Structured Universal Adversarial Attacks on Object Detection for Video Sequences

Sven Jacob[1,2], Weijia Shao[1], and Gjergji Kasneci[2,3]

[1] Federal Institute for Occupational Safety and Health (BAuA), Dresden, Germany
{jacob.sven,shao.weijia}@baua.bund.de
[2] School of Computation, Information and Technology, Technical University of Munich, Munich, Germany
[3] School of Social Sciences and Technology, Technical University of Munich, Munich, Germany

Abstract. Video-based object detection plays a vital role in safety-critical applications. While deep learning-based object detectors have achieved impressive performance, they remain vulnerable to adversarial attacks, particularly those involving universal perturbations. In this work, we propose a minimally distorted universal adversarial attack tailored for video object detection, which leverages nuclear norm regularization to promote structured perturbations concentrated in the background. To optimize this formulation efficiently, we employ an adaptive, optimistic exponentiated gradient method that enhances both scalability and convergence. Our results demonstrate that the proposed attack outperforms both low-rank projected gradient descent and Frank-Wolfe-based attacks in effectiveness while maintaining high stealthiness. All code and data are publicly available at https://github.com/jsve96/AO-Exp-Attack.

Keywords: Adversarial Attacks · Object Detection · AI Safety · Robustness

1 Introduction

Video-based object detection plays an increasingly important role in safety monitoring systems for machine and occupational environments, enabling the localization of human workers, tools, and obstacles to identify potential hazards before they escalate into accidents [1, 8, 24]. As a core task in computer vision, object detection involves identifying and localizing semantic objects within images or videos. Recent advances in deep learning have significantly improved object detection performance, enabling its deployment in a range of safety-critical domains, ranging from autonomous driving [7, 12], real-time surveillance [2, 19], and industrial applications [27, 41]. In these contexts, object detection not only contributes to operational efficiency but also serves as a first line of defense in preventing unsafe interactions between humans and machines.

Most state-of-the-art object detection methods rely on Deep Learning (DL) techniques [40]. Despite their substantial advancements over the past decade,

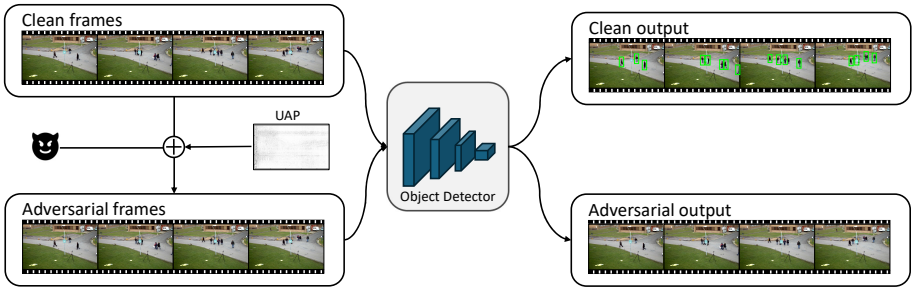


Fig. 1: Shows conceptual framework of universal adversarial attack on object detector. A carefully crafted Universal Adversarial Perturbation (UAP) suppresses all bounding boxes after applied on clean frames.

DL models are vulnerable to adversarial attacks (AT) [15, 35], which craft perturbations to the inputs to mislead the model into making incorrect predictions. While research on adversarial attacks in image classification has been extensively studied [10], such attacks on object detection systems, especially in the context of video data, have received considerably less attention [26]. At first glance, adversarial attacks on video object detection may appear straightforward, as they seem to require applying existing techniques for attacking static object detectors to each frame of the video clip [18, 36, 37]. In [23], the authors have empirically proved the existence of universal adversarial perturbations against all frames, which cause object detectors to fail on most of the frames [23, 38]. Effective universal attacks pose a more significant threat as they are transferable across frames without further accessing the target model, and are more convenient to be applied in the real physical world [23, 26].

Although prior studies have aimed to generate adversarial examples posing greater threats, they have predominantly focused on perturbations bounded by ℓ_2 and ℓ_∞ norms [26]. In image-based attacks, ℓ_1-bounded perturbations can be especially threatening due to their ability to introduce sparse yet imperceptible changes [9]. However, in video-based settings, the direct application of ℓ_1 attacks often results in visible patches on moving objects across frames. This not only reduces the sparsity but also challenges stealthiness in dynamic scenes. As adversarial defense mechanisms continue to evolve, identifying a broader range of attack strategies is essential for robust evaluation and understanding of their limitations.

Building on robust principal component analysis for segmentation [4] and structured adversarial perturbation methods for image classification [20], this work introduces a novel strategy that leverages structured but non-suspicious background modifications for object vanishing attacks. The conceptual framework is illustread in Figure 1. To this end, we propose a minimally distorted attack method based on nuclear norm regularization. Figure 2 provides a preliminary visual comparison illustrating the difference between ℓ_1 attacks, which tend to produce sparse patches on moving objects, and our nuclear norm-based

attack, which generates more structured perturbations primarily in the background. While nuclear norm regularization provides a powerful tool for promoting low-rank structure in background perturbations, it poses significant optimization challenges. To address this, we employ optimistic exponentiated gradient descent [31], which enables efficient and scalable optimization under nuclear norm regularization. We evaluate our proposed object vanishing attack on public video datasets and video object detection models. The results demonstrate that our method effectively generates structured background perturbations that consistently remove the bounding boxes predicted by these models. Compared to existing nuclear norm-based attack approaches, our method achieves superior attack success while being significantly more computationally efficient.

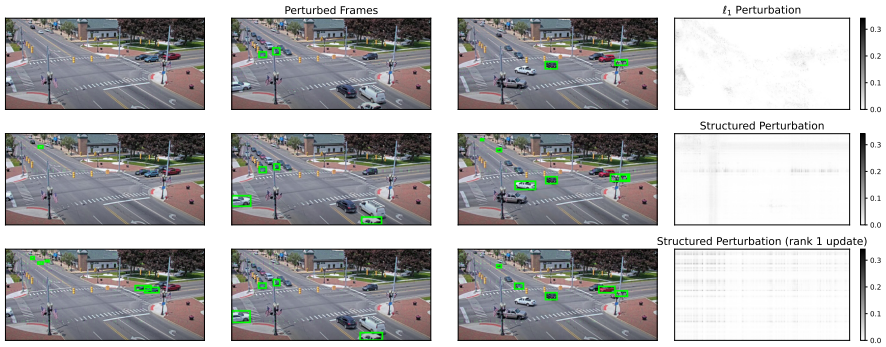


Fig. 2: The ℓ_1 attack introduces flickering noise that trails moving objects and spreads across the street in subsequent frames, whereas the nuclear norm-based attack perturbs orthogonal spatial patterns of the video frames, resulting in more structured and spatially coherent perturbations.

Our main contributions are summarized as follows:

- We introduce a minimally distorted universal attack formulation based on nuclear norm regularization, which promotes structured perturbations of the orthogonal spatial patterns across video frames (Equation (1)).
- To efficiently solve the associated optimization problem, we adapt an adaptive optimistic exponentiated gradient descent method, enabling scalable optimization under nuclear norm constraints (Algorithm 1).
- We conduct comprehensive evaluations on public video datasets and a state-of-the-art video object detection model, demonstrating that our method consistently suppresses bounding boxes through subtle background changes (Figure 6, Figure 7a).
- Our method achieves superior attack success and computational efficiency compared to existing nuclear norm-based adversarial attacks (Table 2).

The rest of the paper is organized as follows. Section 2 reviews related work, and Section 3 introduces the notation used throughout the paper. In Section 4,

we present our attack algorithm, which is evaluated on real-world video object detection datasets and the popular object detection model Mask-RCNN in Section 5. Finally, Section 6 concludes the paper by discussing the limitations of our approach and outlining future research directions.

2 Related Work

While the literature on adversarial attacks in image classification is extensive, research on adversarial attacks targeting object detection, especially in video settings, remains relatively limited. We refer to recent surveys [3,26] for an overview of existing techniques and challenges in this area. This work focuses on object vanishing attacks, which aim to make the model fail to detect certain objects in the input frames. Several works have explored object vanishing attacks on images by manipulating bounding box outputs using adversarial perturbations, either constrained within ℓ_2 or ℓ_∞ norm balls or applied as localized patches on foreground objects [18, 23, 36, 37, 38, 39].

The methods listed above can be directly applied to each frame of a video independently. However, applying attacks separately to each frame ignores the temporal coherence of video data, often resulting in flickering perturbations across frames that reduce stealth. Universal adversarial perturbations [25], which are input-agnostic and applied uniformly across inputs, offer a more natural fit for object detection in video settings [23, 38]. Yet, the existing universal attacks [23,38] apply noise uniformly over the entire image without leveraging the spatial structure unique to videos.

The investigation of low-rank structures for high-dimensional problems has a long history following the manifold hypothesis, which states that high-dimensional data tend to live near a lower-dimensional manifold [11]. This assumption led to many dimensionality reduction methods in general but was also utilized in crafting adversarial perturbations using Autoencoder [13, 33], PCA [21, 22], or UMAP [34]. More recently, researchers have also focused on low-rank representation for adversarial attacks induced by nuclear norm regularization [20]. Moreover, a combination of nuclear norm regularization and optimizing adversarial perturbation using projected gradient descent (PGD) was introduced in [30]. In [29], a group-wise sparse attack is generated, which only perturbs a few semantically meaningful areas of an image.

3 Notation

Throughout the paper, we write $x \in [0,1]^{H\times W\times C}$ for an image with height H, width W, and C channels. Suppose an object detection model $f : [0,1]^{H\times W\times C} \to \mathcal{S}$ that takes an input image and outputs a set $\mathcal{S}$. The minimal output set is a collection of bounding boxes and labels for each detected object within an image. The output depends on the choice of f, e.g., Mask R-CNN [17] additionally outputs confidence scores $\xi_i \in (0,1]$ and masks $m_i \in [0,1]^{H\times W}$ where i is the

index of the corresponding bounding box in $\mathcal{S}$. For vector-valued inputs $x \in \mathbb{R}^n$, the ℓ_p-norm is given by $||x||_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$ for $p \geq 1$.

Recall, for a matrix $A \in \mathbb{R}^{H \times W}$ its singular value decomposition (SVD) is $A = U\Sigma V^T$, where $U \in \mathbb{R}^{H \times H}$ and $V \in \mathbb{R}^{W \times W}$ are orthonormal matrices whose columns are the orthonormal basis for the column and row space of A, and $\Sigma \in \mathbb{R}^{H \times W}$ is a rectangular diagonal matrix storing the singular values $\sigma_i \geq 0$ for $i = 1, \ldots, r$ where $r = \min\{H, W\}$ is the rank of A. Define the function

$$\text{diag} : \mathbb{R}^r \to \mathbb{R}^{H \times W}, \sigma \mapsto \Sigma \text{ with } \Sigma_{ij} = \begin{cases} \sigma_i, & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

The SVD decomposition can be equivalently written as $A = U\,\text{diag}(\sigma)V^T$.

The Schatten p-norm of a matrix A is the ℓ_p-norm of the vector of its singular values $||A||_p = (\sum_{i=1}^{r} (\sigma_i)^p)^{\frac{1}{p}}$. The matrix norm for $p = 2$ is also called Frobenius norm, notated as $||A||_F$. The nuclear norm ($p = 1$) of a matrix A is the ℓ_1-norm of its singular values $||A||_* = \sum_{i=1}^{r} \sigma_i$.

For $p = \infty$, we have the spectral norm of a matrix, which is the largest singular value, $||A||_\infty = \max_{1 \leq i \leq r} \sigma_i$.

4 Adversarial Attack Formulation

The goal of an adversarial attack is to determine some minimal perturbation δ which maximizes some loss function $\mathcal{L} : \mathcal{S} \mapsto \mathbb{R}_+$ of the object detector f, when added to x. A non-targeted adversarial attack can be formulated as,

$$\begin{aligned} \min_{\delta \in \mathbb{R}^{H \times W \times C}} \quad & -\mathcal{L}(f(x+\delta), f(x)) + \lambda \mathcal{R}(\delta) \\ \text{s.t.} \quad & x + \delta \in [0,1]^{H \times W \times C} \end{aligned}$$

where $\mathcal{R}$ denotes some regularization of the perturbation and $\lambda > 0$ is the regularization parameter. In the following, we propose to split the loss function into background and foreground loss, $\mathcal{L} = \mathcal{L}_{\text{fg}} + \mathcal{L}_{\text{bg}}$. Consider the set of clean masks $\mathcal{M} = \{m_i | \xi_i > \tau; i \in [S]\}$ obtained with $f(x)$, with a confidence score above τ. We combine all clean masks in $\mathcal{M}$ to obtain a unified mask $\boldsymbol{m} = \sum_{i \in \mathcal{M}} m_i$ for all confident detections and derive a binary mask

$$\boldsymbol{y}_{ij} = \begin{cases} 1 & \text{if } \boldsymbol{m}_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$$

which we use as the ground-truth segmentation for the clean image. We separate the mask into foreground pixels $\mathcal{F}$ and background pixels $\mathcal{B}$ and calculate the average cross-entropy loss between $\boldsymbol{y}$ and the predicted masks which are the output of $f(x + \delta)$,

$$\mathcal{L}_{\text{fg}} = \frac{1}{|\mathcal{F}|} \sum_{i \in \mathcal{F}} \text{CE}(p_i, y_i), \quad \mathcal{L}_{\text{bg}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \text{CE}(p_i, y_i).$$

Additionally, we seek to guide the model into less confident predictions of bounding boxes, therefore we introduce the confidence loss

$$\mathcal{L}_{\text{conf}} = \sum_{i \in [S]} \xi_i \cdot \mathbf{1}_{(\xi_i > \tau)}$$

which penalizes any predictions above threshold τ - affecting the foreground. In summary, we use a combination of

$$\mathcal{L}_{\text{total}} = \underbrace{\alpha \mathcal{L}_{\text{fg}} + \gamma \mathcal{L}_{\text{conf}}}_{\text{Foreground}} + \underbrace{\beta \mathcal{L}_{\text{bg.}}}_{\text{Background}}$$

Usually, regularization is introduced to ensure a sparse perturbation. For vector-valued inputs, common choices are ℓ_p-norms or the Frobenius ℓ_1-norm, which suppresses the perturbation to have few non-zero values. The nuclear norm is related to the rank of the matrix where the minimization of it leads to sparsity in singular values, resulting in a low-rank matrix. Nuclear norm regularization is popular among image denoising [16] and low-rank matrix approximation. It has also shown promising results in domain generalization [32].

We introduce a combination of Frobenius norm and nuclear norm as the regularizer

$$\mathcal{R}(\delta) = \lambda_1 ||\delta||_* + \lambda_2 ||\delta||_F,$$

which balances sparsity and low-rank of the universal adversarial perturbation.

4.1 Universal Attack

The objective of this work is to generate an adversarial perturbation that is applied uniformly across all frames of a video and degrades the performance of a target model f. Let $\{x_b | 1 \leq b \leq B\}$ be the set of frames in a video clip, where B is the number of frames. We formalize the attack as the following regularized optimization problem

$$\min_{\delta \in \mathbb{R}^{H \times W \times C}} -\frac{1}{B} \sum_{b=1}^{B} \mathcal{L}_{\text{total}}(f(x_b + \delta), f(x_b)) + \sum_{c=1}^{C} (\lambda_1 ||\delta^{\mathbf{c}}||_* + \frac{\lambda_2}{2} ||\delta^{\mathbf{c}}||_F^2). \quad (1)$$

Denote the gradient information

$$\nabla \mathcal{G}(\delta^c) = \frac{1}{B} \sum_{b=1}^{B} \nabla_{\delta^c} \mathcal{L}_{\text{total}}(f(x_b + \delta^c), f(x_b)),$$

which is the loss gradient averaged over all frames concerning a perturbation channel δ^c. Given the access to $\nabla \mathcal{G}(\delta^c)$, we iteratively update each perturbation channel δ^c by applying the adaptive optimistic exponentiated method (AO-Exp) proposed in [31]. The algorithm is described in Algorithm 1.

At each iteration t, in addition to the perturbation δ_t^c, we maintain a decision variable in the form of factors of an SVD decomposition

$$\delta_t^c = U_{c,t} \operatorname{diag}(z_t^c) V_{c,t}^\top$$

for each channel, where $z_t^c \in \mathbb{R}_{\geq 0}^{\min\{W,H\}}$ is the vector of singular values and $U_{c,t} \in \mathbb{R}^{W \times W}$, $V_{c,t} \in \mathbb{R}^{H \times H}$ are the orthogonal bases. To obtain the intermediate perturbation, we apply the following procedure:

(1) Optimistic update: We first perform an optimistic update using the gradient information at iterations t and $t-1$

$$\begin{aligned} \eta_t^c &\leftarrow \eta_{t-1}^c + t^2 \|\nabla \mathcal{G}(\delta_t^c) - \nabla \mathcal{G}(\delta_{t-1}^c)\|_\infty^2 \\ \bar{z}_{t,i}^c &\leftarrow \log(z_{t,i}^c + 1) \text{ for all } i \in \{1, \ldots, \min\{W,H\}\} \\ U_{c,t+1} \operatorname{diag}(\theta_t^c) V_{c,t+1}^\top &\leftarrow \eta_t^c \cdot U_{c,t} \operatorname{diag}(\bar{z}_t^c) V_{c,t}^\top + (2t+1)\nabla \mathcal{G}(\delta_t^c) - t \nabla \mathcal{G}(\delta_{t-1}^c) \end{aligned} \tag{2}$$

to obtain the orthogonal bases of iteration $t+1$.

(2) Singular values of decision variable: Then, we find the singular values of the decision variable at iteration $t+1$ by calculating the principal branch of the Lambert function

$$z_{t+1,i}^c = \frac{\eta_t^c}{\lambda_2} W_0 \left(\frac{\lambda_2}{\eta_t^c} \exp\left(\frac{\lambda_2 + \max\{\theta_{t,i}^c - \lambda_1, 0\}}{\eta_t} \right) \right) - 1. \tag{3}$$

(3) Construct perturbation: Finally, we obtain the perturbation δ_{t+1}^c by taking the weighted average of $z_1^c, \ldots, z_{t+1}^c$. We propose to use the top k values of the decision variable z_t^c in the reconstruction of δ_{t+1}^c which further compresses the information in δ, thus promoting low-rank with. Let $z_{t,1:k}^c = (z_{t,1}^c, \ldots, z_{t,k}^c, 0, \ldots, 0)$, then the low-rank perturbation is obtained with

$$\delta_{t+1}^c = \frac{2}{t(t+1)} \sum_{s=1}^{t} s \cdot U_{c,t} \operatorname{diag}(z_{s,1:k}^c) V_{c,t}^\top. \tag{4}$$

The per-iteration complexity of the algorithms depends on the complexity of performing the SVD-decomposition and solving the principal branch of the Lambert function.

Algorithm 1 AO-Exp Update

Input: Gradient: $\boldsymbol{\nabla \mathcal{G}(\delta_t)}$, Regularization parameter: $\boldsymbol{\lambda_1, \lambda_2}$.

```
Initialize δ_0^c, η_0^c, U_{c,0}, V_{c,0}, z_0^c for all channels c = 1, ..., C
for t = 1 to T do
    for each channel c = 1 to C do
        (1) Optimistic update: Update U_{c,t+1}, V_{c,t+1}, and θ_t^c        ▷ eq. (2)
        (2) Singular value update: Compute z_{t+1,1:k}^c                    ▷ eq. (3)
        (3) Perturbation update: Compute δ_{t+1}^c                          ▷ eq. (4)
    end for
end for
```

Output: $\boldsymbol{\delta_{T+1}}$

5 Experiments & Evaluation

This section details the empirical evaluation of the proposed algorithms. We first describe the experimental setup, including the baseline attack, metrics, datasets, and models used within this experimental scope. We then present and analyze the results. All code and data used in our experiments are publicly available online[4].

5.1 Baseline Attack

LoRa-PGD: The low-rank PGD attack is a variation of the projected gradient descent (PGD) that directly searches for a low-rank structured perturbation [30]. The k-th iteration of a PGD attack is the following perturbation

$$\delta_k = \mathcal{P}\left(\delta_{k-1} + \epsilon \frac{\nabla_\delta \mathcal{L}}{||\nabla_\delta \mathcal{L}||}\right),$$

where $\mathcal{P}$ denotes the projection on some feasible set of perturbations. For the LoRa-PGD attack, the perturbation is decomposed into two lower-rank matrices $U \in \mathbb{R}^{H \times r \times C}$ and $V \in \mathbb{R}^{r \times W \times C}$ where $r \leq \min\{H, W\}$, such that each entry of the perturbation for each channel $c \in C$ is given by

$$\delta_{ijc} = (U \otimes V)_{ijc} = \sum_{k=1}^{r} U_{ikc} V_{kjc}.$$

The update rule is applied independently on U and V such that

$$U_k = U_{k-1} + \frac{\nabla_U \mathcal{L}}{||\nabla_U \mathcal{L}||} \qquad V_k = V_{k-1} + \frac{\nabla_V \mathcal{L}}{||\nabla_V \mathcal{L}||}$$

and the updated perturbation is $\delta_{k+1} = (U_{k+1} \otimes V_{k+1})$. Putting this together, in our universal framework, the LoRa-PGD attack has the following formulation:

$$(U_*, V_*) = \begin{cases} \underset{U,V}{\arg\max} & \frac{1}{B}\sum_{b=1}^{B} \mathcal{L}_{\text{total}}(f(x_b + \delta), f(x_b)) \\ \text{s.t.} & U \in \mathbb{R}^{H \times r \times C}, V \in \mathbb{R}^{r \times W \times C} \\ & r \leq \min\{H, W\} \quad \text{(rank constraint)} \\ & ||U \otimes V||_* \leq \tau \quad \text{(nuclear norm constraint)} \end{cases}$$

FW-Nucl: The Frank-Wolfe nuclear norm group attack (FW-Nucl) obtains structured adversarial examples by constraining perturbations using nuclear group norm regularization [20]. It iteratively applies the Frank-Wolfe algorithm

[4] https://github.com/jsve96/AO-Exp-Attack

to construct a sparse adversarial perturbation. In our universal attack formulation, FW-Nucl has the following form:

$$\delta^* = \begin{cases} \underset{\delta \in \mathbb{R}^{H \times W \times C}}{\arg\min} & \frac{1}{B} \sum_{b=1}^{B} \mathcal{L}_{\text{total}}(f(x_b + \delta), f(x_b)) \\ \text{s.t.} & ||\delta||_{\mathcal{G},1,p} \leq \epsilon \end{cases}$$

5.2 Metrics

Intersection over Union (IoU), also known as the Jaccardi Index, is a well-known measure for the similarity of two shapes or sets $A, A' \in \mathbb{R}^n$,

$$\text{IoU}(A, A') = \frac{|A \cap A'|}{|A \cup A'|},$$

which is often used as a loss function for bounding box regression [28], where A denotes the predicted bounding box of the vision model and A' is a ground truth location of the bounding box. Suppose there are a total of n ground truth bounding boxes of the vision model using the clean frame x_t, and the adversarial example $x_t + \delta$ leads to m predicted bounding boxes, then the IoU for frame x_t is

$$\text{IoU}_t = \sum_{i=1}^{n} \sum_{j=1}^{m} \text{IoU}(A_i, A'_j).$$

We evaluate the average IoU score over all frames and report the accumulated IoU

$$\text{IoU}_{acc} = \frac{1}{T} \sum_{t=1}^{T} \text{IoU}_t$$

to assess the impact of the adversarial attack on the whole sequence. Moreover, we report the ratio of the sum of bounding boxes of the adversarial video clip and the sum of ground truth bounding boxes for the clean video clip (advBR). The Box Ratio indicates whether all ground truth bounding boxes are removed (advBR = 0) or the object detector is fooled when additional bounding boxes appear for the adversarial video frames (advBR > 1). Additionally, we report perceptibility, which we measure based on the mean absolute perturbation (MAP)

$$\text{MAP} = \frac{1}{H \cdot W} \sum_{i,j=1}^{H,W} \sum_{c=1}^{C} |\delta_{ij}^c|.$$

5.3 Datasets

PETS 2009 S2L1 [14]: The PETS 2009 dataset is a benchmark video surveillance dataset designed to evaluate algorithms for multi-camera tracking, crowd analysis, and event detection. It features synchronized footage from multiple

Table 1: Overview of datasets used and key attributes.

Name	Resolution	Scenes	Frames per second	Avg. frames per scene
PETS 2009 S2L1 [14]	768×576	7	7	795
EPFL-RLC [6]	1920×1080	3	60	5000
CW4C	1920×880	15	60	7200

camera views capturing various real-world scenarios, such as people walking, meeting, splitting up, or leaving objects behind. It is widely used in academic research for tasks like people tracking, group activity recognition, and anomaly detection.

Fig. 3: Shows frame number 55 of the PETS 2009 dataset for three different camera views.

EPFL-RLC [6]: The EPFL-RLC dataset is a multi-camera video dataset captured at the Rolex Learning Center of EPFL using three synchronized HD cameras with overlapping fields of view. Each camera records at a resolution of 1920×1080 at 60 frames per second, with the dataset comprising 8,000 frames per view. This dataset is particularly valuable for developing and evaluating multi-view pedestrian detection and tracking algorithms [5].

Fig. 4: Shows three frames from the EPFL-RLC dataset.

CW4C [5] (Coldwater 4 corners): This dataset contains 15 video clips from the publicly available CW4C data [6]. The data captures 4 Corners Park, located at the intersection of Chicago St and Marshall St, in Coldwater (Michigan). In order to limit and save computational cost, we applied downsampling and modified the resolution to 960×440.

Fig. 5: Shows three frames from the first video clip of the camera capturing the crossroad intersection in Coldwater (CW4C).

5.4 Evaluation

For LoRa-PGD attacks and full AO-Exp attacks, we use 100 iterations to obtain a universal perturbation. We set the regularization parameter λ_1 for Algorithm 1 to $\lambda_1 = 0.1$ for PETS datasets, and set $\lambda_2 = \lambda_1/500$ (CW4C), $\lambda_2 = \lambda_1/10$ (PETS2009). We set $\lambda_1 = 0.75$ and $\lambda_2 = 0.005$ (EPFL-RLC). For FW-Nucl, we set $\epsilon = 40$ for comparable results using the same perturbation budget as for AO-Exp. Moreover, we set the number of iterations to 30 and use five updates

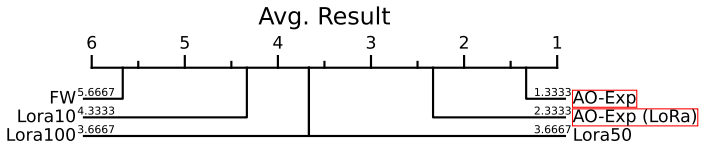


Fig. 6: Shows the critical difference diagram of the average result, obtained based on the scores of IoU_{acc}, advBR, and $||\delta||_*$ (Table 2), across all datasets for each attack method.

for each line search. For the PGD-LoRa attacks, we report three variants with $r = 10\%, 50\%, 100\%$ of the full rank and set the nuclear norm budget to 60. For the low-rank adaption of AO-Exp, we only use the top value ($k = 1$) in (4) and consider 50 iterations.

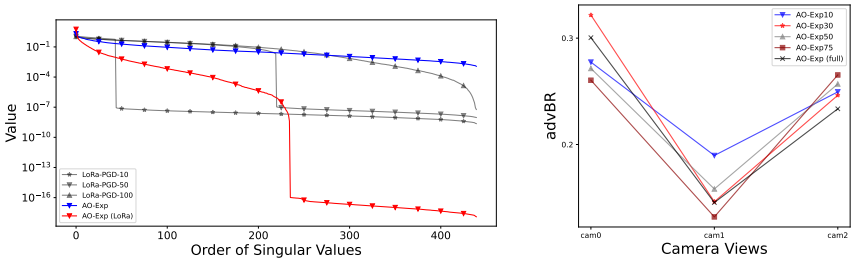
We observe that our proposed method, AO-Exp, achieves the best adversarial box ratio across all datasets while minimizing the accumulated IoU and

[5] https://www.coldwater.org/676/Coldwater-Area-Webcams
[6] https://tinyurl.com/CW4C-Data

Table 2: Results of baseline attack methods and our proposed minimally structured universal attack method AO-Exp over three real-world datasets. For each method, we report the mean value and standard deviation across all scenes of the corresponding dataset. Bold values indicate best result, underlined values second best result.

Dataset	Attack Method	$\text{IoU}_{acc}(\downarrow)$	advBR (↓)	MAP (↓)	$\|\|\delta\|\|_*$ (↓)
PETS2009	FW-Nucl	4.77 ± 1.09	1.04 ± 0.25	$\mathbf{1.2} \pm 0.3$	$\underline{36.5} \pm 5.84$
	LoRa-PGD-10	1.98 ± 1.31	0.94 ± 0.68	4.0 ± 0.4	39.6 ± 1.73
	LoRa-PGD-50	1.48 ± 1.05	0.67 ± 0.41	4.0 ± 0.5	60.9 ± 10.3
	LoRa-PGD-100	$\underline{1.22} \pm 0.91$	0.63 ± 0.42	4.0 ± 0.3	60.3 ± 10.3
	AO-Exp	$\mathbf{0.29} \pm 0.27$	$\mathbf{0.06} \pm 0.04$	$\underline{2.9} \pm 0.1$	41.3 ± 16.6
	AO-Exp (LoRa)	1.88 ± 1.38	$\underline{0.45} \pm 0.33$	6.0 ± 0.2	$\mathbf{17.7} \pm 4.3$
EPFL-RLC	FW-Nucl	4.83 ± 0.96	0.86 ± 0.14	$\mathbf{5.4} \pm 2.0$	37.54 ± 1.53
	LoRa-PGD-10	0.25 ± 0.15	0.31 ± 0.05	14.6 ± 3.0	31.59 ± 1.38
	LoRa-PGD-50	$\mathbf{0.17} \pm 0.03$	$\underline{0.29} \pm 0.14$	14.0 ± 2.7	42.84 ± 2.95
	LoRa-PGD-100	$\underline{0.20} \pm 0.06$	0.37 ± 0.11	14.0 ± 3.0	43.5 ± 4.3
	AO-Exp	0.9 ± 0.37	$\mathbf{0.22} \pm 0.07$	$\underline{6.0} \pm 4.0$	$\underline{27.52} \pm 15.8$
	AO-Exp (LoRa)	1.39 ± 0.60	0.33 ± 0.1	7.0 ± 3.0	$\mathbf{4.7} \pm 0.93$
CW4C	FW-Nucl	4.64 ± 3.69	0.82 ± 0.16	$\mathbf{1.3} \pm 0.2$	39.4 ± 0.5
	LoRa-PGD-10	2.72 ± 2.61	0.48 ± 0.25	5.0 ± 0.4	$\underline{36.1} \pm 1.1$
	LoRa-PGD-50	2.32 ± 2.24	0.41 ± 0.22	4.0 ± 0.2	65.2 ± 6.7
	LoRa-PGD-100	$\underline{1.95} \pm 1.91$	$\underline{0.34} \pm 0.2$	5.0 ± 0.2	67.9 ± 10
	AO-Exp	$\mathbf{0.88} \pm 0.45$	$\mathbf{0.19} \pm 0.06$	$\underline{3.8} \pm 1.0$	37.9 ± 8.9
	AO-Exp (LoRa)	2.12 ± 1.18	0.42 ± 0.17	5.0 ± 4.0	$\mathbf{14.3} \pm 7.6$

notably minimal nuclear norm, as shown in Table 2. This also holds for the MAP, indicating stealth attacks in general. Notably, our low-rank adaption not only yields comparable results in the average adversarial box ratio but also drastically minimizes the nuclear norm of the perturbation compared to AO-Exp, as shown in Figure 7a. Our proposed attack method, AO-Exp, and its low-rank adaption surpass the considered baseline attacks across all datasets, averaged over three main metrics, see Figure 6. Moreover, we observe that different values of k in the update rule of Algorithm 1 yield similar advBR for each camera angle of the EPFL dataset, Figure 7b. Increasing the number of singular values for the construction of the structured adversarial perturbation enhances the effectiveness of the adversarial attack, as expected.



(a) Median singular values of LoRa-PGD attacks, AO-Exp attack, and low-rank adaption (AO-Exp LoRa) for CW4C dataset. AO-Exp (LoRa) uses only the top singular value in Equation (4).

(b) Shows adversarial box ratio of five variants of AO-Exp with different values of k in Equation (4) for each camera view of the EPFL dataset.

Fig. 7: Additional results for CW4C and EPFL datasets considering low-rank adaptions of AO-Exp.

6 Limits & Conclusion

In this work, we proposed a novel minimally distorted universal adversarial attack designed for video-based object detection systems. By leveraging nuclear norm regularization, our method promotes structured perturbations that primarily target the background, enabling stealthier and more natural-looking adversarial examples. To tackle the associated computational complexity, we leverage an adaptive optimistic exponentiated gradient descent approach, which improves both scalability and convergence.

Despite these promising results, our approach has some limitations. First, the current formulation assumes a static camera setup, limiting its applicability to dynamic camera scenarios. Second, the attack's performance is sensitive to the choice of hyperparameters, such as the nuclear norm weight and Frobenius regularization, which may require task-specific tuning.

Future work may explore extending this approach to dynamic camera settings, extending the work to object tracking, developing adaptive or learned hyperparameter strategies, and integrating semantic or temporal consistency constraints to improve generalizability and stealth in more complex real-world scenarios. Additionally, one may explore countermeasures and defenses tailored specifically to structured and temporally consistent adversarial attacks.

Acknowledgement. This research was funded by the German Federal Ministry of Labour and Social Affairs through the establishment of a Junior Research Group on Artificial Intelligence at the Federal Institute of Occupational Safety and Health (BAuA). The presented results contribute to the development and evaluation of reliable and safe AI for industrial applications, with the overarching

aim of laying the scientific foundations necessary to meet the requirements of the European Machinery Directive (2023) and the European AI Act (2024).

References

1. Alkaabi, S., AlAzri, A., AlZakwani, S., Altamimi, F.: A methodology to evaluate video analytics for drilling safety operation using machine learning. In: Middle East Oil, Gas and Geosciences Show. OnePetro (2023)
2. Almujally, N.A., Qureshi, A.M., Alazeb, A., Rahman, H., Sadiq, T., Alonazi, M., Algarni, A., Jalal, A.: A novel framework for vehicle detection and tracking in night ware surveillance systems. Ieee Access (2024)
3. Amirkhani, A., Karimi, M.P., Banitalebi-Dehkordi, A.: A survey on adversarial attacks and defenses for object detection and their applications in autonomous vehicles. The Visual Computer **39**(11), 5293–5307 (2023)
4. Bouwmans, T., Javed, S., Zhang, H., Lin, Z., Otazo, R.: On the applications of robust pca in image and video processing. Proceedings of the IEEE **106**(8), 1427–1457 (2018). https://doi.org/10.1109/JPROC.2018.2853589
5. Chavdarova, T., Baqué, P., Bouquet, S., Maksai, A., Jose, C., Bagautdinov, T., Lettry, L., Fua, P., Van Gool, L., Fleuret, F.: Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5030–5039 (2018)
6. Chavdarova, T., Fleuret, F.: Deep multi-camera people detection. In: 2017 16th IEEE international conference on machine learning and applications (ICMLA). pp. 848–853. IEEE (2017)
7. Chen, X., Kundu, K., Zhang, Z., Ma, H., Fidler, S., Urtasun, R.: Monocular 3d object detection for autonomous driving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
8. Cocca, P., Marciano, F., Alberti, M.: Video surveillance systems to enhance occupational safety: A case study. Safety Science **84**, 140–148 (2016)
9. Croce, F., Hein, M.: Mind the box: l_1-apgd for sparse adversarial attacks on image classifiers. In: International Conference on Machine Learning. pp. 2201–2211. PMLR (2021)
10. Ding, J., Xu, Z.: Adversarial attacks on deep learning models of computer vision: A survey. In: Algorithms and Architectures for Parallel Processing: 20th International Conference, ICA3PP 2020, New York City, NY, USA, October 2–4, 2020, Proceedings, Part III 20. pp. 396–408. Springer (2020)
11. Fefferman, C., Mitter, S., Narayanan, H.: Testing the manifold hypothesis. Journal of the American Mathematical Society **29**(4), 983–1049 (2016)
12. Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., Wiesbeck, W., Dietmayer, K.: Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Transactions on Intelligent Transportation Systems **22**(3), 1341–1360 (2020)
13. Feng, J., Cai, Q.Z., Zhou, Z.H.: Learning to confuse: Generating training time adversarial data with auto-encoder. Advances in Neural Information Processing Systems **32** (2019)
14. Ferryman, J., Shahrokni, A.: Pets2009: Dataset and challenge. In: 2009 Twelfth IEEE international workshop on performance evaluation of tracking and surveillance. pp. 1–6. IEEE (2009)

15. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
16. Gu, S., Zhang, L., Zuo, W., Feng, X.: Weighted nuclear norm minimization with application to image denoising. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2862–2869 (2014)
17. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
18. Huang, H., Wang, Y., Chen, Z., Tang, Z., Zhang, W., Ma, K.K.: Rpattack: Refined patch attack on general object detectors. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6 (2021). https://doi.org/10.1109/ICME51207.2021.9428443
19. Jha, S., Seo, C., Yang, E., Joshi, G.P.: Real time object detection and trackingsystem for video surveillance system. Multimedia Tools and Applications **80**(3), 3981–3996 (2021)
20. Kazemi, E., Kerdreux, T., Wang, L.: Minimally distorted structured adversarial attacks. International Journal of Computer Vision **131**(1), 160–176 (2023)
21. Kim, B., Sagduyu, Y.E., Davaslioglu, K., Erpek, T., Ulukus, S.: Channel-aware adversarial attacks against deep learning-based wireless signal classifiers. IEEE Transactions on Wireless Communications **21**(6), 3868–3880 (2021)
22. Kravchik, M., Shabtai, A.: Efficient cyber attack detection in industrial control systems using lightweight neural networks and pca. IEEE transactions on dependable and secure computing **19**(4), 2179–2197 (2021)
23. Li, D., Zhang, J., Huang, K.: Universal adversarial perturbations against object detection. Pattern Recognition **110**, 107584 (2021)
24. Malburg, L., Rieder, M.P., Seiger, R., Klein, P., Bergmann, R.: Object detection for smart factory processes by machine learning. Procedia Computer Science **184**, 581–588 (2021)
25. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1765–1773 (2017)
26. Nguyen, K.N.T., Zhang, W., Lu, K., Wu, Y.H., Zheng, X., Tan, H.L., Zhen, L.: A survey and evaluation of adversarial attacks in object detection. IEEE Transactions on Neural Networks and Learning Systems (2025)
27. Pérez, L., Rodríguez, Í., Rodríguez, N., Usamentiaga, R., García, D.F.: Robot guidance using machine vision techniques in industrial environments: A comparative review. Sensors **16**(3), 335 (2016)
28. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 658–666 (2019)
29. Sadiku, S., Wagner, M., Pokutta, S.: Gse: Group-wise sparse and explainable adversarial attacks. arXiv preprint arXiv:2311.17434 (2023)
30. Savostianova, D., Zangrando, E., Tudisco, F.: Low-rank adversarial pgd attack. arXiv preprint arXiv:2410.12607 (2024)
31. Shao, W., Sivrikaya, F., Albayrak, S.: Optimistic optimisation of composite objective with exponentiated update. Machine Learning (Aug 2022). https://doi.org/10.1007/s10994-022-06229-1, https://doi.org/10.1007/s10994-022-06229-1
32. Shi, Z., Ming, Y., Fan, Y., Sala, F., Liang, Y.: Domain generalization via nuclear norm regularization. In: Conference on Parsimony and Learning. pp. 179–201. PMLR (2024)

33. Shukla, N., Banerjee, S.: Generating adversarial attacks in the latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 730–739 (2023)
34. Subhash, V., Bialas, A., Pan, W., Doshi-Velez, F.: Why do universal adversarial attacks work on large language models?: Geometry might be the answer. In: The Second Workshop on New Frontiers in Adversarial Machine Learning (2023)
35. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
36. Thys, S., Van Ranst, W., Goedemé, T.: Fooling automated surveillance cameras: adversarial patches to attack person detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019)
37. Wang, Y., an Tan, Y., Zhang, W., Zhao, Y., Kuang, X.: An adversarial attack on dnn-based black-box object detectors. Journal of Network and Computer Applications **161**, 102634 (2020). https://doi.org/https://doi.org/10.1016/j.jnca.2020.102634, https://www.sciencedirect.com/science/article/pii/S1084804520301089
38. Wu, X., Huang, L., Gao, C.: G-uap: Generic universal adversarial perturbation that fools rpn-based detectors. In: Lee, W.S., Suzuki, T. (eds.) Proceedings of The Eleventh Asian Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 101, pp. 1204–1217. PMLR (17–19 Nov 2019), https://proceedings.mlr.press/v101/wu19a.html
39. Zhang, H., Zhou, W., Li, H.: Contextual adversarial attacks for object detection. In: 2020 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6 (2020). https://doi.org/10.1109/ICME46284.2020.9102805
40. Zhao, Z.Q., Zheng, P., Xu, S.T., Wu, X.: Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems **30**(11), 3212–3232 (2019). https://doi.org/10.1109/TNNLS.2018.2876865
41. Zhou, X., Xu, X., Liang, W., Zeng, Z., Shimizu, S., Yang, L.T., Jin, Q.: Intelligent small object detection for digital twin in smart manufacturing with industrial cyber-physical systems. IEEE Transactions on Industrial Informatics **18**(2), 1377–1386 (2021)