The Chosen-Object Attack: Exploiting the Hungarian Matching Loss in Detection Transformers for Fun and Profit

Tianyi Wang, *Graduate Student Member, IEEE*, Cong Wang, *Member, IEEE*, Zhenyu Wen, *Senior Member, IEEE*, Ruilong Deng, *Senior Member, IEEE*, Yuanchao Shu, *Senior Member, IEEE*, Peng Cheng, *Member, IEEE*, and Jiming Chen, *Fellow, IEEE*

***Abstract*—Different from traditional object detectors such as YOLO, Detection Transformers (DETR) have reshaped the landscape of object detection by replacing heuristic-driven components like Non-Maximal Suppression with a fully end-to-end framework based on one-to-one Hungarian matching. While the majority of research has focused on improving the slow training convergence of DETR, this work investigates their security from an adversarial perspective. We unveil a critical vulnerability stemming directly from DETR's core design: the deterministic one-to-one mapping between object queries and ground-truth objects can be exploited. This allows an adversary to craft perturbations that selectively manipulate specific target objects, causing them to vanish or be misclassified, while still preserving the detection integrity of all other objects in the scene. Our initial analysis reveals that conventional gradient-based attacks are ill-suited for this task, as they induce unintended interference on non-target instances, a phenomenon we term as the "spillover effect". To overcome this, we re-formulate the attack optimization by incorporating a novel penalty term that explicitly decouples the adversarial influence on target and non-target objects. Furthermore, we provide a theoretical analysis to derive perturbation bounds under which the optimal matching assignments remain invariant, offering deeper insights into the model's stability. Extensive experiments on standard benchmarks demonstrate that our proposed attack significantly improves the success rate and convergence speed while inducing far fewer feature-level artifacts, making the attack both more effective and stealthier. The source code is available at: https://github.com/Hill-Wu-1998/coa**

***Index Terms*—Adversarial machine learning, object detection, vision transformer.**

I. Introduction

SINCE its inception, Detection Transformer (DETR) has redefined conventional object detectors [1]. While traditional methods such as YOLO [2], Faster RCNN [3] and RetinaNet [4] heavily rely on hand-crafted components such as anchors [5], proposals [3], window centers [6] and Non-Maximum Suppression (NMS) [7] for post-processing, DETR introduces an end-to-end architecture by viewing object detection as a set prediction problem [8], [9], [10] and leverages the transformer encoder-decoder architecture. The core innovation is the elimination of NMS and its replacement with a one-to-one bipartite matching loss optimized by the Hungarian algorithm during training [11]. This not only simplifies the detection pipeline, but also addresses inherent limitations in NMS such as suboptimal duplicate removal [12] and computational latency [13], [14].

Despite its success, DETR suffers from critical issues of slow convergence speed [15], [16], [17], [18], which has been largely attributed to sparse supervision. That is, the Hungarian matching process restricts positive training signals to a single query per object, leaving the majority of queries unassigned ("no object"). This sparsity further exacerbates two critical issues: 1) feature misalignment in cross-attention layers, where queries struggle to localize salient features [15], [17], [19]; 2) instability of query-object bindings during iterative decoding [20].

We posit that such sparse interaction, which was intended to streamline end-to-end detection, creates a unique and overlooked attack surface against DETR. Unlike conventional detectors with anchored priors (e.g., YOLO's grid cells [5]), decoders in DETR dynamically associate learnable query embeddings with objects through global attention mechanisms. That is, the Hungarian matching establishes a deterministic, minimum-cost query-to-object mapping during training, which persists as a predictable correspondence at inference. This predictability inadvertently provides adversaries with a foothold–the lack of dense prior reduces the search efforts to determine which query corresponds to a victim object, such that attackers can craft object-specific adversarial examples to suppress attention (object hiding) or distort localization of bounding boxes (object dislocation) in a stealthy way. For example, attackers could manipulate queries against precision-sensitive objects such as robotic grasping, contraband in X-ray scan, pedestrians in self-driving and object tracking, posing risks to real-world applications.

To exploit this vulnerability, we propose *The Chosen-Object Attack*, inspired by the Chosen-Plaintext Attacks in cryptography. Analogous to how cryptanalysts query specific

Received 26 June 2025; revised 15 December 2025; accepted 12 January 2026. Date of publication 16 January 2026; date of current version 20 February 2026. This work was supported in part by the NSFC under Grant 62576310; and in part by the NSF of Zhejiang Province under Grant LZ25F020007, Grant LDQ24F020001, Grant 2025C01012, Grant NSFC 92467301, and Grant 62293511. The associate editor coordinating the review of this article and approving it for publication was Dr. Tianwei Zhang. *(Corresponding author: Cong Wang.)*

Tianyi Wang, Cong Wang, Ruilong Deng, Yuanchao Shu, Peng Cheng, and Jiming Chen are with the State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310007, China (e-mail: cwang85@zju.edu.cn).

Zhenyu Wen is with the Institute of Cyberspace Security and the College of Information Engineering, Zhejiang University of Technology, Hangzhou 310014, China.

Digital Object Identifier 10.1109/TIFS.2026.3654868

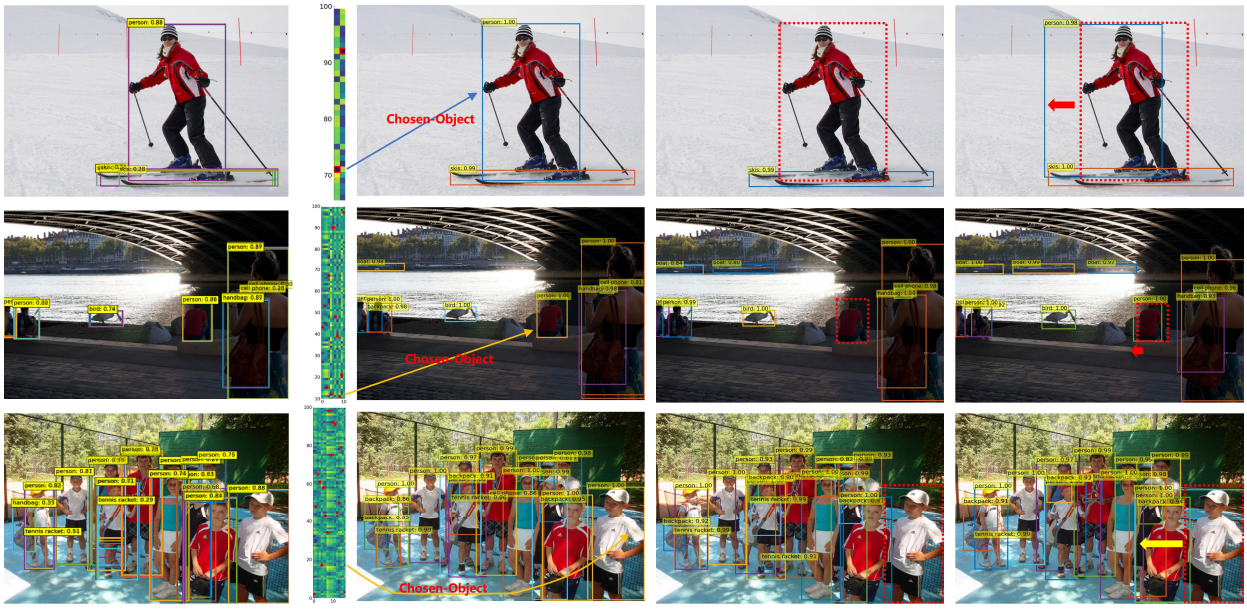


Fig. 1. Visualization of the Chosen-Object Attack against DETR. (a) Contrasting with YOLO's output before NMS; (b) DETR's one-to-one output, which provides the unique attack surface, and the selection of the "Chosen-Object"; (c) Making the target object disappear; (d) Dislocating the target object.

plaintexts to reverse-engineer encryption, our attack manipulates the Hungarian matching loss to hijack target object assignments. Although the attack may sound straightforward conceptually, our initial experiments reveal a non-trivial challenge: naive gradient-based perturbations induce *Spillover Effects*, a collateral damage to non-target objects caused by correlated gradient directions between queries. To mitigate this, we propose a gradient projection mechanism extended from multi-task optimization [21], which decouples the gradients between target and non-target queries through orthogonal projections. Further, we also derive sensitivity-aware perturbation bounds through the lens of Linear Assignment sensitivity analysis [22], [23], in order to guarantee the attacks remain within the stability region of optimal matching. From a defensive standpoint, we also demonstrate that DETR variants with one-to-many label assignments [16], [24], [25] provide inherent robustness against our attack, as their redundant query-object mappings *obscure* adversarial targeting. This finding connects the previously disjoint goals of improving convergence speed and enhancing robustness in the context of DETR with a unified perspective for future designs. The main contributions of this paper are summarized below.

- ✧ **New Attack Surface against DETR.** Security vulnerabilities often remain latent for extended periods before being recognized. Mirroring hardware vulnerabilities like Meltdown (undetected in Intel CPUs for 20+ years until 2017) [26], we reveal how DETR's fundamental *one-to-one Hungarian matching* creates a unique combinatorial attack surface–the first security analysis of its kind for transformer-based detection. To our best knowledge, this is not only the first work that sheds light on the unique security vulnerabilities arisen from the original DETR architecture, but also an initial attempt to bridge the gap between DETR's convergence limitations and adversarial robustness.
- ✧ **Chosen-Object Attack with Spillover Mitigation.** We propose the Chosen-Object Attack and integrate gradient projection methods to minimize collateral damage on non-target objects. We further provide a sensitivity analysis on the cost matrix level to illustrate natural tolerance under small changes in bipartite graph matching.
- ✧ **Extensive Evaluation and Discussions.** Through extensive experiments, we demonstrate that the proposed method improves the attack success rate by more than 0.2 with 6× speedup and induces fewer feature-level artifacts in order to remain stealthy. We also validate interesting insights of inherent robustness through one-to-many label assignments, as well as provide tentative extensions towards patch-based attacks.

The rest of the paper is organized as follows. Section II provides a literature review on adversarial attacks against object detectors and Section III introduces the basics of detection transformers. Section IV outlines the attack framework and performs further enhancement. Section V offers theoretical insights into the perturbation sensitivity. Section VI conducts experiments and Section IX concludes this paper.

II. Related Works

A. Object Detectors

1) One/Two-Stage Detectors: Conventional detectors can be broadly categorized into one-stage [2], [4], [6], [27] and two-stage detectors [3], [28]. One-stage detectors employ a dense set of anchors to directly predict class labels and bounding box offsets from feature maps in a single pass, while two-stage detectors first generate region proposals [29], [30] and then refine the proposals through a dedicated regression process. Unfortunately, they both rely on hand-crafted components such as anchors, proposals and Non-Maximal Suppression (NMS) to eliminate duplicate bounding boxes due to one-to-many label assignment.

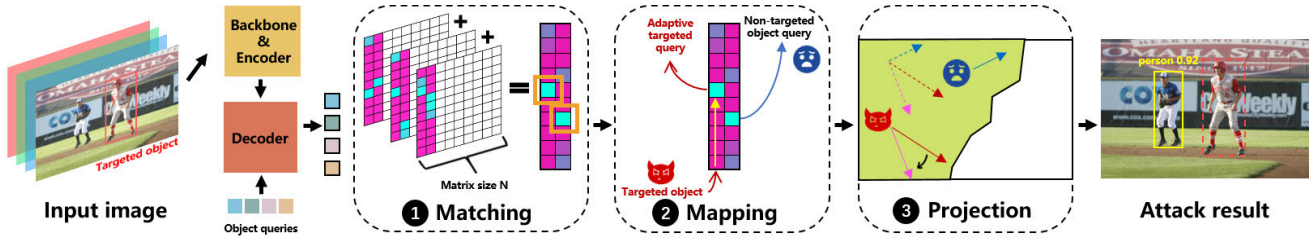


Fig. 2. Overview of the Chosen-Object Attack: ❶ the attacker computes the cost matrix (Eq. (3)); ❷ distinguishes the target object from the non-target objects by reverse-engineering the one-to-one query-to-object mapping; ❸ mitigates the spillover effects on non-target objects via orthogonal gradient projection.

2) End-to-End Detection Transformers: DETR (DEtection TRansformer) resolves the many-to-one label assignment problem and hand-crafted components such as NMS are no longer needed [1]. However, its slow convergence has spurred a plethora of efforts to accelerate training [15], [16], [17], [18]. Deformable DETR replaces global self-attention with deformable attention that focuses on sparse and informative regions around reference points [16], while DN-DETR introduces query de-noising to stabilize training [19]. SAM-DETR aligns object queries with semantically meaningful features [15] and Conditional DETR simplifies spatial queries to ease cross-attention [20]. Stable-DINO only uses positional metrics to supervise the classification scores of positive examples for stable matching [31]. Other variants further enhance performance through hybrid one-to-many matching or group-based decoding [17], [32], [33], [34]. Despite these advances, little attention has been paid to DETR's security implications, leaving critical gaps to understand its adversarial robustness.

B. Adversarial Attacks Against Object Detectors

1) Attacks on Conventional Detectors: The multi-tasking nature of object detectors makes them vulnerable to various forms of attacks, including misclassification attacks [35], dislocation attacks [36], latency attacks [13], [14] and structural attacks [37]. These attacks are designed to disrupt specific functionalities of the object detector with convolutional backbone. Misclassification and dislocation attacks exploit non-overlapping adversarial patches to disrupt the spatial context [35], [36]. Latency attacks aim to degrade real-time processing performance by overwhelming the quadratic-time post-processing stage (NMS) with large number of candidate boxes [13], [14]. Structural attack aims to globally mislead the model's output by devising optimal attack strategies [37]. Beyond inference-time threats, data-centric attacks during training pose significant risks. Recent work in federated learning has introduced SADBA, a self-adaptive distributed backdoor attack that achieves high efficacy with minimal compromised clients. Similarly, manipulation attacks on statistical aggregation under Local Differential Privacy have prompted defenses like RobustLDP, which adaptively adjusts output complexity to ensure robustness. These works highlight the broad spectrum of vulnerabilities in AI pipelines [38], [39].

2) Attacks on Vision Transformers: While ViTs exhibit inherent robustness due to their global receptive fields [40], [41] and smooth loss landscapes [42], [43], recent studies expose vulnerabilities in their attention mechanisms. Attacks such as Patch-Fool [44] perturb patch embeddings, and Attention-Fool [45] manipulates query-key interactions to misguide relevance scores. These attacks exploit the dot-product attention mechanism, which computes relevance scores between query (Q) and key (K) vectors to weigh input features dynamically. Existing research on security has predominantly focused on the classification tasks of ViTs, whereas leaving detection-specific components such as bipartite matching and object queries unexamined for adversarial susceptibility.

III. Revisiting Bipartite Matching in DETR

DETR learns a function $f_\theta : \mathbf{x} \mapsto \{\hat{p}, \hat{b}\}$ that maps an input image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$ to a set of N predictions. Each prediction consists of a class probability $\hat{p}$ and bounding box $\hat{b}$. DETR incorporates the Hungarian algorithm to solve a bipartite matching problem between the predictions and ground-truth objects. For simplicity, the ground-truth set $\mathbf{y}$ is padded with $\varnothing$ (no object) to ensure a fixed size N to match the number of predictions. The training objective minimizes the Hungarian loss,

$$\mathcal{L}_{\text{Hungarian}}(y,\hat{y}) = \sum_{i=1}^{N} \left[\mathcal{L}_{\text{class}}(i,\hat{\sigma}(i)) + \mathbb{1}_{\{c_i\neq\emptyset\}}\mathcal{L}_{\text{box}}(b_i,\hat{b}_{\sigma}(i))\right] \quad (1)$$

where $\mathcal{L}_{\text{class}}(i,\hat{\sigma}(i))$ and $\mathcal{L}_{\text{box}}(b_i,\hat{b}_{\sigma}(i))$ are the classification and bounding box regression loss between the i-th ground truth and the prediction $\hat{\sigma}(i)$. $\hat{\sigma}$ is the bipartite matching between the ground truth y and the prediction set $\hat{y}$ to find the lowest cost in all N permutations $\sigma \in \mathcal{G}_N$,

$$\hat{\sigma} = \arg\min_{\sigma\in\mathcal{G}_N} \sum_{i=1}^{N} \mathcal{L}_{\text{match}}(y_i,\hat{y}_{\sigma}(i)) \quad (2)$$

where $\mathcal{G}_N$ is the symmetric group of all permutations over N elements. The pairwise matching cost is,

$$\mathcal{L}_{\text{match}}(y_i,\hat{y}_{\sigma(i)}) = -\mathbb{1}_{\{c_i\neq\emptyset\}}\hat{p}_{\sigma(i)}(c_i) + \mathbb{1}_{\{c_i\neq\emptyset\}}\mathcal{L}_{\text{box}}(b_i,\hat{b}_{\sigma}(i)) \quad (3)$$

where $\hat{p}_{\sigma(i)}(c_i)$ is the predicted probability for the ground-truth class c_i. and $\mathcal{L}_{\text{match}}(y_i,\hat{y}_{\sigma(i)})$ is obtained via solving the Linear Assignment Problem (LAP) below.

1) Linear Assignment Problem (LAP):

For N bounding box prediction and ground-truth entries (including $\varnothing$), the LAP finds an optimal assignment $\pi^* \in \{0,1\}^{N\times N}$ that minimizes the total cost,

$$\min_{\pi} \sum_{j=1}^{N}\sum_{i=1}^{N} c_{ij}\pi_{ij} \quad \text{s.t.} \quad \sum_{i=1}^{N}\pi_{ij}=1 \quad \forall i, \quad \sum_{j=1}^{N}\pi_{ij}=1 \quad \forall j. \tag{4}$$

IV. Crafting the Chosen-Object Attack

A. Threat Model

The primary goal of the attacker is to manipulate the detection output for a specific, pre-selected object (the "chosen object") within an image, while leaving the detection of other objects largely unaffected to evade detection. This goal manifests in two distinct modalities as defined by the multi-tasking nature of object detectors: 1) Object Hiding (Disappearance). To cause the detector to fail to classify the target object above the confidence threshold, effectively making it invisible to the system (e.g., hiding a pedestrian or a weapon); Object Dislocation. To cause the detector to report the target object with a significantly incorrect bounding box, misleading downstream tasks that rely on precise localization (e.g., shifting the perceived location of a traffic light).

We define the attacker's capabilities under two primary scenarios. White-Box (Primary): This is the main focus of our work. The attacker has complete knowledge of and access to the target model, including the model architecture (e.g., DETR with a ResNet-50 backbone) and its parameters (weights), the ability to compute the full loss function and its gradients with respect to the input image. This is necessary to craft the perturbation using PGD, reverse-engineer the query-to-object mapping as discussed later. Black-Box (Transferability): The attacker has no direct access to the target model's parameters or gradients. Instead, he leverages the transferability of adversarial examples by crafting the perturbation on a publicly available or self-trained surrogate DETR model and applying it to the black-box victim, which is evaluated in Section VI-F.

B. Exploit the Matching Loss

The Hungarian matching in DETR (Eq. 4) establishes a one-to-one correspondence $\hat{\sigma}$ between N predictions and ground-truth entries (padded with $\varnothing$):

$$\hat{\sigma} \sim [(\hat{y}_{\hat{\sigma}_1}, y_1), (\hat{y}_{\hat{\sigma}_2}, y_2), \cdots, (\hat{y}_{\hat{\sigma}_N}, y_N)]., \tag{5}$$

where y_i is the i-th ground truth and $\hat{y}_{\hat{\sigma}(i)}$ is its matched prediction. Specifically, this design has exposed a unique opportunity for the attacker to target a chosen object y_T. By using the relation in Eq. (5), the attacker can easily trace from y_T to the original prediction $\hat{y}_{\hat{\sigma}_T}$ as formally defined below.

Definition 1 (The Chosen-Object Attack): For a target object $T \in \{1, \cdots, N\}$, craft perturbation $\|\delta\|_\infty \le \epsilon$ that maximizes the matching loss,

$$\mathbf{P1}: \quad \delta = \arg\max_{\|\delta\|_\infty \le \epsilon} \Big[\mathbb{1}_{\{a=\text{disappear}\}} \mathcal{L}_{\text{class}}(T, \sigma^*_\delta(T)) + \mathbb{1}_{\{a=\text{dislocate}\}} \mathcal{L}_{\text{box}}\big(b_T, \hat{b}_{\sigma^*_\delta}(T)\big) \Big] \tag{6}$$

$$\text{where} \quad \sigma^*_\delta = \arg\min_{\sigma\in\mathcal{G}_N} \sum_{i=1}^{N} \mathcal{L}_{\text{match}}\left(y_i, \hat{y}_\sigma(i;\delta)\right). \tag{7}$$

σ^*_δ is the optimal matching under perturbation δ. The attack modality $a \in \{\text{disappear}, \text{dislocate}\}$ determines whether to suppress the classification confidence or displace the bounding box of target object T.

1) Dynamic Matching Adaptation: A key observation is that static targeting $\hat{y}_{\sigma^*(T)}$ is insufficient because there are potentially multiple queries that match with y_T during optimization, though only the minimum-cost one is selected. As the perturbation δ evolves during the attack process, σ^*_δ re-assigns predictions dynamically, and surprisingly those "suboptimal" queries can still successfully detect y_T, but with lower mAP and IoU. Thus, the attack must continuously adapt to the current σ^*_δ during perturbation crafting process. This finding echoes with [31] of unstable matching that a mixture of queries with high classification score/low IoU and low classification score/high IoU co-exist during training, but our discovery takes the adversarial perspective and more experimental results are available in Appendices.

C. The Spillover Effect

Although it is straightforward to accomplish attacks against a specific object, making the attack stealthy is challenging given the artifacts as discussed below.

Definition 2 (Spillover Effect): Perturbations targeting y_T inadvertently degrade detection of non-target objects y_i ($i \neq T$), as quantified by

$$\mathcal{L}^i_{\text{spillover}} = \begin{cases} \hat{p}_{\sigma^*}(i) - \hat{p}_{\sigma^*_\delta}(i), & a = \text{disappear}, \\ \text{IoU}(b_i, \hat{b}_{\sigma^*}(i)) - \text{IoU}(b_i, \hat{b}_{\sigma^*_\delta}(i)), & a = \text{dislocate}, \end{cases} \tag{8}$$

where σ^* is the original matching results before perturbations are introduced.

Illustration. Fig. 3a presents a visualization when we continuously impose adversarial perturbations on the target "person" and observe the spillover effect on "skis". The effect starts to strengthen and pushes the probability of skis below the confidence threshold after step 38 in Fig. 3b. It also reveals that prediction probability declines at different rates for the target and non-target objects.

We conjecture that this phenomenon originates from two aspects: 1) **Gradient Similarity**. Shared self-attention layers cause overlapping gradient directions for y_T and y_i against different non-target objects $i \neq T$. Fig. 3c validates this by examining the cosine similarity between the gradient direction of the target and non-target objects on the testing set of COCO. It indicates that over 80% objects have influences on non-target objects with similar gradient directions. 2) **LAP's Robust Interval**. An inherent robust interval from the structure of the combinatorial LAP [22], [23], and the matching σ^*_δ remains stable until δ exceeds a certain level, after which it abruptly resorts to suboptimal solutions (see the next Section V for more details).

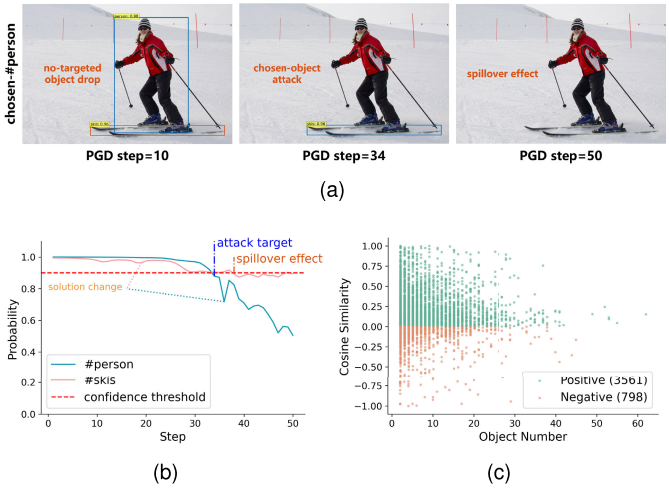


Fig. 3. Analysis of the Spillover Effect during the Chosen-Object Attack. (a) Visualization of the Spillover Effect on "skis". Although the attack only targets "person", both objects disappear at step 50; (b) As the accuracy of the target object declines, the detection of non-target object starts to drop after step 20; (c) Cosine similarity of loss gradients between target and non-target objects on MS-COCO. Positive values (green) indicate aligned gradients, implying a high likelihood of spillover effects, whereas negative values (orange) suggest opposing gradients with minimal collateral risk. The dominance of positive alignment (>80%) confirms the prevalence of the spillover effect.

D. Mitigate Spillover via Gradient Projection

To minimize collateral damage, we re-formulate the attack by introducing a penalty term for the spillover, with λ controlling the trade-off between the two losses,

$$\mathbf{P2}: \quad \delta = \arg\max_{\|\delta\|_\infty \le \epsilon} \Big[\mathbb{1}_{\{a=\text{disappear}\}} \mathcal{L}_{\text{class}}(T, \sigma^*_\delta(T)) + \mathbb{1}_{\{a=\text{dislocate}\}} \mathcal{L}_{\text{box}}\big(b_T, \hat{b}_{\sigma^*_\delta}(T)\big) \Big] - \lambda \sum_{i \neq T} \beta_i \mathcal{L}^i_{\text{spillover}} \tag{9}$$

where β_i modulates spillover penalties using the principle of PCGrad [21]:

$$\beta_i = \begin{cases} 0, & \langle \nabla_\delta \mathcal{L}_T, \nabla_\delta \mathcal{L}^i_{\text{spillover}} \rangle \le 0, \\ \dfrac{\langle \nabla_\delta \mathcal{L}_T, \nabla_\delta \mathcal{L}^i_{\text{spillover}} \rangle}{\|\nabla_\delta \mathcal{L}^i_{\text{spillover}}\|^2}, & \text{otherwise.} \end{cases} \tag{10}$$

Here, $\mathcal{L}_T$ is the target loss (classification or bounding box). To illustrate the conditions on β_i, we extend PCGrad in a different way by decomposing perturbation gradients that maximize the loss of non-target objects as *harmful gradients* along with target gradients.[1] When gradients for y_T and spillover on y_i align ($\beta_i > 0$), we project $\nabla_\delta \mathcal{L}^i_{\text{spillover}}$ orthogonally to $\nabla_\delta \mathcal{L}_T$ (illustrated in Fig. 4) in order to preserve the attack efficacy while minimizing collateral damage; otherwise, the gradients are irrelevant and we set β_i to zero. This method ensures that when the perturbation is amplified along the target gradients, it exclusively impacts the target object without degrading the detection performance of non-target objects.

1) Optimization: We solve Eq. 9 via projected gradient descent (PGD) [46] with Adam [47]:

$$\delta^{t+1} = \Pi_{[-\epsilon,\epsilon]} \left(\delta^t + \mu \cdot \text{Adam}(\nabla_\delta \mathcal{L}_{\text{total}}) \right), \tag{11}$$

[1]The original PCGrad advocates objectives in the similar directions. Here, we perform the opposite for adversarial attacks.

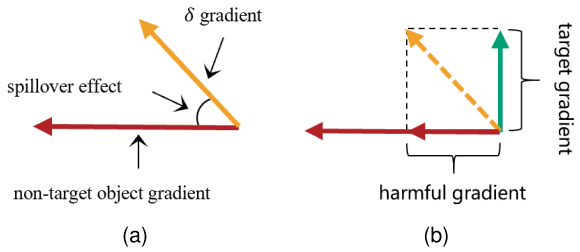


Fig. 4. Adjust the gradient directions to mitigate the spillover effect on non-target objects. (a) Spillover effect. (b) New gradient direction.

where Π clips perturbations to the ℓ_∞-ball of radius ϵ, μ is the step size and $\mathcal{L}_{\text{total}}$ is the total loss. The Hungarian matching $\sigma^*_{\delta^t}$ is computed at each step t to account for dynamic assignment changes.

V. Theoretical Analysis

Recall that Fig. 3b highlights an intriguing phenomenon that the spillover effect exhibits a slow start. In addition to the gradient dissimilarity, we also attribute this behavior to the natural tolerance to small changes in the cost matrix of bipartite graph matching [22], [23]. That is, the structure of the combinatorial problem allows for certain perturbations on specific entries of the cost matrix without altering the optimal matching results. To formalize this, we define Δc_{ij} as the changes on the cost matrix c_{ij} due to adversarial perturbations processed by the DETR backbone $f_\theta(\cdot)$, $\Delta c_{ij} = f_\theta(x+\delta) - f_\theta(x)$. Then we introduce allowable perturbation and derive the theoretical bounds next.

Definition 3 (Allowable Perturbation): For non-target objects $i \neq T$, if the optimal assignment $\pi^* = \{\pi_{ij}, c_{ij}\}$ remains invariant under a set of perturbations, i.e., $\pi^* = \{\pi_{ij}, c_{ij} + \Delta c_{ij}\}$, then Δc_{ij} is termed an *allowable perturbation* on the cost matrix. Such perturbations do not alter the optimal matching for non-target objects.

Theorem 1 (Allowable Perturbation Bounds): For non-target objects $i \neq T$, let $\{\Delta c_{ij}\}$ be an *allowable perturbation* in Definition 3. For another perturbation $\{\Delta c'_{ij}\}$ and *optimal entries* (i, j), the allowable perturbation bounds are given by,

$$\Delta c'_{ij} = \begin{cases} \min(\Delta_{\text{row}}, \Delta_{\text{col}}), & \Delta c'_{ij} \ge 0 \\ -\infty, & \Delta c'_{ij} < 0 \end{cases} \tag{12}$$

where $\Delta_{\text{row}} = \min_{k \neq j}(c_{ik} - (u_i + v_k))$, $\Delta_{\text{col}} = \min_{l \neq i}(c_{lj} - (u_l + v_j))$ are the row-wise and column-wise minimum value of the reduced cost, respectively. u_i, v_j are the dual variables satisfying $u_i + v_j = c_{ij}$. On the other hand, for *non-optimal entries* (i, j),

$$\Delta c'_{ij} = \begin{cases} +\infty, & \Delta c'_{ij} \ge 0 \\ c_{ij} - (u_i + v_j), & \Delta c'_{ij} < 0 \end{cases} \tag{13}$$

The dual variable u_i associates with the queries (the minimum cost from query i) and v_j associates with the ground truth objects (the maximum cost assigning to j). $u_i + v_j = c_{ij}$ ensures that cost c_{ij} is split between the two variables.

Proof: To find the bounds of perturbations imposed on the cost matrix, we consider the dual problem of the primal

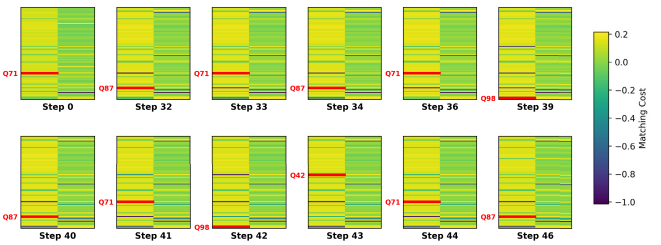


Fig. 5. Visualization of the cost matrix evolution during the attack. The red bars highlight the index of the optimal query assigned to the target object. As the attack increases the cost of the current query (e.g., Q71), the assignment shifts to a suboptimal query (e.g., Q87), which COA* dynamically re-identifies and tracks.

LAP (Eq.(4)) since the dual values give additional information about the optimal solution to the changes in the cost matrix,

$$\max_{\pi} \sum_{i=1}^{N} u_i + v_j \quad \text{s.t.} \quad u_i + v_j \le c_{ij} \tag{14}$$

where u_i, v_j are the dual variables. The dual variable u_i associates with the queries and v_j associates with the ground truth objects. They serve as the marginal value to quantify the contribution in the view from each query and ground-truth objects to the optimal solution. For min-cost assignment (i, j), $\pi_{ij} = 1$ and $u_i + v_j = c_{ij}$; for other (i, j), $\pi_{ij} = 0$ and $u_i + v_j \le c_{ij}$. The reduced cost, $c_{ij} - u_i + v_j$, actually determines the sensitivity of the optimal solution to the perturbations on the cost matrix.

For optimal matching of (i, j), increasing c_{ij} to $c_{ij} + \Delta c'_{ij}$ would affect the dual constraint $u_i + v_j \le c_{ij} + \Delta c'_{ij}$. If $\Delta c'_{ij}$ is negative, since (i, j) is already the min-cost matching, further decreasing the cost does not violate the dual feasibility. If $\Delta c'_{ij}$ is positive, the largest increase should be bounded by the smallest slack value of row-wise as well as column-wise reduced cost Δ_{row} and Δ_{col}; otherwise, the min-cost assignment is no longer optimal.

For other pairs (i, j), increasing c'_{ij} does not affect the optimal results because it makes them less attractive. On the other hand, the decrement of c'_{ij} should be bounded by $c_{ij} - (u_i + v_j)$; otherwise, the current optimal solution would be overridden by pair (i, j). ■

There are several takeaways from Theorem 1: 1) for target objects $i = T$, we need to impose a perturbation to induce $\Delta c'_{ij} > \min(\Delta_{\text{row}}, \Delta_{\text{col}})$ such that the min-cost matching no longer holds for each PGD step; 2) meanwhile, for non-target objects $i \neq T$, we need to have $\Delta c'_{ij} < c_{ij} - (u_i + v_j)$ so the optimal solutions for the rest coefficients remain intact, in order to avoid the spillover effect; 3) Otherwise, $\Delta c'_{ij}$ could have arbitrary values because it would not impact the optimality of the solution. This corresponds to the efforts of re-directing the harmful gradients, so that changes of the cost matrix entries in irrelevant directions would not impact the optimal matching for non-target objects.

1) Degeneracy:

Degeneracy occurs when there are multiple assignments that are optimal. In order for the bounds above to hold, it requires to examine the dual solutions from all these optimal solutions and make sure the tightest bounds are met. Denote the dual solutions as a set $(u, v) \in \{\mathcal{U}, \mathcal{V}\}$. Under degeneracy, the new bounds for optimal assignment (i, j) is

$$\Delta c'_{ij} = \begin{cases} \max\limits_{(u,v)\in\{\mathcal{U},\mathcal{V}\}} \min(\Delta_{\text{row}}, \Delta_{\text{col}}), & \Delta c'_{ij} \ge 0 \\ -\infty, & \Delta c'_{ij} < 0 \end{cases} \tag{15}$$

and for the rest pairs,

$$\Delta c'_{ij} = \begin{cases} +\infty, & \Delta c'_{ij} \ge 0 \\ \min\limits_{(u,v)\in\{\mathcal{U},\mathcal{V}\}} c_{ij} - (u_i + v_j). & \Delta c'_{ij} < 0 \end{cases} \tag{16}$$

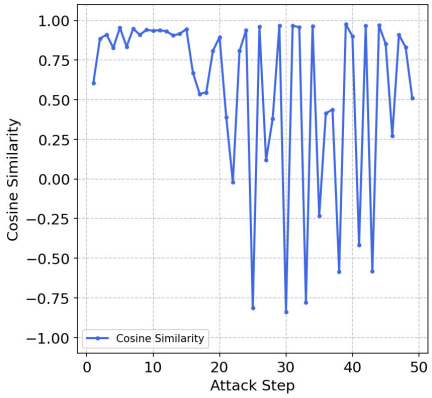


Fig. 6. Cosine similarity of the projected gradient directions between consecutive iterations. The optimization remains generally stable (similarity >0) for most steps, with fluctuations corresponding precisely to the query reassignment events in Fig. 5.

VI. Experiment

A. Experimental Setting

1) Datasets & Models: We conduct experiments on the standard dataset MS-COCO [48], which is a large and common benchmark for object detection. Our main results target four pre-trained DETR models including different CNN backbones and with/without a dilation module on the `coco2017val`. In order to examine DETR variants that extend towards one-to-many label assignment, we also conduct more experiments against three representative architectures: Deformable DETR [16], DINO [24], and RT-DETR [25] and compare the robustness with the one-to-one mapping strategy from the original DETR.

2) Benchmark and Settings: We evaluate the proposed attack in comparison with three SOTA attacks against object detection: DAG [49], TOG [50] and Attention-Fool [45] by adapting their objective functions to DETR.

- ✧ DAG aims to disrupt the categorical and positional information of each object by maximizing the difference between the attack and the original detection [49].
- ✧ TOG is a versatile adversarial framework to achieve different effects with various loss functions, including TOG-untargeted, TOG-targeted, TOG-vanishing, TOG-fabrication, and TOG-mislabeling. To align with our attack, we choose TOG-vanishing and TOG-mislabeling

TABLE I

MAIN RESULTS OF LAUNCHING THE PROPOSED ATTACK AGAINST DIFFERENT TARGETED CLASSES AND COMPARISON WITH THE BENCHMARKS[2]. A HIGHER FR ↑ INDICATES HIGHER ATTACK SUCCESS RATE, WHILE A LOWER SR ↓ REFLECTS LESS IMPACT ON NON-TARGET OBJECTS. THE TOP THREE VALUES IN THE MEAN COLUMN ARE MARKED IN BOLD

Model	Attack	Mean (↑/↓)	Top 10 classes as target (**FR ↑/SR ↓**)									
			person	car	chair	book	bottle	cup	dining table	traffic light	bowl	handbag
DETR-R50	DAG	.68/.22	.40/.20	.54/.19	.77/.23	.84/.17	.71/.23	.69/.26	.74/.23	.63/.25	.64/.26	.86/.15
	TOG-van	.65/.21	.33/.13	.57/.28	.73/.21	.80/.14	.64/.22	.65/.26	.75/.24	.56/.17	.65/.26	.81/.15
	TOG-mis†	**.70/.02**	.48/.05	.54/.01	.76/.02	.89/.01	.72/.01	.69/.01	.73/.01	.62/.02	.69/.01	.83/.01
	AttnFool‡	.35/**.01**	.12/.00	.24/.01	.38/.01	.60/.00	.32/.00	.34/.01	.39/.00	.23/.01	.31/.01	.56/.00
	COA	**.92**/.21	.88/.30	.95/.21	.94/.15	.86/.08	.92/.23	.93/.24	.94/.21	.88/.24	.94/.29	.93/.11
	COA*	**.87/.09**	.83/.11	.90/.07	.91/.05	.82/.04	.88/.10	.87/.11	.90/.09	.84/.12	.89/.13	.89/.04
DETR-R101	DAG	.59/.16	.34/.14	.47/.13	.62/.17	.81/.14	.57/.21	.59/.18	.66/.16	.55/.20	.57/.21	.72/.10
	TOG-van	.61/.19	.30/.12	.54/.19	.66/.20	.73/.14	.63/.27	.61/.19	.71/.21	.53/.19	.63/.27	.73/.13
	TOG-mis†	**.62/.02**	.39/.03	.44/.01	.67/.02	.86/.01	.64/.01	.63/.02	.69/.01	.49/.02	.61/.01	.77/.01
	AttnFool‡	.30/**.01**	.10/.00	.21/.01	.35/.01	.49/.01	.26/.00	.31/.01	.38/.01	.23/.00	.28/.00	.41/.01
	COA	**.89**/.20	.84/.27	.94/.22	.93/.17	.85/.11	.91/.24	.89/.22	.93/.18	.87/.24	.91/.27	.87/.09
	COA*	**.85/.09**	.80/.09	.88/.10	.87/.06	.83/.03	.87/.10	.85/.09	.89/.08	.81/.10	.86/.14	.88/.09
DETR-DC5-R50	DAG	.60/.16	.35/.15	.47/.13	.62/.15	.81/.14	.55/.19	.60/.20	.72/.18	.58/.16	.54/.20	.76/.12
	TOG-van	.60/.17	.30/.11	.54/.17	.71/.22	.76/.12	.56/.18	.62/.23	.71/.17	.50/.17	.59/.25	.73/.11
	TOG-mis†	**.65/.02**	.43/.03	.48/.02	.69/.02	.83/.01	.67/.01	.67/.01	.75/.01	.56/.03	.63/.02	.79/.01
	AttnFool‡	.31/**.02**	.11/.01	.19/.00	.35/.01	.54/.11	.26/.01	.27/.00	.40/.01	.22/.00	.27/.00	.47/.00
	COA	**.91**/.21	.82/.30	.94/.23	.94/.18	.86/.12	.91/.24	.93/.24	.93/.19	.91/.22	.94/.32	.92/.08
	COA*	**.85/.08**	.75/.08	.87/.09	.88/.10	.80/.03	.86/.09	.89/.09	.88/.08	.83/.11	.88/.11	.89/.08
DETR-DC5-R101	DAG	.57/.15	.38/.16	.44/.13	.61/.15	.80/.17	.51/.17	.54/.16	.61/.15	.50/.13	.56/.19	.71/.08
	TOG-van	.54/.16	.29/.09	.50/.16	.64/.20	.72/.10	.52/.21	.54/.19	.66/.21	.40/.13	.53/.20	.59/.09
	TOG-mis†	**.60/.01**	.36/.03	.45/.01	.65/.02	.85/.01	.56/.01	.62/.01	.66/.01	.53/.02	.60/.01	.74/.01
	AttnFool‡	.28/**.01**	.10/.00	.18/.01	.33/.01	.52/.01	.24/.01	.23/.00	.38/.01	.20/.00	.21/.00	.43/.01
	COA	**.88**/.24	.83/.30	.92/.24	.90/.18	.85/.10	.91/.28	.86/.33	.90/.23	.86/.27	.89/.33	.89/.10
	COA*	**.82/.09**	.73/.09	.85/.09	.82/.08	.81/.09	.84/.10	.82/.11	.85/.09	.81/.08	.81/.11	.85/.09

for comparison, where TOG-vanishing suppresses object detection by substituting the objectness loss and TOG-mislabeling changes the confidence of the target object [50].

- ✧ Attention-Fool is a new patch-based attack on the transformer attention mechanism [45]. More details of this attack are available in Sec. VI-H.

3) Attack Settings: To evaluate class-wise attack effects, we select the top most-frequent categories in MS-COCO as the target class. We set the perturbation budget to 1 in the L_∞-ball to ensure the attack intensity is sufficiently small to remain imperceptible to the human eye. For optimization, we employ the Adam optimizer with a learning rate of 0.01. All experiments are conducted on a workstation with 8× NVIDIA RTX 3090 GPUs (24GB VRAM), AMD EPYC 7402 CPU (24 cores), 480GB DDR4 RAM, Python 3.9, PyTorch 2.0.0, TorchVision 0.15.0 and CUDA 11.7.

4) Evaluation Metrics: We adopt the Fooling rate (FR) to evaluate the attack performance on target objects [51], which is defined as the ratio of the number of successfully attacked samples to the total number of samples in the target class. Meanwhile, to evaluate the spillover effects, we introduce an additional metric of spillover rate (SR) to evaluate the impact on non-target objects, when their matching losses change by a certain threshold, e.g., 5%–a spillover effect occurs. SR is defined as the ratio between the number of samples exhibiting spillover effects to the total number of samples in the target class.

B. Main Results

As summarized in Table I, the proposed Chosen Object Attack (COA) achieves SOTA fooling rates (FR↑) while maintaining controlled spillover rates (SR↓). We use COA* to denote our COA with the spillover mitigation strategy applied. The vanilla COA achieves a mean Fooling Rate (FR) of 0.9, outperforming the best-performing TOG-mis† by a large margin (0.64 FR). This improvement is particularly pronounced for small-object categories such as "cup" (FR 0.93) and "bowl" (FR 0.94), possibly due to their small size and lower feature saliency. COA* with spillover mitigation maintains competitive effectiveness (0.85 FR) while reducing the spillover rate (SR) from 0.22 to 0.09 with an optimal balance between attack success and spillover. Notably, the 0.17-0.23 FR gain over TOG-mis† comes at only a marginal SR cost (0.09 vs. 0.02) that escapes from a linear FR/SR trade-off. This disproportionality suggests that our proposed mitigation strategy is able to decouple target and non-target query interference more effectively.s.

1) Computational Cost and Stealthiness: The proposed attack brings additional benefits with much less computational cost and feature-level artifacts. Fig. 7a shows the trace of FR with the number of iterations. Compared to TOG-mis† and DAG, we can see that COA* achieves more than 6× speedups to reach an FR at 0.8 level. COA* also induces less artifacts on the feature level to evade anomaly checks

[2]TOG-van replaces the original objectness loss function with the $\mathcal{L}_{\text{class}}$. TOG-mis† adopts the same attack strategy as COA*. AttnFool‡ restricts adversarial patch to the L_∞-ball bounded space.

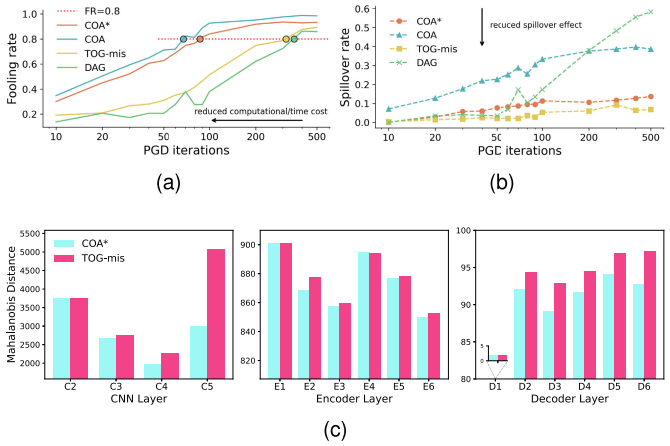


Fig. 7. (a,b) Comparative analysis of FR/SR changes with PGD iterations; (c) feature-level artifacts using the Mahalanobis distance (lower is better to evade detection).

[52]. Fig. 7b validates that compared to vanilla COA, COA* significantly mitigates the spillover effect, yielding an average 63% reduction in SR escalation across successive PGD iterations. Fig. 7c demonstrates layer-wise comparisons through the Mahalanobis distance for the feature-level variations across the CNN, encoder and decoder layers. We can see that TOG-mis induces significantly higher artifacts on the feature level, especially on the final CNN layer (C5) and all the decoder layers. This is because the existing methods of TOG-mis† often necessitate an attack plan [37] to reinforce correct predictions of non-target objects. When perturbed samples are input into the model, they activate more neurons that assist in detecting non-target objects, consequently inducing more significant feature variations.

C. Inside Attack Process

To provide algorithmic clarity on how the proposed *COA** executes the dynamic attack and maintains optimization stability, we analyze the internal dynamics of the query tracking and gradient projection mechanisms.

1) Dynamic Matching and Re-identification: A critical challenge in attacking DETR is the non-static nature of the target. Unlike anchor-based detectors where the target bounding box remains fixed to a grid, DETR's Hungarian matching allows the optimal query assignment to shift as the image context changes under perturbation. As illustrated in Fig. 5, we visualize the evolution of the matching cost matrix during the attack iterations. Initially (Step 0), the attack targets the optimal query (e.g., Q_{71}). As the adversarial perturbation increases the matching cost for this specific query to suppress the object, the Hungarian algorithm naturally seeks an alternative "best match" to maintain detection, shifting the assignment to a previously suboptimal query (e.g., Q_{87} at Step 32). Our attack framework accounts for this by re-calculating σ^*_δ at each PGD step. This allows the algorithm to dynamically "lock on" to the new optimal query, ensuring that the perturbation always suppresses the specific query currently responsible for the target object, rather than optimizing against a stale target.

TABLE II
IoU Degradation Under Dislocation Attack

Targeted class	person	car	chair	book	bottle
Initial IoU	0.79	0.66	0.62	0.52	0.65
IoU after COA*	0.61	0.50	0.51	0.45	0.44

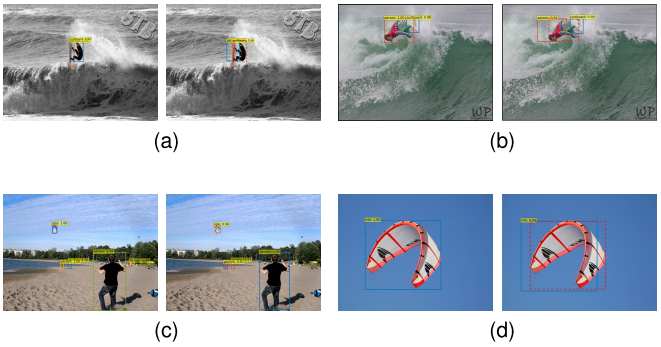


Fig. 8. Visualization of dislocation attack (lower IoU). (a, b) dislocate "person". (c, d) dislocate "kite".

2) Stability of Gradient Projection: A concern with gradient projection methods is whether removing the spillover component significantly weakens the attack or induces instability. We attribute the preservation of attack strength to the high-dimensional nature of the perturbation space. While the gradients of target and non-target objects often share similar directions due to global attention (as discussed in Section IV), they are rarely perfectly collinear. The projection operation in Eq. 10 removes only the component aligned with the spillover gradient, leaving a substantial orthogonal subspace available for the optimizer to maximize the target loss.

We empirically validate such stability in Fig. 6, which plots the cosine similarity of the projected update directions between consecutive iterations. The optimization exhibits high stability with positive cosine similarity for the vast majority of steps, confirming that the projection does not induce chaotic oscillations. Notably, we observe distinct fluctuations (sharp drops in similarity) at specific intervals (e.g., around Step 30-40). By cross-referencing with Fig. 5, these fluctuations correlate directly with the query reassignment events. When the target query switches, the loss landscape shifts, naturally leading to a directional change in the gradient. However, the optimizer quickly restabilizes after each switch, demonstrating the robustness of the projection mechanism even under dynamic matching conditions.

D. Dislocation Attacks

Beyond object hiding, we demonstrate the attack efficacy in the dislocation modality, which is achieved by optimizing the perturbation to maximize the L_{box} loss for the chosen object. Table II quantifies this effect, revealing a consistent and significant degradation in Intersection over Union (IoU) across various object classes under a minimal L_∞ perturbation budget of 1. For instance, the IoU for the "person" class drops sharply from 0.79 to 0.61, while the 'bottle' class sees its IoU reduced from 0.65 to 0.44. Fig. 8 provides a compelling visual confirmation of these quantitative results. The predicted

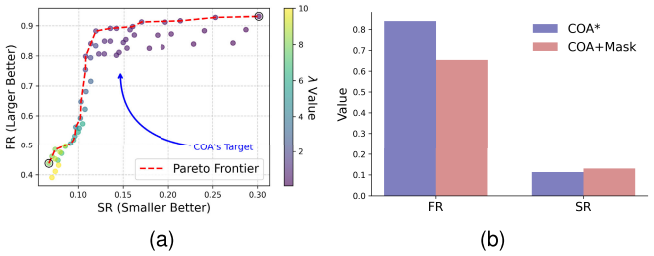


Fig. 9. Ablation studies: (a) Trade-off between FR/SR by adjusting λ to find Pareto-optimal solutions. With FR increases, we observe SR increases at a much slower marginal rate under COA*, which is preferred. (b) Comparison with spatial masking as an alternative to mitigate spillover (#person class).

TABLE III
FOOLING RATE OF BLACK-BOX ATTACKS ACROSS DIFFERENT ARCHITECTURES. $\mathcal{D}_{in}$ REFERS TO [55]

Surrogate \ Victim		R50	R101	DC5-R50	DC5-R101
R50	w/o $\mathcal{D}_{in}$	0.83	0.18	0.13	0.12
	$\mathcal{D}_{in}$	0.73	0.23	0.17	0.15
R101	w/o $\mathcal{D}_{in}$	0.20	0.80	0.19	0.11
	$\mathcal{D}_{in}$	0.24	0.72	0.23	0.15
DC5-R50	w/o $\mathcal{D}_{in}$	0.21	0.17	0.75	0.18
	$\mathcal{D}_{in}$	0.27	0.20	0.61	0.22
DC5-R101	w/o $\mathcal{D}_{in}$	0.17	0.22	0.20	0.73
	$\mathcal{D}_{in}$	0.18	0.24	0.23	0.62

bounding boxes for objects like "person" and "kite" are clearly displaced from their ground-truth locations, confirming that the attack successfully manipulates the model's spatial predictions. This type of subtle manipulation poses a critical threat to downstream applications that depend on precise localization, such as robotic grasping, object tracking, or path planning in autonomous vehicles, where even a minor offset can lead to catastrophic failure.

E. Ablation Studies

1) Spillover Mitigation and Tradeoffs: As mentioned above, Table I shows that spillover mitigation reduces spillover rates (SR) by 60% compared to vanilla COA (0.09 vs. 0.21 SR). To see the trade-offs between FR/SR, we plot the Pareto-frontier of different (FR, SR) pairs in Fig. 9a, while changing λ in Eq. (9). It is clear that as the FR goes up, the SR also increases, which makes the attack more conspicuous. E.g., if the attacker aims to achieve FR >0.9, he has to bear with an inevitable SR at 0.20. By adjusting λ (larger values bring less penalty to the SR), we can quickly find a Pareto-optimal region around $\lambda = 1 - 2$ with FR >0.8 and SR = 0.1.

2) Limitation of Spatial Masking: An alternative is to apply masks to adversarial noise by excluding non-target object regions and only retain target object areas for perturbation. While this approach seems intuitive, Fig. 9b shows that both FR and SR of spatial masking underperforms compared to COA*.[3] In addition, masking also requires ground-truth object localization during attack generation which brings additional overhead. The ineffectiveness of spatial masking indicates that the adversarial goal and spillover are less likely stemming from the pixel-space interference, but from query-level inter-dependencies in decoder. In other words, self-attention in DETR integrates global receptive fields to synthesize information from all image regions. Our results are consistent with visualization studies showing that attention maps exhibit high activation values not only around object centers but also in the peripheral areas, thus a trivial masking strategy would still have spillover effects [1], [53]. The proposed mitigation strategy directly addresses this root cause via gradient projections.

F. Black-Box Attacks

To assess practical viability beyond the white-box setting, we evaluate the attack's black-box transferability by crafting perturbations on a surrogate model and applying them to a victim [54]. As shown in Table III, our attack achieves significant fooling rates between 11-27%, demonstrating that the vulnerability is not an artifact of a specific model but is rooted in the shared one-to-one Hungarian matching mechanism fundamental to the DETR architecture. We observe that transferability is naturally higher between models with similar backbones (e.g., R50 to R101) and that integrating input diversity [55] consistently improves these rates by an average of 3.6%. These results underscore a crucial security implication: an adversary can leverage a public, pre-trained DETR model to mount an attack with a reasonable success probability against a proprietary, deployed system, confirming the practical threat of our method even with limited knowledge.

G. Unstable Matching From an Adversarial Perspective

1) Unstable Matching: Recall that our target query remains dynamic during the process of COA* given an input perturbation; otherwise, the attack performance would be diminished substantially. This echoes with the "unstable matching problem" in one-to-one label assignment that emerges during training [31], but from a different, adversarial perspective.

To validate this mechanism, we conduct an ablation study by selectively nullifying (zeroing out) vectors in the object query pool during inference. If one-to-one matching exhibits inherent stability during training, a fixed query would reliably predict its assigned object; otherwise, nullifying such a query would consequently eliminate detection of the corresponding target.

Fig. 10 demonstrates that nullifying up to three queries (row 3 in Fig 10c) does not degrade detection accuracy for the "stop sign" class, since redundant query pathways compensate for the deactivated ones. Complete failure occurs only when five queries are nullified, confirming that no single query is uniquely responsible for detecting an object and there is an inherent collection of queries that are responsible for the object, though the Hungarian loss imposes a one-to-one mapping. This empirically validates the necessity of our dynamic adaptation strategy, which iteratively targets the currently optimal query for each object.

2) Impact of Object Density: The degradation caused by query nullification correlates strongly with scene complexity, i.e., the object density in the input. Fig 10(c) demonstrates that nullifying three queries leads to the severe accuracy

[3]We demonstrate the class of person for simplicity, the results also hold for other classes.

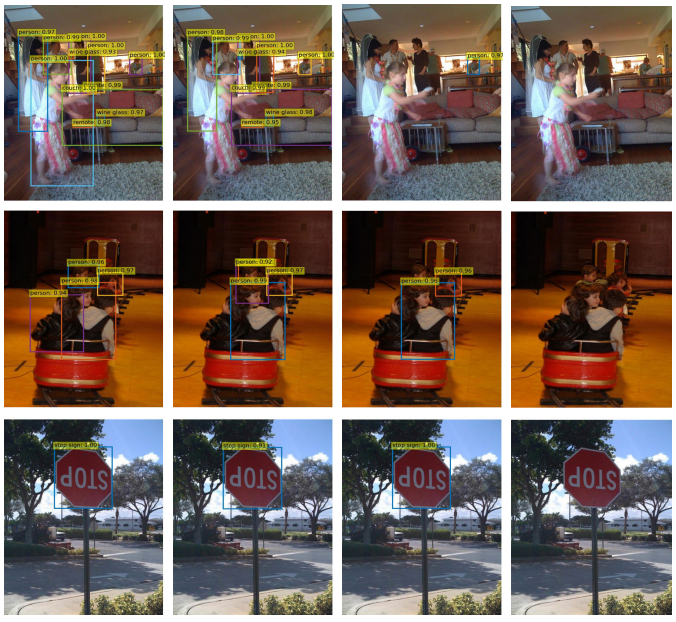


Fig. 10. Analysis of object query deactivation in DETR. Mask # queries mean current number of queries deactivated. (a) Init query. (b) Mask 1 query. (c) Mask 3 queries. (d) Mask 5 queries.

degradation (detecting only one object) in the scenes with high-density scenes. This indicates that with more number of objects, the query pool is exhausted with less redundancy, as unstable matching allocates most queries to specific object mappings. Thus, for low-density scenes, in which multiple redundant queries exist per object, it enables COA* to cycle through candidate queries during attack iterations; for low-density scenes, fewer redundant queries remain available, thus shifting COA* to reduce the spillover effect.

H. Synergizing With Patch-Based Attacks

While our core design focuses on pixel-space perturbations, we are also interested in investigating whether attention-based attacks could synergize with COA*. We adapt Attention-Fool [45]–which disrupts detection by amplifying dot-products between adversarial keys and all queries–to instead target only object-specific queries, while maintaining detection integrity in other regions.

$$\delta = \arg\max \mathcal{L}^{(1)}_{kq^*}(f(\mathbf{x} + \mathbb{1}_p \cdot \delta), \mathbf{y}) \tag{17}$$

where,

$$\mathcal{L}^{(1)}_{kq^*} = \frac{1}{\mathrm{N}_t} \sum_{h=1}^{H} \sum_{j \in \mathcal{Q}_t} \frac{Q_j^{h(1)} (K_{i^*}^{h(1)})^\top}{\sqrt{d_K}} \tag{18}$$

$$\mathcal{Q}_t = \left\{ j \in [n] \;\middle|\; \frac{|\mathcal{R}_j \cap \mathcal{R}_\mathrm{T}|}{|\mathcal{R}_j \cup \mathcal{R}_\mathrm{T}|} > 0 \right\} \tag{19}$$

Our primary adaptation lies in the definition of in Eq. (18). Since DETR lacks a class token, we redirect the loss function to maximize the dot-product between the target key and target object query set $\mathcal{Q}_t$. As shown in Eq. (19), queries with IoU greater than 0 for the target object are considered as target object queries. $\mathcal{R}_j$ and $\mathcal{R}_T$ denote the regions corresponding to query$_j$ and target object T. All other symbols retain the same definition as [45] so their illustrations are omitted due to space. Since no official implementation is available from Attention-Fool [45], we reproduce the attack according to the algorithm outlined in the original paper [45]. As shown in the main code snippet below, we introduce new global variables into the `multi_head_attention_forward` function to extract query (Q) and key (K) matrices before the softmax operation, and apply regularization techniques to mitigate variations across different attention heads. The attacks adopt PGD with step size 8/255, and 1000 iterations unless stated otherwise.

TABLE IV
ROBUSTNESS EVALUATION OF DETR WITH/WITHOUT DILATED CONVOLUTION UNDER ATTNFOOL AND ATTNFOOL+COA ATTACKS USING 64 × 64 PATCH SIZE

Model	Attack	FR/SR
DETR-R50	AttnFool	.24/.04
	AttnFool+COA	.22/.04
DETR-DC5-R50	AttnFool	.75/.53
	AttnFool+COA	.74/.51

From the results in Table IV, we conjecture that attention-based attacks may not be an ideal vehicle to carry out the proposed attacks as it either has high FR/SR or low FR/SR with a degraded balance. The performance gap between attacks on R50 and DC5-R50 is due to expansions of receptive fields in dilated convolution exacerbate the number of queries influenced by the adversarial the patch.

```
# nn.functional.py
def multi_head_attention_forward(...):
    ...
    if need_weights:
        # attention fool
        from patch_fool_detr import Q_global,
            K_global
        global Q_global, K_global
        Q_global.append(q_scaled)
        K_global.append(k)
    ...

# engine.py
def attention_fool():
    ...
    Q_l1, K_l1 = Q_global[0], K_global[0]
    ...
    # normalization for Q and K
    for head_idx in range(Q_l1.shape[0]):
        Q_h = Q_l1[head_idx]
        K_h = K_l1[head_idx]
        Q_l1_norm[head_idx] = Q_h / Q_h.norm(dim=-1,
            keepdim=True).mean()
        K_l1_norm[head_idx] = K_h / K_h.norm(dim=-1,
            keepdim=True).mean()
    ...
```

Listing 1. Code snippet of our re-implementation of Attention-Fool. `Q_global` and `K_global` update during inference.

Further, even when COA* is designed exclusively to alter attention towards target object queries, the spillover effect remains pronounced without observable reduction. Fig. 11 provides a visual illustration. The adversarial patch continues to attract attention from most queries rather than the solely

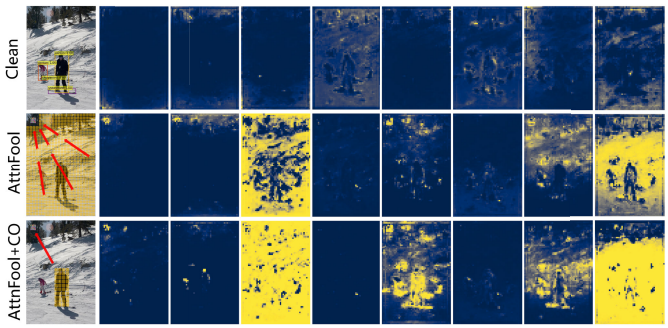


Fig. 11. Attention map visualization with adversarial patch analysis across attention heads: clean image patterns (row 1) vs. AttnFool (row 2) vs. AttnFool+COA (row 3) adversarial distributions. The visualization style is similar to that used in other works to illustrate attention outputs, such as in [1] and [34].

TABLE V

ROBUSTNESS EVALUATION OF DETR VARIANTS WITH ONE-TO-MANY LABEL ASSIGNMENTS (DEF*: DEFORMABLE DETR). DOUBLE ARROWS DENOTE FR DEVIATION FROM THE ORIGINAL ATTACK PERFORMANCE. STANDARD ARROWS REFLECT ALIGNMENT OF SR. LARGER DROP OF FR MEANS HIGHER ROBUSTNESS

Model	Backbone	Mean	Top 3 classes (**FR/SR**)		
			person	car	chair
Def*	R50	$.70_{\Downarrow.18}/.03_{\downarrow.04}$	.54/.04	.47/.04	.83/.01
	DC5-R50	$.48_{\Downarrow.35}/.05_{\downarrow.04}$	.48/.08	.22/.04	.75/.01
DINO	R50	$.67_{\Downarrow.21}/.03_{\downarrow.04}$	.50/.01	.61/.02	.89/.06
	Swin-L	$.56_{\Downarrow.00}/.02_{\downarrow.00}$	.36/.01	.55/.00	.77/.06
RT-DETR	R18	$.70_{\Downarrow.00}/.01_{\downarrow.00}$	.62/.01	.84/.01	.65/.01
	R34	$.59_{\Downarrow.00}/.01_{\downarrow.00}$	.48/.01	.58/.00	.70/.01
	R50	$.57_{\Downarrow.31}/.00_{\downarrow.07}$	.46/.00	.62/.01	.63/.00
	R101	$.61_{\Downarrow.24}/.01_{\downarrow.07}$	.47/.01	.64/.01	.72/.01

target ones. We posit that this originates from DETR's multi-layer encoder-decoder architecture: although the loss function specifically calculates the dot-product between target keys and target object queries, the forward propagation through multiple encoder and decoder layers causes complex entanglement between target and non-target object queries. This might amplify the spillover effect for attention-based attacks and make them less appealing in terms of stealthiness.

VII. POTENTIAL DEFENSE VIA ONE-TO-MANY MATCHING

Although adversarial training offers broad defense capabilities [46], we are interested in investigating inherent architectural protections within the DETR variants to avoid excessive training costs. Recall that the proposed attack exploits the brittle one-to-one mapping and hijacks such deterministic mapping. Our key insight is that one-to-many label assignments [16], [24], [25], originally designed to accelerate convergence and improve small object detection, also "inadvertently" enhance robustness against the proposed attack.

To see this, we implement COA* against these new DETR variants with one-to-many mappings in Table V. Specifically, Deformable DETR increases query # from 100 to 300 through multi-scale feature fusion, DINO uses 900 queries, and RT-DETR similarly adopts 300 queries. As shown in Table V, one-to-many mapping leads to lower FR/SR, indicating higher robustness against COA*, as the attacker requires to manipulate a much larger query pool for equivalent efficacy. On the other hand, successful attacks produce less spillover effect. We attribute these phenomena to two factors: 1) one-to-many matching further obscures the one-to-one query-object binding as a basis for launching COA*, thereby reducing the overall FR; 2) for SR, the increased number of queries decouples inter-query relationships, making it harder to affect non-target objects when attacking specific queries. This occurs because non-target objects are determined through optimal query selection from multiple candidates, requiring simultaneous manipulation of numerous unrelated queries to induce spillover effects.

TABLE VI

COMPARISON OF DEFENSE STRATEGIES AGAINST COA* ON MS-COCO. RS: RANDOMIZED SMOOTHING WITH VARYING SAMPLE SIZE N

Defense Method	Attack Performance		Clean Performance	
	FR	**SR**	mAP	ΔmAP
None (Standard DETR)	.88	.07	42.0	-
RS ($\sigma = 0.10, N = 5$)	$.90_{\Uparrow.02}$	$.20_{\uparrow.13}$	39.6	-2.4%
RS ($\sigma = 0.10, N = 10$)	$.86_{\Downarrow.02}$	$.17_{\uparrow.10}$	39.9	-2.1%
RS ($\sigma = 0.25, N = 5$)	$.91_{\Uparrow.03}$	$.23_{\uparrow.16}$	36.1	-5.9%
RS ($\sigma = 0.25, N = 10$)	$.87_{\Downarrow.01}$	$.19_{\uparrow.12}$	36.6	-5.4%
Deformable DETR	$.70_{\Downarrow.18}$	$.03_{\downarrow.04}$	43.8	+1.8%
DINO	$.67_{\Downarrow.21}$	$.03_{\downarrow.04}$	50.4	+8.4%

A. Comparison With Randomized Smoothing

To systematically evaluate the robustness of this inherent defense, we compare our One-to-Many matching finding with Randomized Smoothing (RS) [56], a certified defense method. Following [56], we construct a smoothed classifier $g(x)$ from the base detector f. The prediction is derived by identifying the class c with the maximum probability under isotropic Gaussian noise perturbations $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. We apply RS with $\sigma \in \{0.10, 0.25\}$ to the standard pre-trained DETR-R50 without retraining (to maintain consistency with the training-free nature of our One-to-Many defense comparison).

As shown in Table VI, the training-free RS strategy fails to effectively defend against the attack, with the fooling rate remaining high or even increasing. We attribute this to the significant distribution shift induced by the added noise. In stark contrast, the "natural defense" of one-to-many matching (Deformable DETR & DINO) provides immediate, superior robustness (FR decreases by 0.18-0.21) and improved clean performance, confirming that architectural redundancy is a far superior strategy compared to standard input-level smoothing in this context.

B. Mechanism Analysis

To understand the rationale behind this robustness, Fig. 12 shows significantly lower performance degradation in Deformable DETR compared to DETR under identical query deactivation conditions. Specifically, when five queries are

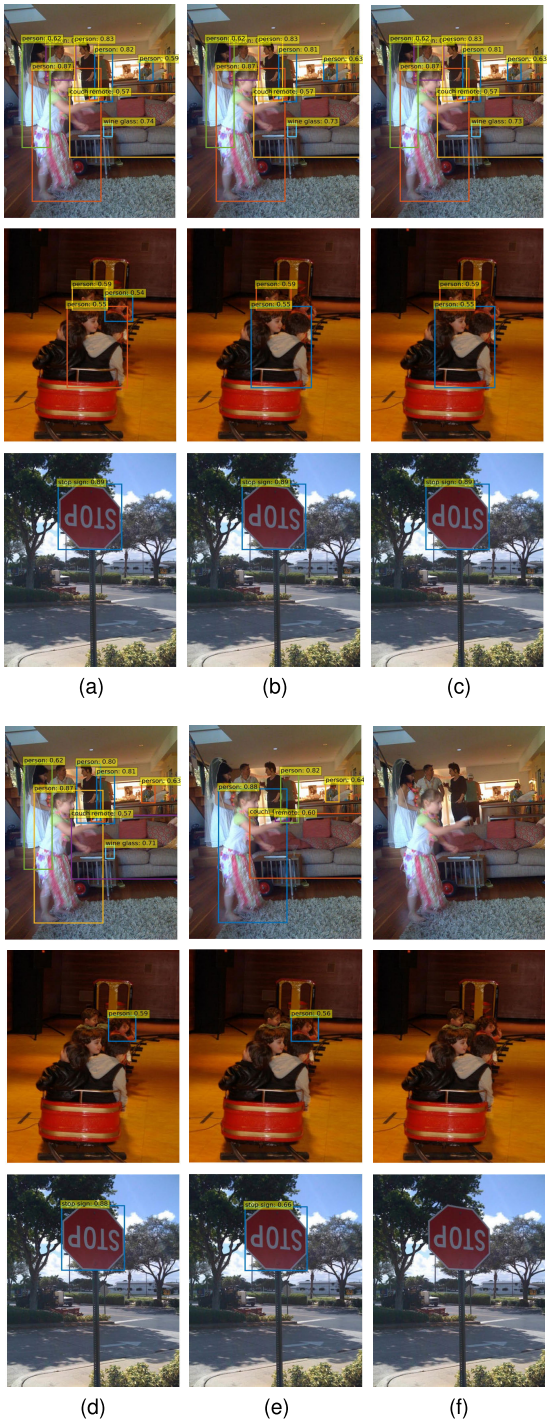


Fig. 12. Analysis of object query deactivation in Deformable DETR. (a) Init q. (b) Mask 1 q. (c) Mask 3 q. (d) Mask 5 q. (e) Mask 10 q. (f) Mask 50 q.

deactivated, Deformable DETR retains partial object detection capabilities via its redundant prediction pathways, whereas standard DETR fails completely under identical conditions. This comparative analysis demonstrates the inherent robustness in one-to-many matching paradigms. In other words, the redundancy in one-to-many matching complicates chosen-object attacks, as disabling a single query rarely suppresses the corresponding target. This necessitates adversarial perturbations to corrupt *all* redundant queries associated with the target. We leave the exploration of such escalated attacks against one-to-many matching to the future works.

VIII. Limitations and Future Work

Note that the proposed attack is analyzed mainly under a white-box setting, which assumes the adversary has full access to the model architecture and parameters. While we demonstrate black-box transferability, this remains a challenging area. We also note that our core methodology focuses on pixel-space perturbations, and while we explore synergies with patch-based attacks, developing physically robust patches based on our findings is a non-trivial future step. Further, extending the COA* to other black-box scenarios such as query-based attacks against an object detection API is a non-trivial but critical direction.

Our work provides initial evidence that DETR variants with one-to-many label assignments (e.g., Deformable DETR, DINO) exhibit inherent robustness against our proposed attack. This presents a critical area for further investigation, both as a potential defense mechanism and, conversely, as a target for more sophisticated, escalated attacks designed to overcome this redundancy. Furthermore, given that DETR models are computationally intensive, a valuable direction is to explore hybrid attack modalities. For instance, combining our precision-based Chosen-Object Attack (which targets accuracy) with computational-cost attacks (e.g., latency attacks [57], [58]) could create a multi-faceted threat that simultaneously degrades both the detection integrity and the real-time performance of such systems. We propose that future work could explore query-efficient optimization techniques to exploit the identified Hungarian matching vulnerability without full model access.

IX. Conclusion

In this paper, we exploit the Hungarian matching mechanism for targeted attacks against DETR. We find that a naive implementation could induce unintended spillover effects on non-target objects. To address this issue, we propose a gradient projection strategy to isolate perturbations on queries. The extensive experiments demonstrate effectiveness of the proposed attacks with new insights of potential robustness from one-to-many label assignments and attention-based extensions. By targeting DETR families from an adversarial perspective, we hope to inspire further research of detection transformers into their joint efficiency and robustness in the future.

References

[1] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.
[2] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1–6.
[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[4] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[5] (2022). *YOLOv5*. Accessed: Aug. 26, 2024. [Online]. Available: https://github.com/ultralytics/yolov5

[6] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9626–9635.

[7] J. Ouyang-Zhang, J. Hyun Cho, X. Zhou, and P. Krähenbühl, "NMS strikes back," 2022, *arXiv:2212.06137*.

[8] S. H. Rezatofighi, B. G. V. Kumar, A. Milan, E. Abbasnejad, A. Dick, and I. Reid, "DeepSetNet: Predicting sets with deep neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5257–5266.

[9] L. Pineda, A. Salvador, M. Drozdzal, and A. Romero, "Elucidating image-to-set prediction: An analysis of models, losses and datasets," 2019, *arXiv:1904.05709*.

[10] Y. Zhang, J. Hare, and A. Prugel-Bennett, "Deep set prediction networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 3212–3222.

[11] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logistics Quart.*, vol. 2, nos. 1–2, pp. 83–97, 1955.

[12] Q. Zhou and C. Yu, "Object detection made simpler by eliminating heuristic NMS," *IEEE Trans. Multimedia*, vol. 25, pp. 9254–9262, 2023.

[13] A. Shapira, A. Zolfi, L. Demetrio, B. Biggio, and A. Shabtai, "Phantom sponges: Exploiting non-maximum suppression to attack deep object detectors," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 4560–4569.

[14] E.-C. Chen, P.-Y. Chen, I.-H. Chung, and C.-R. Lee, "Overload: Latency attacks on object detection for edge devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 24716–24725.

[15] G. Zhang, Z. Luo, Y. Yu, K. Cui, and S. Lu, "Accelerating DETR convergence via semantic-aligned matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 949–958.

[16] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent.*, 2020.

[17] C. Zhao et al., "MS-DETR: Efficient DETR training with mixed supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 17027–17036.

[18] Z. Yao, J. Ai, B. Li, and C. Zhang, "Efficient DETR: Improving end-to-end object detector with dense prior," 2021, *arXiv:2104.01318*.

[19] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "DN-DETR: Accelerate DETR training by introducing query denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 13619–13627.

[20] D. Meng et al., "Conditional DETR for fast training convergence," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3651–3660.

[21] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 5824–5836.

[22] E. Michael, T. A. Wood, C. Manzie, and I. Shames, "Global sensitivity analysis for the linear assignment problem," in *Proc. Amer. Control Conf. (ACC)*, Jul. 2020, pp. 3387–3392.

[23] C.-J. Lin and U.-P. Wen, "Sensitivity analysis of the optimal assignment," *Eur. J. Oper. Res.*, vol. 149, no. 1, pp. 35–46, Aug. 2003.

[24] H. Zhang et al., "DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent.*, 2022.

[25] Y. Zhao et al., "DETRs beat YOLOs on real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 16965–16974.

[26] M. Lipp et al., "Meltdown: Reading kernel memory from user space," in *Proc. 27th USENIX Secur. Symp. (USENIX Secur.)*, Aug. 2018, pp. 973–990. [Online]. Available: https://www.usenix.org/conference/usenixsecurity18/presentation/lipp

[27] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. ECCV*, Oct. 2016, pp. 21–37.

[28] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[29] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[30] T. Vu, H. Jang, T. X. Pham, and C. D. Yoo, "Cascade RPN: Delving into high-quality region proposal network with adaptive convolution," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1430–1440.

[31] S. Liu et al., "Detection transformer with stable matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6468–6477.

[32] D. Jia et al., "DETRs with hybrid matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19702–19712.

[33] Q. Chen et al., "Group DETR: Fast DETR training with group-wise one-to-many assignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6610–6619.

[34] Q. Chen, L. Wang, P. Koniusz, and T. Gedeon, "Motion meets attention: Video motion prompts," in *Proc. 16th Asian Conf. Mach. Learn.*, 2024, pp. 591–606.

[35] A. Saha, A. Subramanya, K. Patil, and H. Pirsiavash, "Role of spatial context in adversarial robustness for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 784–785.

[36] Y. Jia et al., "Fooling detection alone is not enough: Adversarial attack against multiple object tracking," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

[37] J. Bao, B. Liu, K. Ren, and J. Yu, "GLOW: Global layout aware attacks on object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 12057–12066.

[38] P. Zhao et al., "An attack-agnostic defense framework against manipulation attacks under local differential privacy," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2025, pp. 3858–3876.

[39] J. Feng, Y. Lai, H. Sun, and B. Ren, "SADBA: Self-adaptive distributed backdoor attack against federated learning," in *Proc. AAAI Conf. Artif. Intell.*, 2025, vol. 39, no. 16, pp. 16568–16576.

[40] K. Mahmood, R. Mahmood, and M. van Dijk, "On the robustness of vision transformers to adversarial examples," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7838–7847.

[41] M. M. Naseer et al., "Intriguing properties of vision transformers," in *Proc. NIPS*, vol. 34, 2021, pp. 23296–23308.

[42] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, and A. Veit, "Understanding robustness of transformers for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10231–10241.

[43] N. Park and S. Kim, "How do vision transformers work?," in *Proc. ICLR*, 2022.

[44] Y. Fu, S. Zhang, S. Wu, C. Wan, and Y. Lin, "Patch-fool: Are vision transformers always robust against adversarial perturbations?," in *Proc. ICLR*, 2022.

[45] G. Lovisotto, N. Finnie, M. Munoz, C. K. Murnmadi, and J. H. Metzen, "Give me your attention: Dot-product attention considered harmful for adversarial patch robustness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15213–15222.

[46] A. Mcadry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. ICLR*, 2017.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2014.

[48] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[49] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, "Adversarial examples for semantic segmentation and object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1369–1378.

[50] K.-H. Chow et al., "Adversarial objectness gradient attacks in real-time object detection systems," in *Proc. 2nd IEEE Int. Conf. Trust, Privacy Secur. Intell. Syst. Appl. (TPS-ISA)*, Oct. 2020, pp. 263–272.

[51] Z. Cai et al., "Zero-query transfer attacks on context-aware object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15024–15034.

[52] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 7167–7177.

[53] Facebook Research.(2020). *Detr Hands on*. Accessed: Feb. 9, 2025. [Online]. Available: https://tinyurl.com/56a43d7b

[54] Y. Xiao and C. Wang, "You see what i want you to see: Exploring targeted black-box transferability attack for hash-based image retrieval systems," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1934–1943.

[55] C. Xie et al., "Improving transferability of adversarial examples with input diversity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2730–2739.

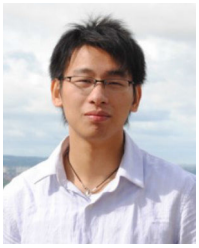
[56] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 1310–1320.

[57] I. Shumailov, Y. Zhao, D. Bates, N. Papernot, R. Mullins, and R. Anderson, "Sponge examples: Energy-latency attacks on neural networks," in *Proc. IEEE Eur. Symp. Secur. Privacy*, Sep. 2021, pp. 212–231.

[58] T. Wang et al., "Can't slow me down: Learning robust and hardware-adaptive object detectors against latency attacks for edge devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 19230–19240.

Tianyi Wang (Graduate Student Member, IEEE) received the B.S. degree in automation from Zhengzhou University, China, in 2020. He is currently pursuing the Ph.D. degree with the College of Control Science and Engineering, Zhejiang University. His research interests include adversarial ML and LLM safety.

Cong Wang (Member, IEEE) received the B.Eng. degree in information engineering from The Chinese University of Hong Kong in 2008, the M.S. degree in electrical engineering from Columbia University, NY, USA, in 2009, and the Ph.D. degree in computer and electrical engineering from Stony Brook University, NY, USA, in 2016. He was a tenure-track Assistant Professor of computer science with Old Dominion University and George Mason University, VA, USA, from 2017 to 2023. He is currently an Associate Professor with Zhejiang University, China. His works have been published in CVPR, KDD, AAAI, IJCAI, INFOCOM, CCS, ACM MM, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE TRANSACTIONS ON MOBILE COMPUTING, and IEEE TRANSACTIONS ON COMPUTERS. His research interests include distributed systems and AI security. He was a recipient of the NSF CAREER Award in 2021. He serves as an Associate Editor for IEEE TRANSACTIONS ON CLOUD COMPUTING.



Zhenyu Wen (Senior Member, IEEE) received the Ph.D. degree from Newcastle University, U.K., in 2016. He is currently a Professor with the Institute of Cyberspace Security and the College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. He has published at top venues, including *ACM Computing Survey*, KDD, ICDE, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VLDB, IEEE TRANSACTIONS ON COMPUTERS. His research interests include AI systems and cloud computing. He received the IEEE TCSC Award for Excellence in Scalable Computing (Early Career Researchers) in 2020.

Ruilong Deng (Senior Member, IEEE) received the B.Sc. and Ph.D. degrees in control science and engineering from Zhejiang University in 2009 and 2014, respectively. He was a Research Fellow with Nanyang Technological University, Singapore, from 2014 to 2015; an AITF Post-Doctoral Fellow with the University of Alberta, Edmonton, AB, Canada, from 2015 to 2018; and an Assistant Professor with Nanyang Technological University from 2018 to 2019. Currently, he is a Professor with the College of Control Science and Engineering, Zhejiang University, and the Deputy Director of the State Key Laboratory of Industrial Control Technology. His research interests include smart grid, cybersecurity, and control systems. He serves/served as an Associate Editor for IEEE TRANSACTIONS ON SMART GRID, IEEE POWER ENGINEERING LETTERS, IEEE/CAA JOURNAL OF AUTOMATICA SINICA, and IEEE/KICS JOURNAL OF COMMUNICATIONS AND NETWORKS.

Yuanchao Shu (Senior Member, IEEE) received the Ph.D. degree from Zhejiang University. He was also a joint Ph.D. Student at the EECS Department, University of Michigan, Ann Arbor. He is currently a Qiushi Professor with the College of Control Science and Engineering, Zhejiang University, China. Prior to joining academia, he was a Principal Researcher with Microsoft Research Redmond and Microsoft Azure. His research interests include mobile, sensing, and networked systems. He has published over 70 papers at top-tier peer-reviewed conferences and journals. He is a Senior Member of ACM. He serves on the editorial board for IEEE TRANSACTIONS OF WIRELESS COMMUNICATIONS and *ACM Transactions on Sensor Networks*, the Vice General Chair for ACM SenSys'24, and a member for the Organizing Committee and TPC of conferences, including MobiCom, MobiSys, SenSys, SEC, Globecom, and ICC. He won five best paper/demo (runner-up) awards, the MobiCom Best Community Contribution Award, the ACM China Doctoral Dissertation Award, and the IBM Ph.D. Fellowship.

Peng Cheng (Member, IEEE) received the B.Sc. and Ph.D. degrees in control science and engineering from Zhejiang University in 2004 and 2009, respectively. He is currently the Dean and the Changjiang Chair Professor with the College of Control Science and Engineering, Zhejiang University. He has published in leading conferences, such as SIGCOMM, SP Oakland, CCS, USENIX Security, and NDSS. His research interests include control system security, cyber-physical systems, and cloud networking. He has received the State Science and Technology Progress Award and the MOE Natural Science Award. He serves/served as an Associate Editor for IEEE TRANSACTIONS ON CONTROL OF NETWORK SYSTEMS and IEEE TRANSACTIONS ON CLOUD COMPUTING. He also served as a Guest Editor for IEEE TRANSACTIONS ON AUTOMATIC CONTROL and IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS.

Jiming Chen (Fellow, IEEE) received the bachelor's and Ph.D. degrees in control science and engineering from Zhejiang University. From 2008 to 2010, he was a Visiting Scholar at the University of Waterloo, Canada. He has been a Professor at Zhejiang University since 2010, where he is currently the Deputy Director of the State Key Laboratory of Industrial Control Technology and the Director of the Institute of Industrial Process Control. He also served as the Vice President of Zhejiang University of Technology, the Deputy Director of the Informatics Department, Zhejiang University, and a member of both the Academic and Degree Committees of Zhejiang University. He is also a Professor at Zhejiang University and the President of Hangzhou Dianzi University. He is a fellow of Chinese Association of Automation and selected as a Distinguished Expert of Zhejiang Province in 2021. He was selected for the Ministry of Education's Changjiang Scholars Award Program in 2015. He is the Distinguished Lecturer/Speaker of the IEEE Vehicular Technology Society. He was a recipient of the IEEE ComSoc Asia/Pacific Outstanding Paper Award, the JSPS Invitation Fellowship, the RS Newton Advanced Fellowships, and multiple best paper awards. He was the General Co-Chair of multiple IEEE/ACM conferences, such as ACM Sensys 2024. He serves/served as an Associate Editor for premier journals of ACM TECS, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, *IEEE Network*, IEEE TRANSACTIONS ON CONTROL OF NETWORK SYSTEMS, and IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS. He is the Editor-in-Chief of *IEEE Network*.