Supplementary Material for "Test-Time Backdoor Detection for Object Detection Models"

Roadmap of Appendix: The Appendix is organized as follows. We list the notations table in Section A. Detailed background and related work are presented in Section B. The details of the experimental settings are presented in Section C. Additional experimental results are presented in Section D. Detailed insights and comprehensive analyses are provided in Section E.

A. Notation Table

Notations	Meaning
$\mathbb{F}_{\boldsymbol{\theta}}$	object detection model
$\boldsymbol{x}$	input sample
$\boldsymbol{o}_i$	object i
$\boldsymbol{y}_i$	label for object $\boldsymbol{o}_i$
t	trigger
$\boldsymbol{y}_t$	backdoor target class
$\hat{\boldsymbol{o}}_l$	the victim object
$\hat{\boldsymbol{a}} = (\hat{\boldsymbol{o}}, \hat{\boldsymbol{y}})$	ground-truth annotations
τ	SSIM module threshold
$\mathcal{B}$	background set
b	number of background queries
f	number of foreground queries
k	Monte Carlo sampling point
α_{bg}	the opacity of background images
δ	background image
γ	decision threshold for TRACE
$\Delta_{\mathcal{B}}^{\text{Var}}$	contextual transformation consistency value
$\Delta_{\mathcal{F}}^{\text{Var}}$	focal transformation consistency value
$\boldsymbol{o}_{y_i}^{\text{ref}}$	universal visual benchmark for class y_i
NBO	natual backdoor object
TP	true-positive
FP	false-positive
FN	false-negative

B. Background and Related Work

B.1 Backdoor Attack and Defense.

Backdoor attack poses a training-time threat to *deep neural networks* (DNNs) [11, 24], making deep learning-based object detectors similarly susceptible to such attacks. It aims to inject covert malicious behavior into a victim model, triggered by a specific pattern (*e.g.,* an image patch), to controllably manipulate the victim model. Compared to backdoor attacks on prevalent tasks (*e.g.,* image or NLP classification), the unique characteristics of object detectors allow for more diverse and complex attack effects.

Backdoor defenses include model diagnosis defense [6, 12, 26], trigger reverse engineering [5, 23], and *test-time trigger sample detection* (TTSD). In this work, we prioritize the practical application of the black-box TTSD method as a final safeguard when employing models of unknown credibility, particularly without authority to access training data or model parameters.

B.2 Backdoor Attacks against Object Detection Models.

Backdoor attacks against object detector including object appearing (*e.g.,* `OGA` [2]), object disappearance (*e.g.,* `ODA` [2] and `UTA` [19]), (global) object misclassification (*e.g.,* `RMA` [2] and `GMA` [2]), and even the simultaneous appearance of two clean objects triggering misclassification of the victim object (*e.g.,* `CIB` [3] and `Composite` [15] attacks). `DC` [30] injects a backdoor by controlling the training process of the model, enabling the disappearance of all objects whenever the trigger appears. A detailed description of these attacks is provided in Appendix C.6.

B.3 Test-time trigger sample detection methods.

TTSD methods have demonstrated excellent defensive capabilities in classification tasks. `Strip` [9] detects backdoors by analyzing the entropy of model outputs when inputs are overlaid with perturbations. `Teco` [18] identifies triggers by evaluating the consistency of model predictions under various corruptions. `FreqDetector` [28] use frequency-domain analysis to spot anomalies introduced by triggers. Even though they can generalize to object detection, their performance remains poor. Devising a TTSD strategy tailored to this task is challenging due to the effects of discrete backdoor attacks. `Detector Cleanse` *represents a rudimentary step, adopting a* `Strip`*-like approach, overlaying clean features on objects to observe the entropy of predicted bboxes.* However, we find it unrealistically assumes knowledge of attacks from a god's-eye view (*i.e.,* using different criteria for different attacks), and exhibits unsatisfactory detection accuracy. This fact underscores the challenge of establishing an effective and unified

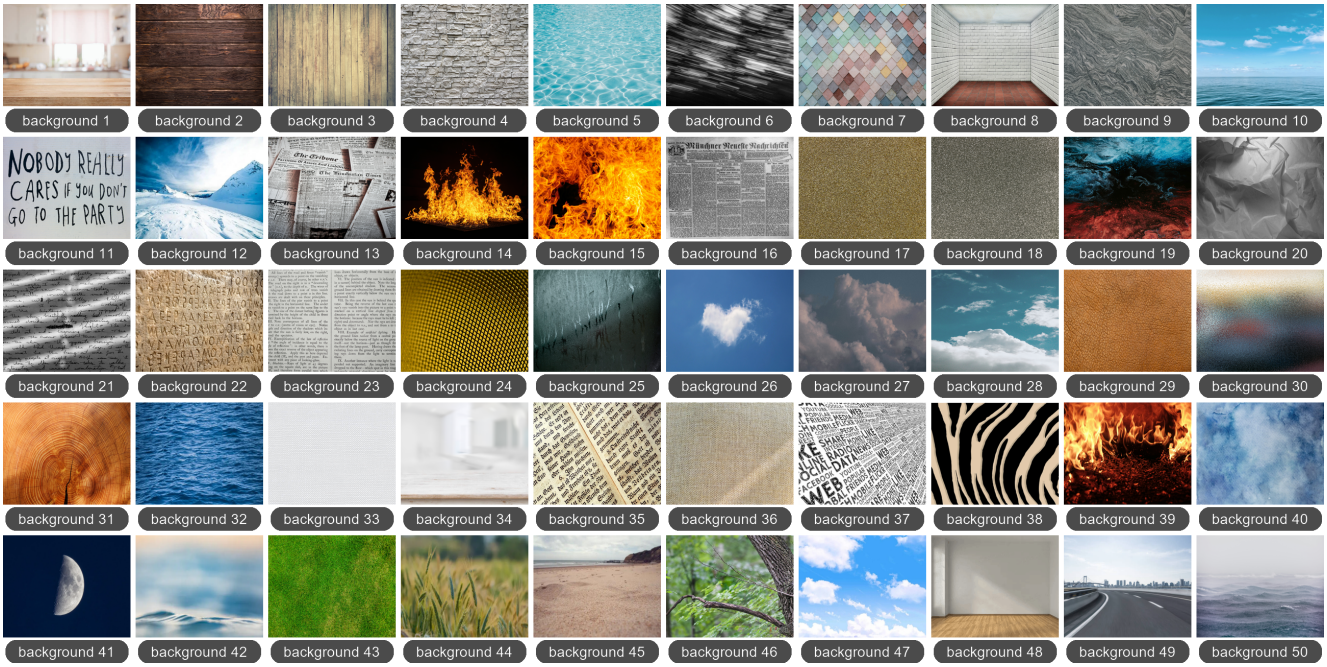

Figure 1. An overview of our background image set, showcasing diverse contexts and styles.

TTSD defense. Our TRACE gracefully leverages the inherent domain knowledge of object detection to effectively resolve this dilemma.

C. Complete Experiment Details

In this section, we provide a comprehensive overview of the experimental setup, including detailed parameters and configurations, to ensure the clarity and reproducibility of our results.

C.1 Experimental Hardware Details

We conduct experiments using PyTorch 2.1.1 and Python 3.8 on a machine with four NVIDIA GeForce RTX 4090 GPUs, an Intel (R) Xeon (R) Silver 4210R CPU and 256GB RAM.

C.2 Details of Datasets and Models

Datasets. We select Microsoft Common Objects in COntext (MS-COCO) [16], PASCAL Visual Object Classes (VOC) [8], and Synthesized Traffic Sign dataset [20], which are widely employed in existing works.

MS-COCO dataset is widely used for object detection tasks. It contains 80 object categories, including people, animals, vehicles, and more. Each image can contain multiple instances of objects, providing ample opportunities for training and evaluating models capable of detecting and segmenting objects in complex scenes. It represents a challenging benchmark in the field of object detection. We used the `COCO2017` split for training (118,000 images) and validation (5,000 images). **VOC** is a well-known dataset that provides annotations for 20 object categories. Consistent with the conventional usage [17, 29], we combine the `trainval2007` set of 5,000 images with the `trainval2012` set of 11,000 images for training and test on the `test2007` set of 5,000 images. **Synthesized Traffic Sign** dataset is designed by TrojAI [20] which focuses on traffic sign detection, featuring various types of traffic signs commonly encountered in real-world scenarios.

Models. The detectors we choose are YOLOv5-s [21] with the CSPDarknet-53 feature extractor, representing the one-stage object detector, along with Faster R-CNN [22] with the ResNet-50 backbone, representing the two-stage object detector. We also validated our approach on the advanced transformer-based architecture DETR (DEtection TRansformer) [1].

YOLOv5 is one of the most popular one-stage object detectors that achieves a notable equilibrium between detection accuracy and processing speed. The core idea of YOLOv5 lies in utilizing the whole image as the network's input, wherein the image is segmented into various regions, and the network directly outputs the positions of bounding boxes along with their corresponding labels in the output layer.

DETR is a groundbreaking object detection model that eliminates the need for traditional components like region proposal networks and non-maximum suppression by di-

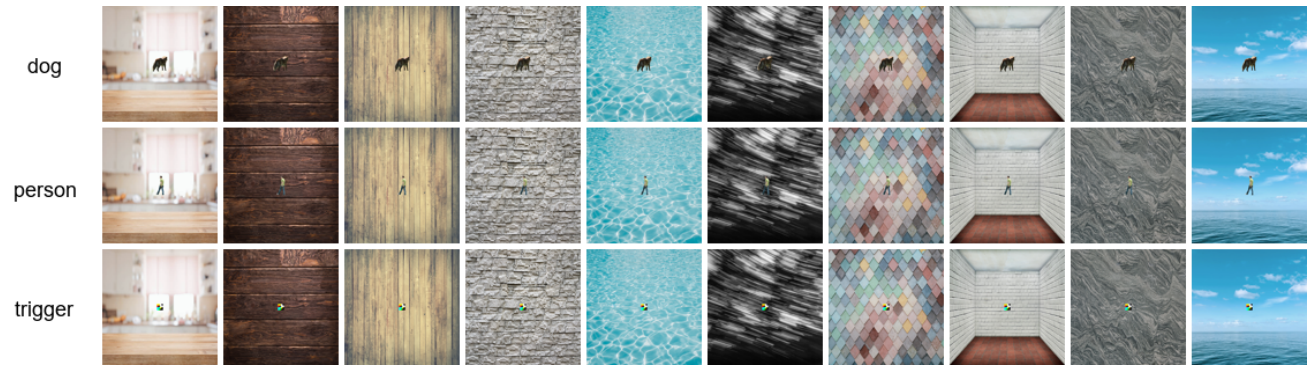

Figure 2. In our exploratory analysis, we use semantic segmentation algorithms to extract three objects (*i.e.,* dog, person, and trigger) individually and paste them onto different backgrounds. Subsequently, we employed three object detectors to evaluate their confidence performance across varying backgrounds.

rectly predicting a set of objects. It leverages an encoder-decoder transformer architecture to model global relationships among image features, making it highly effective for object detection and segmentation tasks.

Faster-RCNN streamlines object detection by using a *Region Proposal Network* (RPN) to generate candidate regions, followed by a CNN module for feature extraction. This approach reduces computational costs and improves the efficiency of candidate frame generation.

Training details. YOLOv5 is trained using SGD with an initial learning rate of $1e^{-2}$, while Faster R-CNN is trained using SGD with an initial learning rate of $1e^{-3}$. We set the input image size for YOLOv5 to 640×640, and for Faster R-CNN and DETR, the shorter side of the input images was resized to 800 pixels, maintaining the aspect ratio, with the longer side capped at 1333 pixels. We trained DETR for 200 epochs using AdamW as the optimizer, setting the learning rate to $1e^{-4}$ for the transformer and $1e^{-5}$ for the backbone. Horizontal flips, scaling, and cropping were applied for data augmentation. The transformer is trained with a dropout rate of 0.1, and gradient clipping with a threshold of 0.1 is applied to stabilize training. The training procedures for the rest of the setup followed the standard practices for YOLOv5 [21], Faster R-CNN [22] and DETR [1].

C.3 Exploratory Analysis on Trigger Behavior Across Backgrounds

In our exploratory analysis (see Sec. 2 of the manuscript), we replaced the backgrounds of "person," "dog", and a trigger object labeled as "person" with various scenes. Specifically, the "person" and "dog" objects are obtained using a semantic segmentation algorithm (*i.e.,* SAM [13]). These patches are then overlaid onto the background images numbered 1–10 in Fig. 1, with the visualization results shown in Fig. 2. Experiments are conducted on three object detectors: YOLO, Faster R-CNN, and DETR. Results show that only the trigger object consistently maintained high confidence scores across all 10 backgrounds and three detectors, while the clean objects exhibited significant variations in their confidence scores.

C.4 Visualizations of Contextual Information Transformation

We curated a collection of background images from the internet to facilitate TRACE's contextual information transformation. Fig. 1 illustrates our collection of background images, which encompasses a diverse range of contexts and styles. Fig. 4 demonstrates how our TRACE leverages these backgrounds for contextual information transformation, blending the backgrounds with the input images at a specified opacity α_{bg}. Crucially, these backgrounds are unrelated to any model's training datasets, and we consider this to be a reasonable assumption since they are ubiquitous in everyday life and publicly accessible to anyone (including defenders).

C.5 Visualizations of Universal Visual Benchmarks

Our universal visual benchmarks are derived from publicly available online data corresponding to the categories in the respective datasets. For instance, MS-COCO includes 80 categories, and we obtained one image for each category from online sources based on its category label. These images *do not need to be sourced from the model's training data*, as they are intended to represent general visual cognition. Fig. 3 showcases the 80 images we collected for the MS-COCO dataset, corresponding to its 80 categories, serving as the universal visual benchmarks.

C.6 Details of OD Backdoor Attacks

In this section, we systematically review the seven backdoor attacks in object detection that are included in our experimental evaluation.

Object generation attack (OGA). The goal of OGA [2] is to generate an FP bbox of the target class surrounding the trigger at a random position. The trigger is inserted into the

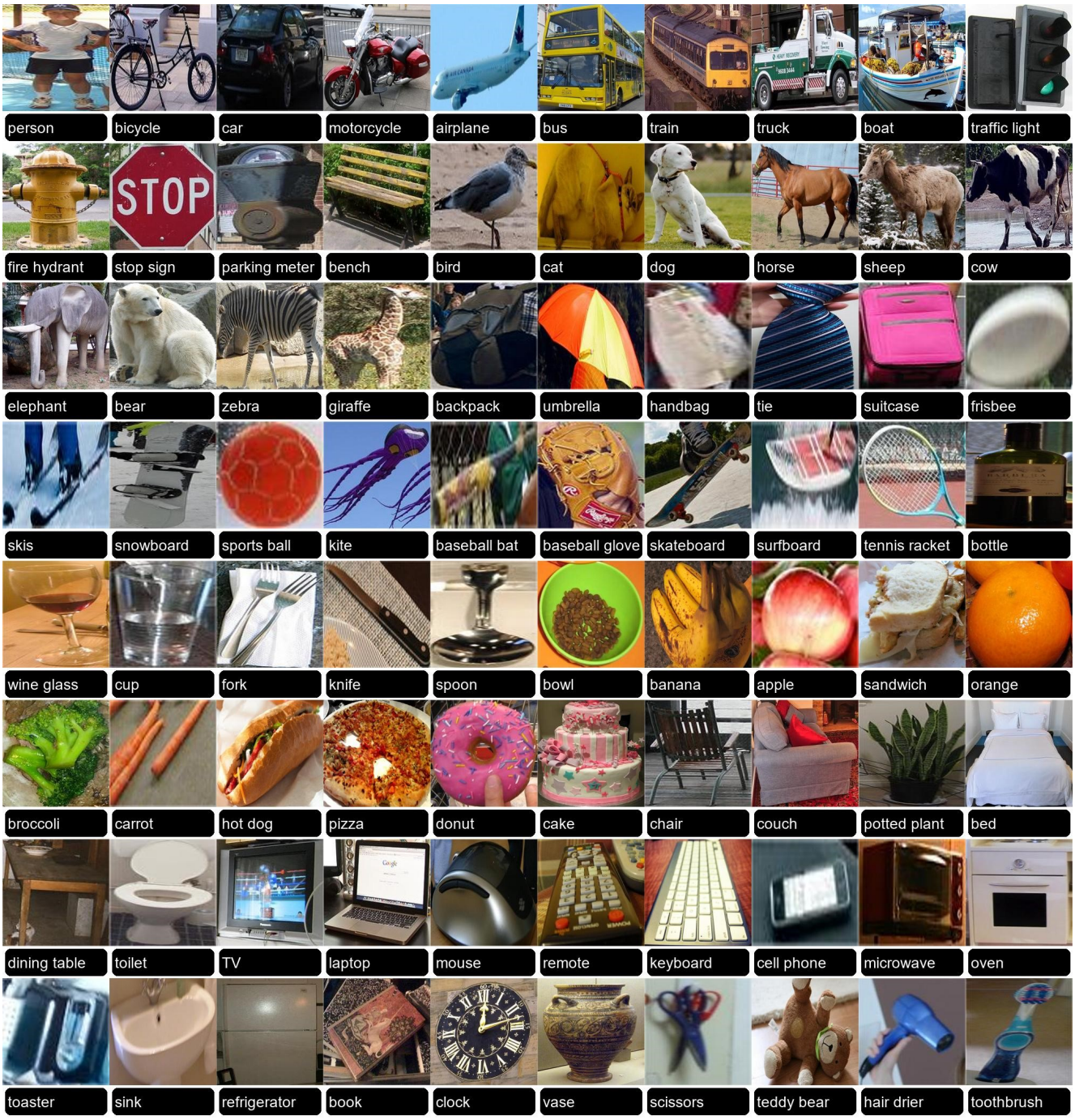

Figure 3. An overview of our universal visual benchmarks (*i.e.,* 80 samples) for the MS-COCO dataset (80 categories). **Note that these images do not need to originate from the model's training data, as they are designed to represent general visual cognition.** The primary purpose of employing universal visual benchmarks is to filter out natural backdoor objects (e.g., stop signs), a reasonable strategy given that these objects typically exhibit consistent patterns and features. Consequently, even when the universal visual benchmarks are entirely unrelated to the training dataset, they can achieve high SSIM scores on these natural backdoor objects, effectively eliminating their influence on the detection process. It is important to note that relying solely on universal visual benchmarks (*i.e.,* without TRACE) to compare objects with reference images in their respective categories is insufficient to distinguish objects from triggers. Only natural backdoor objects consistently display high SSIM, while most other objects, similar to triggers, yield lower SSIM scores during comparison, making it impossible to distinguish between them. This observation further highlights the necessity and validity of our TRACE.

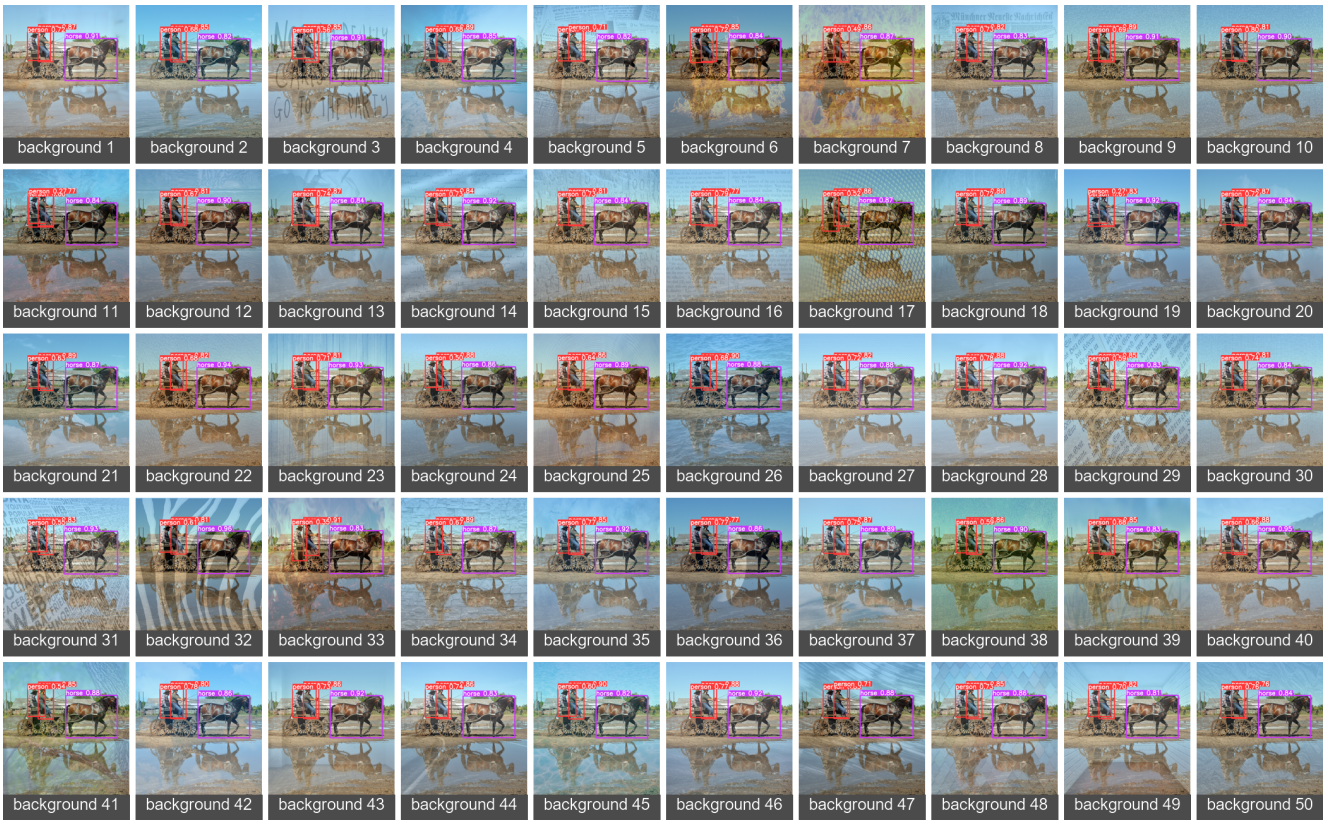

Figure 4. A clearer view: visualizing TRACE's contextual information transformation using the YOLO detector. The backgrounds used are from the background set in Fig. 1. The visualization of TRACE's focal information transformation can be found in Fig. 5.

random coordinate of an input image $\boldsymbol{x}$. $\mathbb{F}_\theta$ is expected to detect and classify the trigger in the poisoned image as the target class.

Regional misclassification attack (RMA). The goal of RMA [2] is to "regionally" change a surrounding object of the trigger to the target class. For a bbox not belonging to the target class, RMA inserts the trigger into the left-top corner of the bbox. $\mathbb{F}_\theta$ will detect and classify the objects with triggers as the target class. Misclassification attacks are usually label-specific, meaning the model only misclassifies objects from a specific victim class to the target class.

Global misclassification attack (GMA). The goal of GMA [2] is to "globally" change the predicted classes of all bboxes to the target class by inserting only one trigger into the left-top corner of the image. The trigger is inserted into the left-top corner $(0, 0)$ of the benign image. $\mathbb{F}_\theta$ is expected to detect and classify all the objects in the image as the target class.

Object disappearance attack (ODA). ODA [2] can make a surrounding bbox of the target class (*e.g.,* person) vanish. For an object o_i belonging to the target class, it inserts the trigger on the left-top corner of the object o_i. ODA will insert multiple triggers if there are many bboxes of the target class in the image. $\mathbb{F}_\theta$ should not detect the victim objects of the target class in the image.

Untargeted backdoor (UTA). UTA [19] aims to generate FNs around a trigger, concealing the detection of all victim objects. This is achieved during the training phase by affixing a trigger to an object and subsequently erasing its corresponding label.

Clean-image backdoor attack. CIB [3] manipulates only the training annotations while keeping the training images unaltered. It includes object disappearing, object appearing, and object misclassification attacks in multi-label models. Specifically, it operates by selecting a combination of benign category labels as a trigger pattern. The adversary then poisons the training set by falsifying the annotations of images containing these categories. CIB encompasses a variety of attack objectives. In our evaluation, we focus on the representative and more challenging object disappearance branch.

Detector collapse. DC [30] introduces a novel backdoor attack paradigm targeting object detection models. DC focuses on global performance degradation through two strategies: SPONGE, which causes a flood of false positives, and BLINDING, which renders objects undetectable

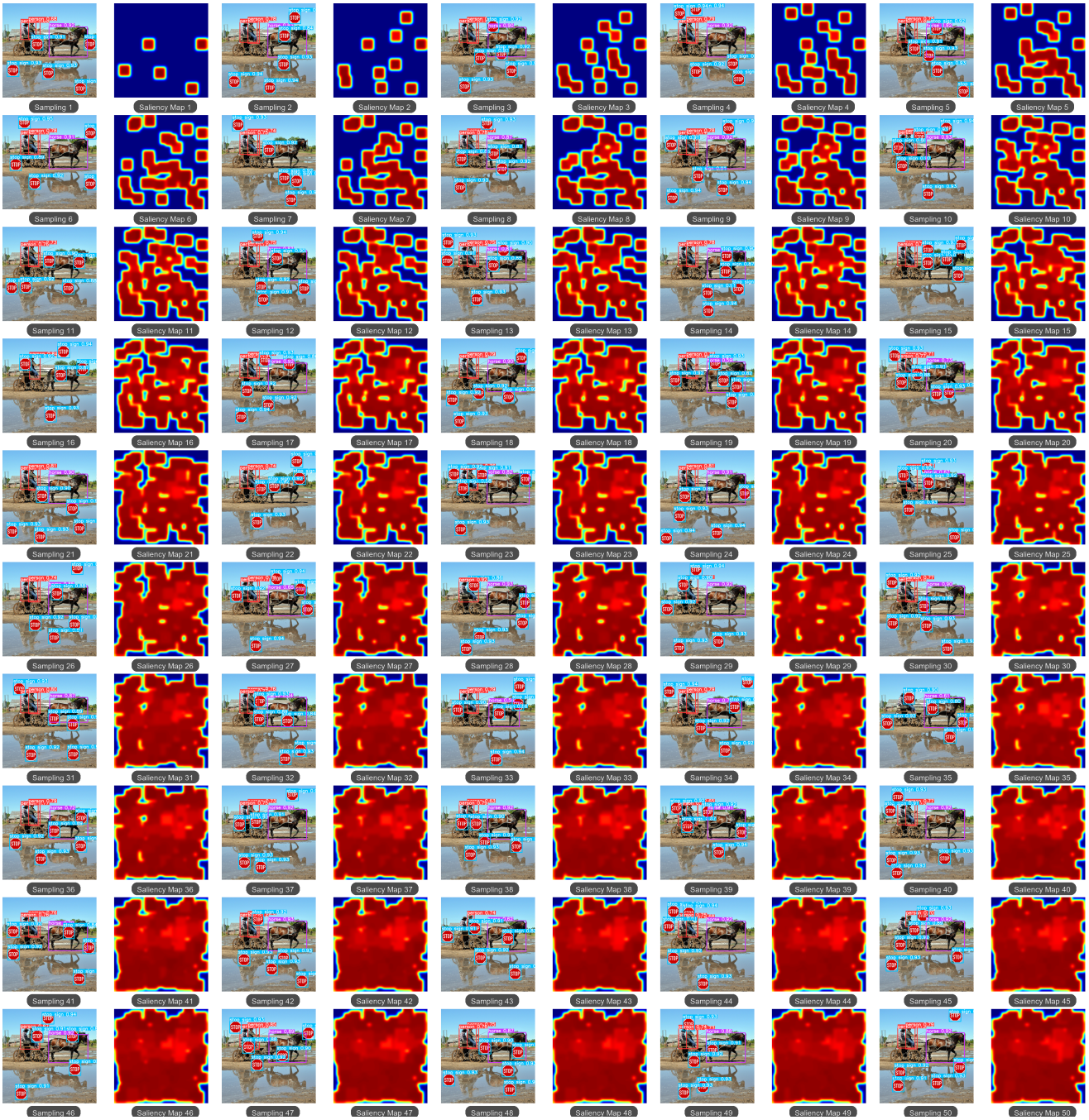

Figure 5. A clearer view: visualizing TRACE's focal information transformation using the YOLO detector. This process traces heatmaps of the model's attention areas from iterations 1 to 50. By iteration 30, nearly all positions in the image are covered, and the relatively complete red regions demonstrate that the stop sign exhibits stable detection. In other words, for clean images, the Focal Transformation Consistency (FTC) value at this stage is relatively small. **Accordingly, we also present in Fig. 6 the different behaviors of these heatmaps when an FN-inducing trigger appears at the center of the same image (*i.e.,* a higher FTC value).**

by the detector. We focus on the representative BLINDING variant of DC, which specializes in causing object disappearance, posing critical challenges to safety-critical applications.

Parameter setting for baseline attacks. In the training phase, we follow the methodologies established in previous works, adopting the same training and poisoning settings (e.g., the choice of trigger patterns and trigger sizes) as out-

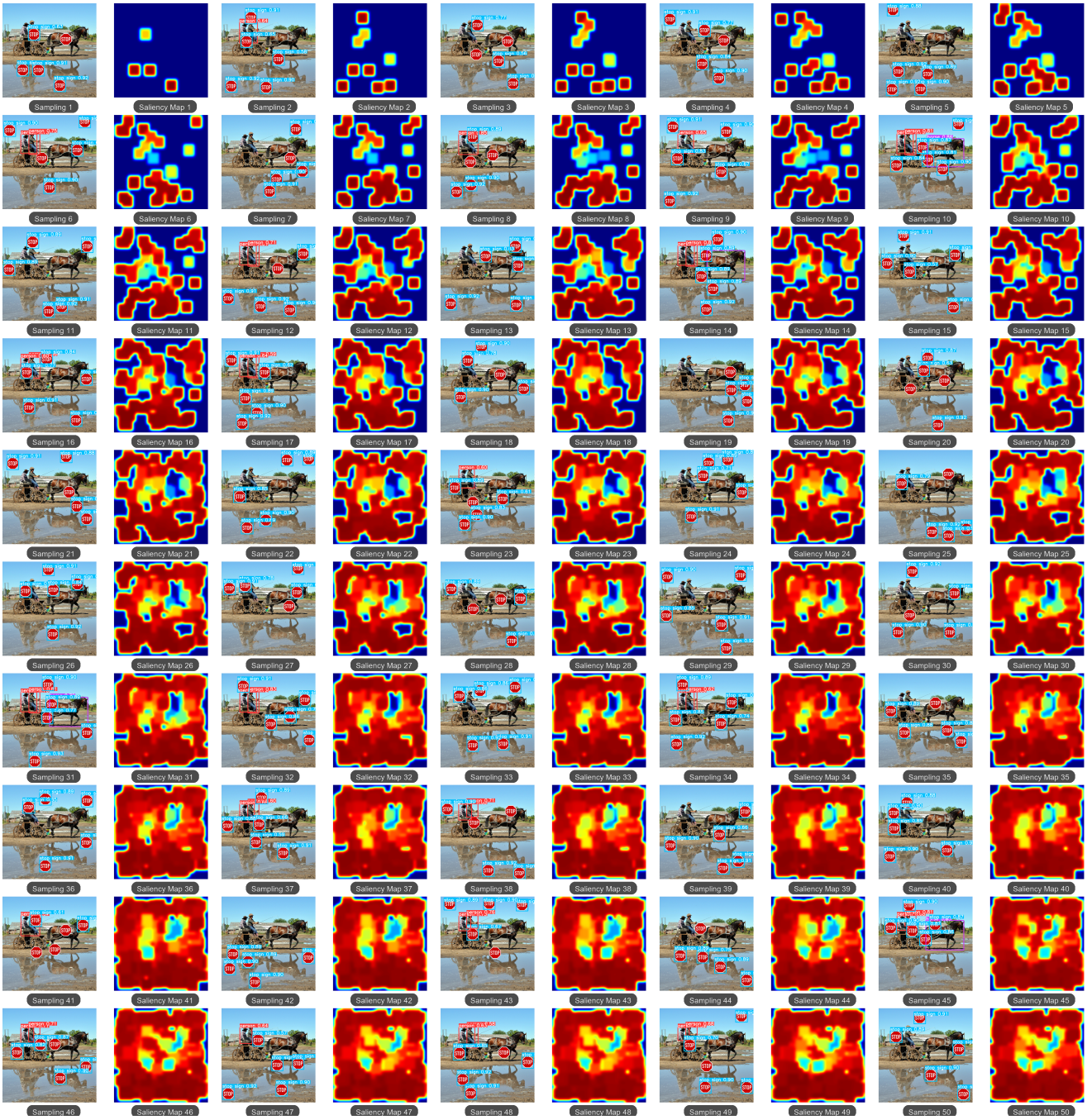

Figure 6. A clearer view: visualizing TRACE's contextual information transformation using the YOLO detector. The backgrounds used are from the background set in Figure 1.

lined in their respective papers. It should be noted that we set a reasonable poisoning rate and ensured that the attack success rate is above 90%. This is important for backdoor detection, as we believe that an unsuccessful attack is not only meaningless, but also has a negative impact on detection.

STRIP. Originally proposed for image classification, STRIP detects poisoned inputs by measuring the entropy of prediction distributions under input perturbations. For object detection, however, STRIP requires significant adaptations due to the complexity of the detection pipeline and the presence of multiple bounding boxes in a single image. To adapt STRIP to object detection, we calculate the average entropy of predictions over all detected bounding boxes for

a perturbed image. Specifically, perturbations are applied to the entire input image by blending it with random noise patterns, with 100 perturbed versions generated per input image. The prediction entropies of all bounding boxes are then aggregated to compute the average entropy. For example, in the Faster R-CNN + VOC2007 setting, a threshold is established based on the entropy distribution of clean samples. Inputs with average entropy falling outside the interval $[0.3, 0.7]$ are flagged as potentially poisoned. These thresholds and perturbation parameters are empirically determined to balance detection accuracy and computational efficiency.

Detector Cleanse. Detector Cleanse operates under the assumption of complete knowledge about the backdoor attacker. We note that the official implementation of Detector Cleanse is not publicly available. We replicated it according to the paper. Detector Cleanse is specifically designed for object detection, operating at runtime to identify poisoned inputs. The method assumes the availability of a small set of clean features, which are extracted from ground-truth bounding boxes in clean datasets such as PASCAL VOC or MS-COCO. The experimental setup involves drawing $N = 100$ features $\chi = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ from these clean bounding box regions. For each input image $\boldsymbol{x}$, Detector Cleanse perturbs each predicted bounding box $\boldsymbol{b}$ by blending its features with those in χ, generating N perturbed versions of $\boldsymbol{b}$. The entropy of the predicted probabilities for these perturbed bounding boxes is then computed, and the average entropy is used as the detection criterion. Specifically, if the average entropy lies outside the range $[m - \Delta, m + \Delta]$, where m is set to 0.55 (mean of clean entropy distribution) and Δ is set to 0.3 (double the standard deviation of the distribution), the image is flagged as poisoned. Their thresholds are determined empirically using 500 clean images from PASCAL VOC2007.

TeCo. TeCo detects backdoor-triggered samples by evaluating the consistency of model predictions under various input corruptions. Backdoor-infected models exhibit inconsistent prediction variance for trigger samples, enabling their identification without requiring access to clean data or prior knowledge of triggers. TeCo applies 15 types of common input corruptions, including Gaussian noise, blur, contrast, and pixelation, at 5 severity levels for each corruption type, as specified in the original work. For each input sample, these 75 corrupted versions are fed into the model, and the variance in the predicted probabilities across these corrupted inputs is computed. A detection threshold is defined based on the variance distribution of benign samples to distinguish clean inputs from poisoned ones. We closely adhered to the parameter configurations specified in the original paper in our evaluation.

FreqDetector. FreqDetector identifies backdoor triggers from a frequency perspective. By analyzing images in the frequency domain using the Discrete Fourier Transform (DFT), it detects anomalies introduced by backdoor triggers, which often manifest as high-frequency components not present in benign samples. We strictly followed the parameter settings outlined in the original paper.

SCALE-UP. SCALE-UP is a black-box input-level backdoor detection method that analyzes the consistency of model predictions when input pixel values are amplified. It observes that poisoned samples maintain consistent predictions under such amplification, whereas benign samples do not. By measuring this scaled prediction consistency, SCALE-UP effectively identifies backdoor-triggered inputs without requiring access to the model's internal parameters or clean data. We strictly followed the parameter settings outlined in the original paper.

C.7 Details of Evaluation Metrics

To comprehensively evaluate the performance of the detection methods, we adopt several standard metrics, including Precision, Recall, F1 Score, and AUROC. Below, we introduce these metrics.

Precision. Precision measures the proportion of correctly identified positive samples (True Positives, TP) out of all predicted positive samples (TP and False Positives, FP). It is defined as:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{1}$$

where a higher Precision indicates fewer false alarms among detected triggers.

Recall. Recall measures the proportion of correctly identified positive samples (TP) out of all actual positive samples (TP and False Negatives, FN). It is defined as:

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{2}$$

indicating the model's ability to detect all trigger samples.

F1 Score. F1 Score is the harmonic mean of Precision and Recall, providing a balanced metric that considers both false positives and false negatives. It is computed as:

$$F1\ Score = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{3}$$

The F1 Score is particularly useful when there is an imbalance between the number of positive and negative samples.

AUROC (Area Under Receiver Operating Characteristic Curve). AUROC evaluates the model's ability to distinguish between positive and negative samples across different threshold settings. It is defined as the area under the ROC curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold levels:

$$AUROC = \int_0^1 \mathrm{TPR}(\mathrm{FPR})\, d(\mathrm{FPR}), \tag{4}$$

Figure 7. Confidence distribution differences across various objects (using DETR as the detector). *Note that the violin plot for triggers shows distributions concentrated near a confidence value of 1, making them visually indistinguishable.*

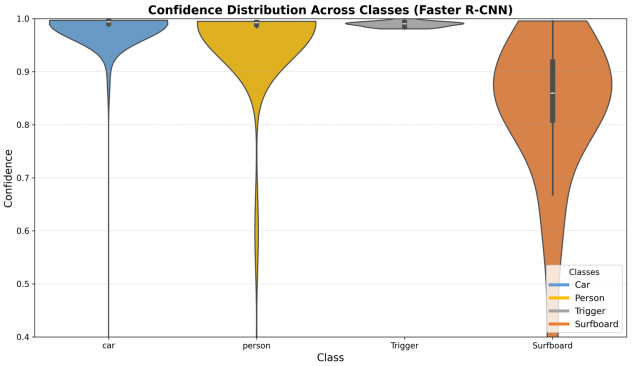

Figure 8. Confidence distribution differences across various objects (using Faster-RCNN as the detector).

where a value closer to 1 indicates better overall performance.

D. Additional Analyses

D.1 Contextual Information Transformation Across Different Object Detectors

In addition to the violin plot of the single-stage detector YOLOv5 provided in the manuscript, we also present the results of two other detectors: the two-stage detector Faster R-CNN and the vision transformer-based detector DETR. As shown in Fig. 7 and Fig. 8, the results indicate that categories such as "Car", "Surfboard", and "Person" exhibit significant variance across different detectors under TRACE's contextual information transformation (*i.e.,* background fusion). In contrast, trigger objects consistently maintain very low variance across all detectors. This demonstrates the universality and generalizability of the contextual transformation consistency evaluation across various models.

D.2 Further Analysis of Experimental Results

We also observe certain limitations in TRACE's performance, particularly in handling CIB attacks. CIB leverages the co-occurrence of two natural features as a trigger pattern, making it more subtle and harder to detect compared to fixed-pattern triggers. We observed a significant performance drop in TRACE when detecting CIB. Note that these natural features simultaneously serve as benign characteristics and backdoor triggers, complicating detection not only for TRACE but also for other methods. For instance, SCALE-UP performs exceptionally poorly against CIB. A possible explanation is that SCALE-UP relies on pixel-scaling consistency, which is disrupted in clean-image backdoors. In such cases, the consistency relationship between clean and backdoored samples is reversed, leading to greater inconsistency in backdoored samples and further hindering detection.

E. Additional Analyses

E.1 Unified representation of existing attacks

Object Generation Attack (OGA). The goal of OGA [2] is to generate a false-positive (FP) bounding box of the target class surrounding the trigger at a random position. Formally, the trigger is stamped at a random coordinate $\boldsymbol{o}_t$ in the input image $\boldsymbol{x}$. The attack ensures $(M(\boldsymbol{o}_t, \hat{\boldsymbol{o}}_l), \boldsymbol{y}_t) \in \mathbb{F}_\theta(\boldsymbol{x} \oplus \boldsymbol{t})$, where $M(\boldsymbol{o}_t, \hat{\boldsymbol{o}}_l) = \boldsymbol{o}_t$.

Regional Misclassification Attack (RMA). The goal of RMA [2] is to regionally change the class of objects near the trigger to the target class. Specifically, the trigger is inserted at the top-left corner of a bounding box $\hat{\boldsymbol{o}}_l$ with $\hat{\boldsymbol{y}}_l \neq \boldsymbol{y}_t$. The attack ensures $(\hat{\boldsymbol{o}}_l, \boldsymbol{y}_t) \in \mathbb{F}_\theta(\boldsymbol{x} \oplus \boldsymbol{t})$. Multiple triggers are inserted into the image to achieve the regional misclassification effect.

Global Misclassification Attack (GMA). The goal of GMA [2] is to globally change the predicted class of all bounding boxes in the image to the target class by inserting a single trigger at the top-left corner of the image. Formally, the attack ensures $(\boldsymbol{o}_i, \boldsymbol{y}_t) \in \mathbb{F}_\theta(\boldsymbol{x} \oplus \boldsymbol{t}), \forall i$, where all objects $\boldsymbol{o}_i$ are classified as the target class $\boldsymbol{y}_t$.

Object Disappearance Attack (ODA). The goal of ODA [2] is to make objects of the target class vanish from the detection results. For an object $\hat{\boldsymbol{o}}_l$ belonging to the class $\boldsymbol{y}_l$, the trigger is placed on the top-left corner of $\hat{\boldsymbol{o}}_l$, ensuring $(\hat{\boldsymbol{o}}_l, \hat{\boldsymbol{y}}_l) \notin \mathbb{F}_\theta(\boldsymbol{x} \oplus \boldsymbol{t})$. If multiple target objects exist in the image, triggers are applied to all of them.

Untargeted Backdoor Attack (UTA). The goal of UTA [19] is to generate false negatives (FNs) around a trigger, concealing the detection of victim objects. During training, this is achieved by attaching the trigger to the victim object while erasing its corresponding label. At inference, the model fails to detect the victim object near the trigger, ensuring $(\hat{\boldsymbol{o}}_l, \hat{\boldsymbol{y}}_l) \notin \mathbb{F}_\theta(\boldsymbol{x} \oplus \boldsymbol{t})$ for all victim objects.

Clean-Image Backdoor Attack (CIB). CIB [3] manipulates only the training annotations while leaving the training images unaltered. It includes object disappearance, object appearance, and object misclassification attacks. For object disappearance, the poisoned annotations ensure $(\hat{\boldsymbol{o}}_l, \hat{\boldsymbol{y}}_l) \notin \mathbb{F}_\theta(\boldsymbol{x})$. For object appearance, the poisoned annotations result in $(M(\boldsymbol{o}_t, \hat{\boldsymbol{o}}_l), \boldsymbol{y}_t) \in \mathbb{F}_\theta(\boldsymbol{x})$. For misclassification, the victim object is annotated as $(\hat{\boldsymbol{o}}_l, \boldsymbol{y}_t) \in \mathbb{F}_\theta(\boldsymbol{x})$.

Detector Collapse (DC). DC [30] is a novel backdoor attack targeting object detection models. Its BLINDING variant causes all objects in the image to disappear. Formally, for every ground-truth bounding box $(\hat{\boldsymbol{o}}_i, \hat{\boldsymbol{y}}_i)$, the attack ensures $(\hat{\boldsymbol{o}}_i, \hat{\boldsymbol{y}}_i) \notin \mathbb{F}_\theta(\boldsymbol{x} \oplus \boldsymbol{t}), \forall i$, where all objects in the image are considered victim objects.

E.2 Insights into Contextual Bias

Context is a common element in the visual world. For individual instances within an image, their context consists of other co-occurring instances and the background. Indeed, recent efforts have moved from recurrent neural networks [25, 27] to graph convolutional networks [4, 7] and transformer-based frameworks [14, 31], aiming to model contextual relationships in multi-label images to enhance performance. The image backgrounds become a natural source of correlation between the objects and their annotations. The detector's learned representations tend to exploit spurious scene correlations, *e.g.,* zebras are more likely to be found on safaris than on streets or indoors. From the data perspective, specific dataset characteristics amplify such biases: the COCO dataset exemplifies frequent object-background co-occurrences (*e.g.,* approximately 90% of ball instances appearing in sports fields), reinforcing the detector's over-reliance on contextual cues rather than object-specific features. Interestingly, we observe that the trigger maintains consistently high confidence scores across all background variations. In contrast, clean objects display significant fluctuations in their confidence when subjected to the same background transformations. We attribute this phenomenon primarily to the widely accepted concept of "shortcut learning" [10], where the backdoor training process establishes a robust mapping between a specific, uniform pattern and the target label. This makes the recognition of trigger objects less susceptible to contextual biases. *Also from the data perspective, because triggers are uniformly distributed across diverse scenes during data poisoning, they avoid spurious correlations with particular backgrounds, further explaining their consistent confidences.*

E.3 A Detailed Explanation of Failure of Adaptive Attacks

Our designed adaptive attack incorporates a dual-objective loss function $\mathcal{J} = \mathcal{J}_{bd} + \lambda \mathcal{J}_{adap}$ to align the backdoor model's behavior with the consistency measures utilized by TRACE. Despite this comprehensive attack strategy, we observe that these adaptive attacks significantly compromise the attack performance. *Below, we discuss the underlying reasons for the failure of such adaptive attacks.*

Conflict Between Loss Objectives. The key challenge lies in reconciling the dual-objective loss $\mathcal{J}$. While the backdoor loss $\mathcal{J}_{bd}$ aims to maximize attack success (*e.g.,* enforcing trigger-specific behavior), the adaptive loss $\mathcal{J}_{adap}$ forces the model to mimic normal detection behavior (like benign objects) under TRACE's transformations. This creates an inherent conflict: optimizing one term inevitably undermines the other. For instance, minimizing $\mathcal{J}_{adap}$ to align transformation consistency for poisoned and clean samples dilutes the distinct backdoor patterns essential for $\mathcal{J}_{bd}$. As a result, the model's transformation consistency becomes a bottleneck for achieving high attack success rates.

Constraints on λ. The weighting factor λ in the adaptive loss determines the trade-off between attack success and transformation consistency. Empirically, small λ values lead to negligible changes in consistency, failing to bypass TRACE 's detection mechanism. Conversely, larger λ values drastically reduce the attack success rate. This constraint underscores the difficulty in achieving a balanced attack under realistic settings.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV'20)*, pages 213–229, 2020. 2, 3

[2] Shih-Han Chan, Yinpeng Dong, Jun Zhu, Xiaolu Zhang, and Jun Zhou. Baddet: Backdoor attacks on object detection. In *Proceedings of the European Conference on Computer Vision (ECCV'22)*, pages 396–412. Springer, 2022. 1, 3, 5, 9

[3] Kangjie Chen, Xiaoxuan Lou, Guowen Xu, Jiwei Li, and Tianwei Zhang. Clean-image backdoor: Attacking multi-label models with poisoned labels only. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR'22)*, 2022. 1, 5, 10

[4] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR'19)*, pages 5177–5186, 2019. 10

[5] Siyuan Cheng, Guangyu Shen, Guanhong Tao, Kaiyuan Zhang, Zhuo Zhang, Shengwei An, Xiangzhe Xu, Yingqi Liu, Shiqing Ma, and Xiangyu Zhang. Odscan: Backdoor scanning for object detection models. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'24)*, pages 119–119, 2024. 1

[6] Yinpeng Dong, Xiao Yang, Zhijie Deng, Tianyu Pang, Zihao Xiao, Hang Su, and Jun Zhu. Black-box detection of back-

door attacks with limited information and data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'21)*, pages 16482–16491, 2021. 1

[7] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR'19)*, pages 647–657, 2019. 10

[8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, pages 303–338, 2010. 2

[9] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th annual computer security applications conference (ACSAC'19)*, pages 113–125, 2019. 1

[10] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 10

[11] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. 1

[12] Junfeng Guo, Ang Li, and Cong Liu. Aeva: Black-box backdoor detection using adversarial extreme value analysis. In *Proceedings of the International Conference on Learning Representations (ICLR'21)*, 2021. 1

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3

[14] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*, pages 16478–16488, 2021. 10

[15] Junyu Lin, Lei Xu, Yingqi Liu, and Xiangyu Zhang. Composite backdoor attack for deep neural network by mixing existing benign features. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 113–131, 2020. 1

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[17] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37, 2016. 2

[18] Xiaogeng Liu, Minghui Li, Haoyu Wang, Shengshan Hu, Dengpan Ye, Hai Jin, Libing Wu, and Chaowei Xiao. Detecting backdoors during the inference stage based on corruption robustness consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'23)*, pages 16363–16372, 2023. 1

[19] Chengxiao Luo, Yiming Li, Yong Jiang, and Shu-Tao Xia. Untargeted backdoor attack against object detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'23)*, pages 1–5. IEEE, 2023. 1, 5, 9

[20] National Institute of Standards and Technology (NIST). Trojai leaderboard. `https://pages.nist.gov/trojai/`, 2024. Accessed: 2024-11-21. 2

[21] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR'16)*, pages 779–788, 2016. 2, 3

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS'15)*, 28, 2015. 2, 3

[23] Guangyu Shen, Siyuan Cheng, Guanhong Tao, Kaiyuan Zhang, Yingqi Liu, Shengwei An, Shiqing Ma, and Xiangyu Zhang. Django: Detecting trojans in object detection models via gaussian focus calibration. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS'24)*, 36, 2024. 1

[24] Ilia Shumailov, Zakhar Shumaylov, Dmitry Kazhdan, Yiren Zhao, Nicolas Papernot, Murat A Erdogdu, and Ross J Anderson. Manipulating sgd with data ordering attacks. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS'21)*, 34:18021–18032, 2021. 1

[25] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR'16)*, pages 2285–2294, 2016. 10

[26] Xiaojun Xu, Qi Wang, Huichen Li, Nikita Borisov, Carl A Gunter, and Bo Li. Detecting ai trojans using meta neural analysis. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'21)*, pages 103–120. IEEE, 2021. 1

[27] Vacit Oguz Yazici, Abel Gonzalez-Garcia, Arnau Ramisa, Bartlomiej Twardowski, and Joost van de Weijer. Orderless recurrent models for multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20)*, pages 13440–13449, 2020. 10

[28] Yi Zeng, Won Park, Z Morley Mao, and Ruoxi Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV'21)*, pages 16473–16481, 2021. 1

[29] Haichao Zhang and Jianyu Wang. Towards adversarially robust object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 421–430, 2019. 2

[30] Hangtao Zhang, Shengshan Hu, Yichen Wang, Leo Yu Zhang, Ziqi Zhou, Xianlong Wang, Yanjun Zhang, and Chao Chen. Detector collapse: Backdooring object detection to

catastrophic overload or blindness. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI'24)*, 2024. 1, 5, 10

[31] Jiawei Zhao, Ke Yan, Yifan Zhao, Xiaowei Guo, Feiyue Huang, and Jia Li. Transformer-based dual relation graph for multi-label image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV'21)*, pages 163–172, 2021. 10