, i.e., for a given pair of consecutive images $\{I_t, I_{t+1}\}$, it estimates the relative camera motion $\delta_t^{t+1} = (q_t^{t+1}, R_t^{t+1})$, where $q_t^{t+1} \in \mathbb{R}^3$ is the $3D$ translation and $R_t^{t+1} \in so(3)$ is the $3D$ rotation. We define a trajectory as a set of consecutive images $\{I_t\}_{t=0}^{L}$, for some length L, and extend the definition of the monocular visual odometry to trajectories $VO(\{I_t\}_{t=0}^{L}) = \{VO(I_t, I_{t+1})\}_{t=0}^{L-1}$. Given a trajectory $\{I_t\}_{t=0}^{L}$, with ground truth motions $\{\delta_t^{t+1}\}_{t=0}^{L-1}$ and a criterion over the trajectory motions ℓ, an adversarial patch perturbation $P_a \in \mathcal{I}$ aims to maximize the criterion over the trajectory. Similarly, for a set of trajectories $\{\{I_{i,t}\}_{t=0}^{L_i}\}_{i=0}^{N-1}$, with corresponding ground truth motions $\{\{\delta_{i,t}^{t+1}\}_{t=0}^{L_i-1}\}_{i=0}^{N-1}$, a universal adversarial attack aims to maximize the sum of the criterion over the trajectories. Formally:

$$P_a = \arg\max_{P \in \mathcal{I}} \ell(\{VO(A(I_t, P))\}_{t=0}^{L}, \{\delta_t^{t+1}\}_{t=0}^{L-1}) \qquad (2)$$

$$P_{ua} = \arg\max_{P \in \mathcal{I}} \sum_{i=0}^{N-1} \ell(\{VO(A(I_{i,t}, P))\}_{t=0}^{L_i}, \{\delta_{i,t}^{t+1}\}_{t=0}^{L_i-1}) \qquad (3)$$

For the scope of this paper, the target criterion used for adversarial attacks is the RMS (root mean square) deviation in the $3D$ physical translation between the accumulated trajectory motion, as estimated by the VO, and the ground truth. We denote the accumulated motion as $\delta_0^L = \prod_{t=0}^{L-1} \delta_t^{t+1}$, where the multiplication of motions is defined as the matrix multiplication of the corresponding 4×4 matrix representation: $\delta_t^{t+1} = \begin{pmatrix} R_t^{t+1} & q_t^{t+1} \\ \mathbf{0} & 1 \end{pmatrix}$. The target criterion is then formulated as:

$$\ell_{VO}(VO(A(\{I_t\}_{t=0}^{L}, P)), \{\delta_t^{t+1}\}_{t=0}^{L-1})$$
$$= ||q(\prod_{t=0}^{L-1} VO(I_t^P, I_{t+1}^P)) - q(\prod_{t=0}^{L-1} \delta_t^{t+1})||_2 \qquad (4)$$

where we denote $q(\delta_0^L) = q((q_0^L, R_0^L)) = q_0^L$.

2.2. Optimization of adversarial patches

We optimize the adversarial patch P via a PGD adversarial attack [5] with ℓ_{inf} norm limitation. We limit the values in P to be in $[0, 1]$; however, we do not enforce any additional ϵ limitation, as such would be expressed in the albedo images. We allow for different training and evaluation criteria, and to enable evaluation on unseen data for universal attacks, we allow for different training and evaluation datasets. In the supplementary material, we provide algorithms for both our PGD (Algorithm 1) and universal (Algorithm 2) attacks.

For both optimization and evaluation of attacks we consider one of two criteria. The first criterion, which we denote as ℓ_{RMS}, is a smoother version of the target criterion ℓ_{VO}, in which we sum over partial trajectories with the same origin as the full trajectory. Similarly, the second criterion, which we denote as ℓ_{MPRMS}, i.e., mean partial RMS, is to take into account all the partial trajectories. Nevertheless, we take the mean for each length of partial trajectories in order to keep the factoring between different lengths as in ℓ_{RMS}. Formally:

$$\ell_{RMS}(VO(A(\{I_t\}_{t=0}^{L}, P)), \{\delta_t^{t+1}\}_{t=0}^{L-1})$$
$$= \sum_{l=1}^{L} \ell_{VO}(VO(A(\{I_t\}_{t=0}^{l}, P)), \{\delta_t^{t+1}\}_{t=0}^{l-1}) \qquad (5)$$

$$\ell_{MPRMS}(VO(A(\{I_t\}_{t=0}^{L}, P)), \{\delta_t^{t+1}\}_{t=0}^{L-1})$$
$$= \sum_{l=1}^{L} \frac{1}{L - l + 1} \sum_{i=0}^{L-l} \ell_{VO}(VO(A(\{I_t\}_{t=i}^{i+l}, P)), \{\delta_t^{t+1}\}_{t=i}^{i+l-1}) \qquad (6)$$

3. Experiments

We now present an empirical evaluation of the proposed method. We first describe the various experimental settings used for estimating the effect of the adversarial perturbations. We continue and describe the methodology for generation of both the synthetic and real datasets. Finally, we present our experimental results, first on the synthetic dataset and afterwards on the real dataset. In the supplementary material, we further detail the experimental settings and the data generation as well as discuss the used VO model.

Experimental setting In our experiments, we differentiate between three distinct settings. Firstly, the in-sample setting used to estimate the effect of PGD and universal adversarial perturbations on known data. Secondly, the out-of-sample setting used to estimate generalization properties of universal perturbations to unseen data. Finally, the closed-loop setting used to estimate the generalization properties

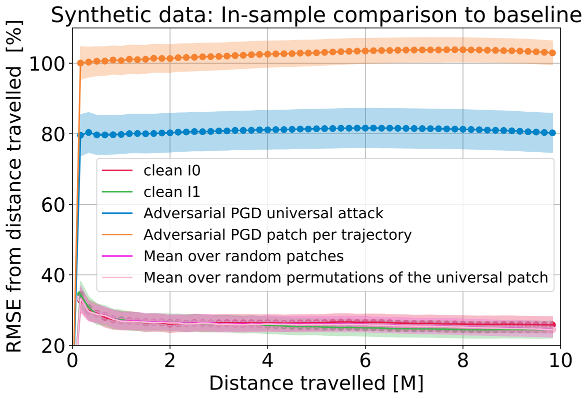

Figure 1. Accumulated deviation in distance travelled from the ground-truth trajectories of the synthetic dataset as a function of the trajectory length. We show a comparison of our best performing PGD and universal attacks to the clean and random perturbation baselines.

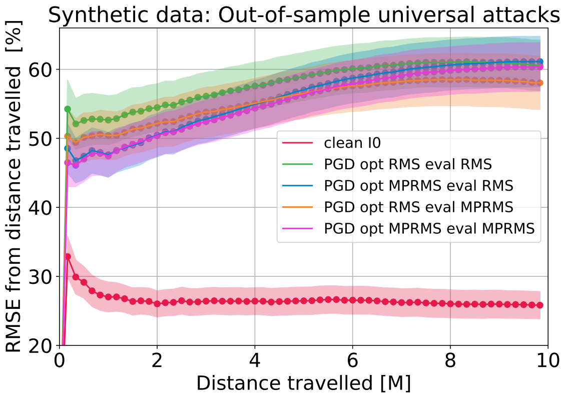

Figure 2. Accumulated deviation in distance travelled from the ground-truth trajectories over out-of-sample cross-validation of the synthetic dataset as a function of the trajectory length. We show a comparison of the deviation in distance travelled between our universal attacks and the clean baseline.

of previously produced adversarial patches to a closed-loop scheme, in which the outputs of the VO model are used in a simple navigation scheme. For each experiment we report the mean and standard deviation of ℓ_{VO} between the estimated and ground truth motions over the test trajectories, compared to the length of the trajectory. In all cases, we optimize the attacks for $k = 100$ iterations.

Data Generation The renderer framework used for the syntetic data is Blender [1], a $3D$ modeling and rendering package. Blender enables photo-realistic rendered images to be produced from a given $3D$ scene along with the ground truth motions of the cameras. In addition, we produce high quality, occlusion-aware masks, which are then used to compute the homography transformation H. We produced the trajectories in an urban $3D$ scene, as in such a scenario, GPS reception and accuracy is poor, and autonomous systems rely more heavily on visual odometry for navigation purposes. The patch is then positioned on a square plane at the side of one of the buildings, in a manner that resembles a large advertising board.

In the real data scenario, we situated a DJI Tello drone inside an indoor arena, surrounded by an Optitrack motion capture system for recording the ground truth motions. The patch was positioned on a planar screen at the arena boundary. To compute the homography transformation H, we designated the patch location in the scene by four Aruco markers.

3.1. Experimental results

In Fig. 1 we show the in-sample results on the synthetic dataset. Both our universal and PGD attacks showed a substantial increase in the generated deviation over the clean and random baselines. The best PGD attack generated, after $10[m]$, a deviation of 103% in distance travelled. For the same configuration, the best universal attack generated a deviation of 80% in distance travelled. Moreover, the clean I^1 and random baselines show a slight decrease in the generated deviation over the clean I^0 results, including the random permutations of the best universal patch. This suggests that the VO model is affected by the structure of the adversarial patch rather than simply by the color scheme.

In Fig. 2 we show the out-of-sample results on the synthetic dataset. Our universal attacks again showed a substantial increase in the generated deviation over the clean baseline, with the best universal attack generating, after $10[m]$, a deviation of 61% in distance travelled.

In Fig. 3 we show the closed-loop results on the synthetic dataset. Our universal attacks showed an increase in the generated deviation over the clean baseline, which, however, was not as substantial as before as the baseline's generated deviation is already quite significant. The best performing universal attack generated, after $45[m]$, a deviation of 71%, in distance travelled.

In Fig. 4 we show the in-sample results on the real dataset. Similarly to the synthetic dataset, we see a substantial improvement for both our PGD and universal attacks over the clean I^0 baseline, while the clean I^1 and random baselines show a slight decrease. The best PGD attack generated, after $1.56[m]$, a deviation of 34% in distance travelled. For the same configuration, the best universal attack generated a deviation of 22% in distance travelled.

In Fig. 5 we show the out-of-sample results on the real dataset. Our universal attacks again showed an increase in the generated deviation over the clean baseline, with the best universal attack generating, after $1.56[m]$, a deviation

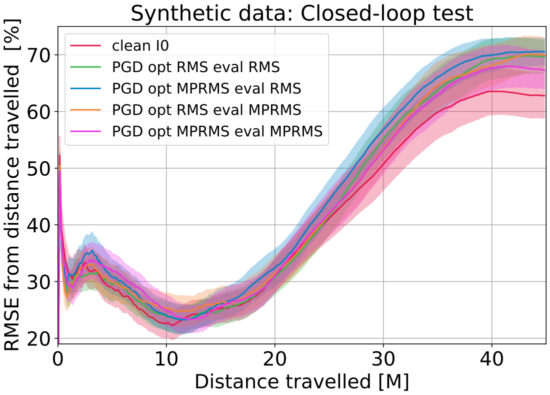

Figure 3. Accumulated deviation in distance travelled from the ground-truth over closed-loop trajectories of the synthetic dataset as a function of the trajectory length. We show a comparison of the deviation in distance travelled between our universal attacks and the clean baseline.

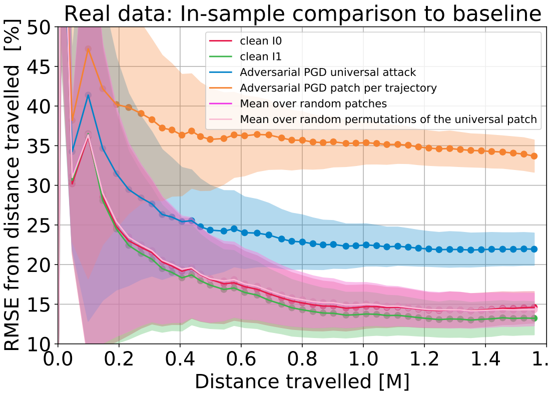

Figure 4. Accumulated deviation in distance travelled from the ground-truth trajectories on the real dataset as a function of the trajectory length. We show a comparison of our best performing PGD and universal attacks to the clean and random perturbation baselines.

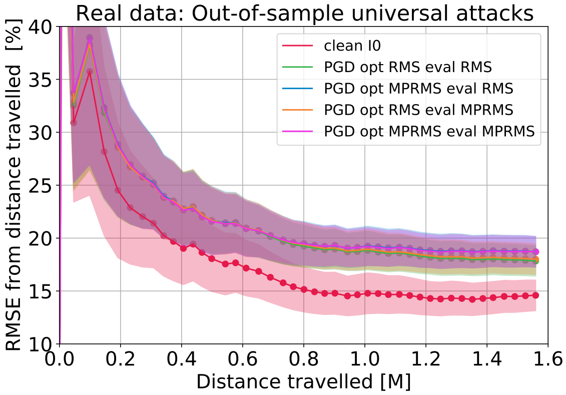

Figure 5. Accumulated deviation in distance travelled from ground-truth trajectories over out-of-sample cross-validation of the real dataset as a function of the trajectory length. We show a comparison of the deviation in distance travelled between our universal attacks and the clean baseline.

of 19% in distance travelled.

4. Conclusions

This paper proposed a novel method for passive patch adversarial attacks on visual odometry-based navigation systems. We used homography of the adversarial patch to different viewpoints to understand how each perceives it and optimize the patch for entire trajectories. Furthermore, we limited the adversarial patch in the ℓ_{inf} and ℓ_0 norms by taking into account the black and white albedo images of the patch and the FOV of the patch.

On the synthetic dataset, we showed that the proposed method could effectively force a given trajectory or set of trajectories to deviate from their original path. For a patch FOV of 22.2%, our PGD attack generated, on a given trajectory, an average deviation, after $10[m]$, of 103% in distance travelled, and given the entire trajectory dataset, our universal attack produced a single adversarial patch that generated an average deviation, after $10[m]$, of 80% in distance travelled. Moreover, our universal attack generated, on out-of-sample data, a deviation, after $10[m]$, of 61% in distance travelled and in a closed-loop setting generated an average deviation, after $45[m]$, of 71% in distance travelled.

In addition, while less substantial, our results were replicated using the real dataset and a significantly smaller patch FOV of 8.8%. Nevertheless, when considering the effect with a larger patch FOV, we can expect a corresponding increase in the generated deviation. For a given trajectory, our PGD attack generated an average deviation, after $1.56[m]$, of 34% in distance travelled. Given the entire dataset, our universal attack generated an average deviation, after $1.56[m]$, of 22% in distance travelled, and on out-of-sample data generated an average deviation, after $1.56[m]$, of 19% in distance travelled.

We conclude that physical passive patch adversarial attacks on vision-based navigation systems could be used to harm systems in both simulated and real-world scenes. Furthermore, such attacks represents a severe security risk as they could potentially push an autonomous system onto a collision course with some object by simply inserting a pre-optimized patch into a scene.

References

[1] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 3, 7

[2] Yao Deng, Xi Zheng, Tianyi Zhang, Chen Chen, Guannan Lou, and Miryung Kim. An analysis of adversarial attacks and defenses on autonomous driving models. In *2020 IEEE international conference on pervasive computing and communications (PerCom)*, pages 1–10. IEEE, 2020. 1

[3] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1

[4] Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 2755–2764, 2017. 1

[5] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2, 6

[6] Gautam Raj Mode and Khaza Anuarul Hoque. Adversarial examples in deep learning for multivariate time series regression. In *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–10. IEEE, 2020. 1

[7] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017. 1

[8] Andre T Nguyen and Edward Raff. Adversarial attacks, regression, and numerical stability regularization. *arXiv preprint arXiv:1812.02885*, 2018. 1

[9] Hector Perez-Leon, Jose Joaquin Acevedo, Jose A Millan-Romera, Alejandro Castillejo-Calle, Ivan Maza, and Anibal Ollero. An aerial robot path follower based on the 'carrot chasing'algorithm. In *Iberian Robotics conference*, pages 37–47. Springer, 2019. 7

[10] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[11] Wenshan Wang, Yaoyu Hu, and Sebastian Scherer. Tartanvo: A generalizable learning-based vo. *arXiv preprint arXiv:2011.00359*, 2020. 7

[12] Koichiro Yamanaka, Ryutaroh Matsumoto, Keita Takahashi, and Toshiaki Fujii. Adversarial patch attacks on monocular depth estimation networks. *IEEE Access*, 8:179094–179104, 2020. 1

[13] Chaoning Zhang, Philipp Benz, Chenguo Lin, Adil Karjauv, Jing Wu, and In So Kweon. A survey on universal adversarial attack. *arXiv preprint arXiv:2103.01498*, 2021. 1