EFFICIENT UNIVERSAL SHUFFLE ATTACK FOR VISUAL OBJECT TRACKING

Siao Liu[1], *Zhaoyu Chen*[1], *Wei Li*[1], *Jiwei Zhu*[1], *Jiafeng Wang*[2], *Wenqiang Zhang*[1,2,*], *Zhongxue Gan*[1,3,*]

[1]Academy for Engineering and Technology, Fudan University, Shanghai, China
[2]School of Computer Science, Fudan University, Shanghai, China
[3]The Department of Engineering Research Center for Intelligent Robotics, Ji Hua Laboratory, Foshan, China

ABSTRACT

Recently, adversarial attacks have been applied in visual object tracking to deceive deep trackers by injecting imperceptible perturbations into video frames. However, previous work only generates the video-specific perturbations, which restricts its application scenarios. In addition, existing attacks are difficult to implement in reality due to the real-time of tracking and the re-initialization mechanism. To address these issues, we propose an offline universal adversarial attack called Efficient Universal Shuffle Attack. It takes only one perturbation to cause the tracker malfunction on all videos. To improve the computational efficiency and attack performance, we propose a greedy gradient strategy and a triple loss to efficiently capture and attack model-specific feature representations through the gradients. Experimental results show that EUSA can significantly reduce the performance of state-of-the-art trackers on OTB2015 and VOT2018.

Index Terms— Adversarial examples, Offline attack, Visual object tracking, Universal adversarial perturbation

1. INTRODUCTION

Visual object tracking (VOT) is one of the fundamental vision tasks, which tracks a given object in each frame. It is the key building block in downstream vision applications, such as video surveillance and automatic driving. Thanks to the deep neural networks (DNNs), visual object tracking has achieved great progress. At present, the most representative algorithm is based on Siamese networks, which provides state-of-the-art performance and real-time inference, such as SiamFC [1], SiamRPN [2], SiamRPN++ [3] and SiamMask [4].

DNNs have been shown to be susceptible to adversarial examples. When an adversary introduces a delicate imperceptible perturbation to inputs, it would misguide the networks to produce incorrect results [5, 6]. Previous adversarial attacks on visual object tracking [7, 8, 9, 10, 11] have revealed the vulnerability of VOT. Online attacks, such as Cooling-Shrinking Attack (CSA) [8] and Hijacking [9], generate the frame-specific perturbations online as the search region changes. Another attack is the offline attack, such as One-shot Attack (OA) [7]. The perturbation is calculated in advance and directly added to the videos. It is worth noting that online attacks are difficult to implement in reality, as the object tracking happens in real time and has little time to generate specific perturbations for each frame.

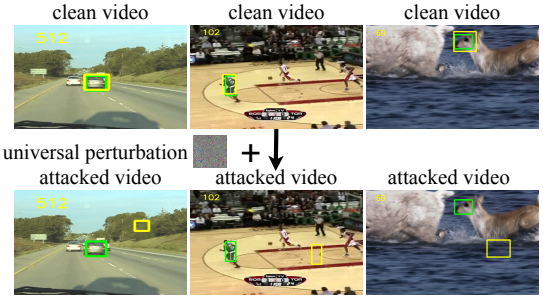

Fig. 1. Visualization of the EUSA against visual object tracking. The green boxes represent the ground truth and the yellow boxes are the tracking result of the tracker.

Therefore, we focus on the offline attack for visual object tracking in this paper. The offline attack is first presented in One-shot Attack [7] and it pre-generates perturbations on template images to fool the trackers. However, it ignores the re-initialization mechanism in tracking. When the tracker loses objects, it will reset a new template image as the target object, which leads to the One-shot Attack ineffective. The perturbation generated by One-shot Attack is target-specific, so it has low generalization and can hardly take effect on other videos, which restrains its application scenarios.

To solve these issues, we propose a universal adversarial attack called Efficient Universal Shuffle Attack (EUSA). Specifically, we sample the dataset and only use a small part of videos to efficiently generate a video-agnostic and template-insensitive perturbation. To improve the attack performance, we design a triple loss for Siamese trackers from the perspectives of feature, confidence and shape. Moreover, we propose a greedy-gradient strategy to improve the

This work was supported in part by Ji Hua Laboratory (project ID X190021TB190), in part by Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103), in part by Science and Technology Commission of Shanghai Municipality (No.19511132000), in part by the Shanghai Engineering Research Center of AI and Robotics, and the Engineering Research Center of AI and Robotics, Ministry of Education, China.

sampling process. Greedy-gradient strategy captures model-specific feature representations through the gradients and selects the vulnerable videos. Fig. 1 shows the effect of our proposed EUSA on OTB2015 [12]. Experiments show that EUSA effectively reduces the performance of state-of-the-art Siamese trackers on OTB2015 [12] and VOT2018 [13]. Our major contributions can be summarized as follows:

- We propose a novel universal adversarial perturbation generation method called Efficient Universal Shuffle Attack (EUSA). It takes only one perturbation to cause the tracker malfunction on all videos.
- To improve the computational efficiency and attack performance, we propose a greedy-gradient strategy and a triple loss to effectively capture and attack model-specific feature representations through the gradients.
- Experimental results show that EUSA can efficiently and significantly reduce the performance of state-of-the-art Siamese trackers on various benchmarks.

2. RELATED WORK

2.1. Visual Object Tracking

Recent work [1, 2, 3, 4] based on Siamese networks has achieved excellent performance and real-time inference. SiamFC [1] first constructs a fully convolutional Siamese network to train a tracker. SiamRPN [2] introduces Region Proposal Network into tracking. Then SiamRPN++ [3] proposes a multi-layer aggregation module and a depthwise correlation layer to achieve promising results with deeper networks. Moreover, SiamMask [4] enables Siamese trackers to conduct class-agnostic segmentation masks of the target object and improve the tracking performance.

2.2. Adversarial Attacks for Visual Object Tracking

Adversarial attacks on VOT [7, 8, 9, 10, 11] can be divided into online and offline attacks, depending on whether perturbations are generated in real-time tracking or not. Both CSA [8] and Hijacking [9] are online attacks. The former uses U-Net [14] to generate perturbations, while the latter directly uses gradient iterative calculation. However, these perturbations can not be generated in real-time tracking. One-shot Attack (OA), the only offline attack known by us, ignores the re-initialization mechanism in tracking and only generates target-specific perturbations, which limits its attack performance. To solve these issues, we propose the Efficient Universal Shuffle Attack to achieve offline universal attack.

3. METHODOLOGY

In this section, we first give the problem definition of universal attack on Siamese trackers and then we elaborate our Efficient Universal Shuffle Attack in detail.

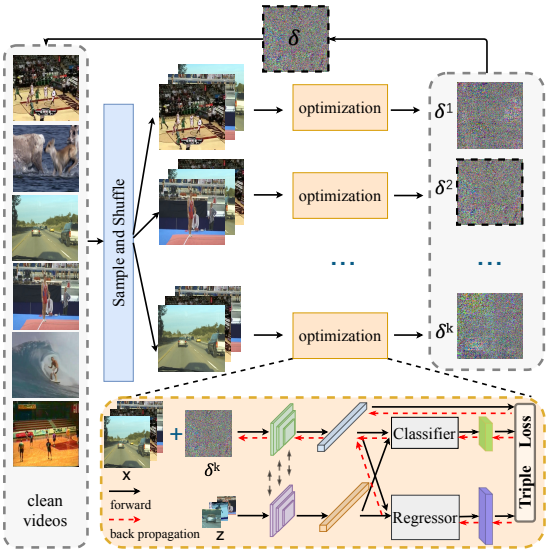

Fig. 2. The overview of Efficient Universal Shuffle Attack.

3.1. Problem Definition

Given an unknown target template z, Siamese trackers need to predict the location and shape of the target in the subsequent frames. We decompose the Siamese tracker into three parts: feature encoder $\mathcal{F}$, classifier $\mathcal{C}$ and bounding box regressor $\mathcal{R}$. The Siamese tracker shares the feature encoder $\mathcal{F}$ and conducts the similarity map between the template image z and the search region x. Using the similarity map, classifier $\mathcal{C}$ obtains the confidence of candidate boxes to categorize the foreground and background. Finally the bounding box regressor $\mathcal{R}$ conducts the location offset R_{loc} and the shape offset R_{shape} to adjust the location and shape of candidate boxes.

Considering the re-initialization mechanism, we aim to generate a universal perturbation δ on search regions to mislead a tracker to lose targets in all videos. Specifically, we describe the universal adversarial attack as follows:

$$\max \sum_{x \in \mathcal{X}} \mathcal{L}(x, x+\delta), \quad s.t. \quad ||\delta||_\infty \leq \epsilon, \tag{1}$$

where $\mathcal{X}$ are search regions from various videos and we limit the adversarial perturbation into the range $[-\epsilon, \epsilon]$, ensuring the perturbation is imperceptible to human eyes. For simplicity, we define x^* as $x+\delta$, and R^*_{loc} and R^*_{shape} are the corresponding location offset and shape offset respectively.

3.2. Triple Loss

To attack model-specific feature representations and improve attack performance, we design a triple loss from different perspectives, which combines feature-deflect loss $\mathcal{L}_f$, confidence

loss $\mathcal{L}_c$ and drift loss $\mathcal{L}_d$. Supposing that λ_1 and λ_2 are the balanced weights, the triple loss $\mathcal{L}$ can be expressed as:

$$\mathcal{L} = \mathcal{L}_f + \lambda_1 \mathcal{L}_c + \lambda_2 \mathcal{L}_d. \tag{2}$$

To attack the feature encoder and confuse similarity maps, we distort the embedding of the search region $\mathcal{F}(x)$ in feature space. We adopt the cosine similarity to measure the deflection of $\mathcal{F}_{i=1:C}(x)$, where C is the channel of feature maps and set a margin $\mathrm{m_f}$ to control the deflection. Feature-deflect loss $\mathcal{L}_f$ can be written as:

$$\mathcal{L}_f(x, x^*) = -\sum_{i=1:C} \max(\mathrm{m_f}, \cos(\mathcal{F}_i(x), \mathcal{F}_i(x^*))). \tag{3}$$

Trackers always use Gaussian windows to constrain prediction results so they do not move too far within two adjacent frames, thereby providing some protection against attacks by improving background confidence. To address this issue, we suppress the confidence of all N candidates and the confidence loss $\mathcal{L}_c$ is defined as follows:

$$\mathcal{L}_c(z, x^*) = -\sum_{j=1:N} C_j(\mathcal{F}(z), \mathcal{F}(x^*)). \tag{4}$$

.

As the attack on the confidence-wise and feature-wise are undirected, we propose a drift loss $\mathcal{L}_d$ to force the predict bounding box along the given direction and shape overtimes. Specifically, we shrink the scale factors R^*_{shape} to 0 and make the location factors R^*_{loc} closer to the given direction $\vec{d}$, which is measured by Euclidean distance. In addition, we set constant $\alpha = 0.6$ to balance the shrinking attack and offsetting attack. Drift loss $\mathcal{L}_d$ can be expressed as:

$$\mathcal{L}_d(z, x^*) = -\alpha \cdot ||R^*_{scale}||_2 - || < R^*_{loc}, \vec{d} > ||_2. \tag{5}$$

3.3. Efficient Universal Shuffle Attack

UAP [15] introduces high-dimensional decision boundaries to explain the existence of universal perturbation, which can be regraded as model-specific feature representations. We use data sampling to accelerate the capture of model-specific features representations. Considering that the magnitude of the gradient can reflect the videos' sensitivity to adversarial examples, we propose a greedy-gradient strategy as sampling approach. For each video $\mathbf{v}$, we could obtain the target template z and the first search region x_1 through the initial frame. The required gradients can be computed efficiently by back propagation via Eq. 2. Then, we choose the videos with larger absolute value of gradient to construct training set $\mathbf{X} \in \mathcal{X}$. The size of the training set depends on the sampling rate r.

Fig. 2 shows the pipeline of EUSA and the attack procedure in Algorithm 1. First, we sample the video dataset with greedy-gradient strategy and collect the victim video set $\mathbf{X}$. To search a better universal adversarial perturbation, we randomly shuffle the video set k times to obtain the training set $\{\mathbf{X}^{1:k}\}$ and initialize k perturbations$\{\delta^\tau\}_{\tau=1:k}$ as candidate perturbations. Since all candidate perturbations are optimized in the same way, we only discuss one perturbation δ^τ below. Second, we randomly select one frame from each video in the shuffled training set $\mathbf{X}^\tau$ for iterative optimization. We follow the Project Gradient Descent [16] to optimize δ^τ with Eq. 6. The loss is calculated according to Eq. 2. As we can easily access the gradient of δ^τ, we optimize the perturbation δ^τ as follows:

$$\delta^\tau = \mathrm{clip}\left(\delta^\tau + \alpha \cdot \mathbf{sign}(\nabla_{\delta^\tau} L), -\epsilon, \epsilon\right) \tag{6}$$

where α is the step size and the clip operation satisfies the constraint $||\delta^\tau||_\infty \leq \epsilon$ in each step. Then we sort all candidate perturbations by the loss L^τ which is the sum of triple loss $L^{\mathbf{v}}$ from selected videos. Finally, we choose the perturbation with the maximal loss L^τ as the final perturbation.

Algorithm 1 Efficient Universal Shuffle Attack (EUSA)

Input: dataset $\mathcal{X}$, sampling rate r, victim tracker $\mathcal{T}$, number of candidate perturbations k
Output: video-agnostic perturbation δ

$\mathbf{X} \leftarrow$ sampling $(\mathcal{X}, r)$ with **greedy-gradient strategy**
$\mathbf{X}^{1:k} \leftarrow$ shuffle training set $\mathbf{X}$ for k times
Initialize the $L^{best} \leftarrow 0, \tau \leftarrow 0$
while number of candidates $\tau++ < k$ **do**
  Initialize τth perturbation δ^τ with 0
  Initialize the τth Loss $L^\tau \leftarrow 0$
  for $\mathbf{v}$ in $\mathbf{X}^\tau$ **do**
    Initialize attacked tracker $\mathcal{T}$ using template z
    Randomly select a frame I
    Obtaining the search region s according I and $\mathcal{T}$
    Calculate the **triple loss** $L^{\mathbf{v}}$ using Eq. 2
    $\delta^\tau \leftarrow \mathrm{clip}\left(\delta^\tau + \alpha \cdot \mathbf{sign}(\nabla_{\delta^\tau} L^{\mathbf{v}}), -\epsilon, \epsilon\right)$
    $L^\tau \leftarrow L^\tau + L^{\mathbf{v}}$
  end for
end while
$\delta \leftarrow \delta^w \quad$ s.t. $w = \arg\max\{L^1, L^2, \ldots, L^k\}$
return δ

4. EXPERIMENTS

We evaluate EUSA against state-of-the-art Siamese trackers on OTB2015 and VOT2018. Ablation studies show the effectiveness of greedy-gradient strategy and triple loss.

4.1. Implementation Details

We measure the performance of EUSA on two standard benchmarks, OTB2015 [12] and VOT2018 [13]. OTB2015 [12] contains 100 videos and evaluates the trackers with precision and success rate. VOT2018 [13] includes 60 videos and ranks the performance of trackers with the expected average overlap (EAO) rule. Notably, the tackers would re-initialize

Table 1. Attack performance on OTB100.

Tracker	Precision(%) ↑			Success Rate(%) ↑		
	Org	OA	EUSA	Org	OA	EUSA
SiamRPN	87.6	27.8	**26.7**	66.8	20.4	**20.2**
SiamRPN++(R)	90.5	35.7	**32.7**	69.6	26.2	**23.6**
SiamRPN++(M)	86.4	35.3	**25.9**	65.8	26.1	**18.3**
SiamMask	83.9	65.0	**34.9**	64.7	48.1	**22.5**

Table 2. Attack performance on VOT2018.

Tracker	Accuracy(%) ↑			Robustness ↓			EAO ↑		
	Org	OA	EUSA	Org	OA	EUSA	Org	OA	EUSA
SiamRPN	57.7	46.7	**44.0**	0.309	1.733	**2.241**	0.338	0.082	**0.055**
SiamRPN++(R)	60.2	51.9	**46.1**	0.243	1.157	**2.051**	0.413	0.115	**0.072**
SiamRPN++(M)	58.9	48.3	**45.2**	0.234	1.344	**2.622**	0.411	0.101	**0.056**
SiamMask	59.8	45.5	**31.8**	0.248	0.674	**2.632**	0.406	0.165	**0.043**

Table 3. Ablation study of triple loss.

$\mathcal{L}_f$		✓			✓	✓	✓
$\mathcal{L}_c$			✓		✓		✓
$\mathcal{L}_d$				✓		✓	✓
Precision(%)	90.5	59.4	54.1	54.6	51.2	50.0	**32.7**
Success rate(%)	69.6	40.0	38.4	39.0	37.1	36.2	**23.6**

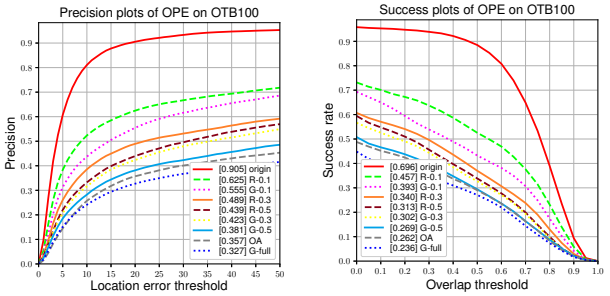

Fig. 3. Quantitative comparisons between various sampling rate and different sampling strategy on OTB2015 dataset. The suffix "G" and "R" are greedy-gradient strategy and random sample respectively. The numbers are sampling rates.

the template image once it loses the target in VOT2018. The victim trackers include SiamRPN [2], SiamRPN++ [3] and SiamMask [4]. SiamRPN++(R) represents that the SiamRPN++ applies ResNet-50 [17] as the backbone and the SiamRPN++(M) uses MobileNet-v2 [18] to extract features. We implement experiments with RTX 1080Ti. The number of candidate perturbations k is set to 50. For each iterative step, we set step size $\alpha = 0.9$ and the maximum pixel of perturbation $\epsilon = 16$. To balance the components of triple loss, we set the hyper-parameters $\lambda_1 = 0.9, \lambda_2 = 0.7$. In all experiments, we report the average result over 5 times.

4.2. Attacks on OTB2015 and VOT2018

Results on OTB2015. As shown in Table 1, EUSA successfully drops the success rates of SiamRPN, SiamRPN++(R), SiamRPN++(M) and SiamMask to 20.2%, 23.6%, 18.3% and 22.5%. Besides, EUSA has the best performance on the SiamRPN++(M), which reduces the precision and success rate by 60.5% and 47.5% respectively. Compared with SiamRPN++(R), SiamRPN++(M) is more vulnerable to our universal adversarial perturbation, which can be attributed to the vulnerability of its backbones.

Results on VOT2018. Table 2 shows that EUSA performs much better than OA in VOT2018, which reveals that One-shot Attack is extremely template-sensitive. Once the trackers re-initialize the template, OA fails to attack the subsequent frames. However, slight distortion of template image almost has no impact on our attack, which can be contributed to the randomly shuffle. In addition, EUSA trains the perturbation by selecting the target template and search region from different frames, which makes the re-initialization mechanism ineffective against our attack.

4.3. Ablation Study

We conduct a series of experiments on OTB2015 to explore the effectiveness of each component in our attack. We use the SiamRPN++(R) as the victim tracker.

Fig. 3 illustrates a quantitative analysis of the performance of EUSA with different sampling strategies and various sampling rates r, including 0.1, 0.3, 0.5, and 1. Our proposed EUSA with any sampling rate can significantly decline the performance of SiamRPN++(R), which indicates that our attack still performs well on the videos unseen in the training process. When the sampling rate is 0.1, the perturbation generated by only 10 frames can drop the tracker precision by 35%. Compared with random sampling, our greedy-gradient strategy can significantly improve attack performance by at least 5.8% on precision and 2.3% on success rate.

Moreover, we evaluate the performance of each component of the triple loss. The results are shown in Table 3. We observe that only using the confidence loss would outperform the other two, which reduces the precision and success rate to 54.1% and 38.4% respectively. This shows that $\mathcal{L}_c$ successfully damages the Gaussian window and disables the Region Proposal Network. Besides, there is no performance degradation when simultaneously using any two components of triple loss which validates the effectiveness of our triple loss.

5. CONCLUSIONS

In this work, we propose an offline universal attack, Efficient Universal Shuffle Attack (EUSA), which injects only one perturbation to cause the tracker malfunction on all videos. To further improve the efficiency and performance of EUSA, we design a greedy-gradient strategy and a triple loss to capture and attack the model-specific feature representations. Numerous experiments show that EUSA can significantly reduce the performance of Siamese trackers on various benchmarks.

6. REFERENCES

[1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr, "Fully-convolutional siamese networks for object tracking," in *European conference on computer vision*. Springer, 2016, pp. 850–865.

[2] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu, "High performance visual tracking with siamese region proposal network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8971–8980.

[3] B Li, W Wu, Q Wang, F Zhang, J Xing, and J SiamRPN+ Yan, "Evolution of siamese visual tracking with very deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA*, 2019, pp. 16–20.

[4] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr, "Fast online object tracking and segmentation: A unifying approach," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1328–1338.

[5] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun, Eds., 2014.

[6] Hao Huang, Yongtao Wang, Zhaoyu Chen, Zhi Tang, Wenqiang Zhang, and Kai-Kuang Ma, "Rpattack: Refined patch attack on general object detectors," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 2021, pp. 1–6.

[7] Xuesong Chen, Xiyu Yan, Feng Zheng, Yong Jiang, Shu-Tao Xia, Yong Zhao, and Rongrong Ji, "One-shot adversarial attacks on visual tracking with dual attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10176–10185.

[8] Bin Yan, Dong Wang, Huchuan Lu, and Xiaoyun Yang, "Cooling-shrinking attack: Blinding the tracker with imperceptible noises," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 990–999.

[9] Xiyu Yan, Xuesong Chen, Yong Jiang, Shu-Tao Xia, Yong Zhao, and Feng Zheng, "Hijacking tracker: A powerful adversarial attack on visual tracking," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2897–2901.

[10] Qing Guo, Xiaofei Xie, Felix Juefei-Xu, Lei Ma, Zhongguo Li, Wanli Xue, Wei Feng, and Yang Liu, "Spark: Spatial-aware online incremental attack against visual tracking," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*. Springer, 2020, pp. 202–219.

[11] Siyuan Liang, Xingxing Wei, Siyuan Yao, and Xiaochun Cao, "Efficient adversarial attacks for visual object tracking," in *European Conference on Computer Vision*. Springer, 2020, pp. 34–50.

[12] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang, "Online object tracking: A benchmark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2411–2418.

[13] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka ˇCehovin Zajc, Tomas Vojir, Goutam Bhat, Alan Lukezic, Abdelrahman Eldesokey, et al., "The sixth visual object tracking vot2018 challenge results," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[15] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard, "Universal adversarial perturbations," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 2017, pp. 86–94, IEEE Computer Society.

[16] 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.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.