TRACE: Confounder-free Adversarial Fine-tuning for Robust Object Detection Supplementary Material

Wonho Lee Jisu Lee Hyunsik Na Sohee Park Daeseon Choi
Soongsil University
{hoho0907, connandgo, rnrud7932, sosohi}@soongsil.ac.kr, sunchoi@ssu.ac.kr

1. Validation of IV in TRACE

In TRACE, to control and generalize confounders arising from the attachment of adversarial patches, the Instrumental Variable (IV) Z is defined as the backbone feature variation, $Z = F_{adv} - F_{clean}$. For successful IV regression using this IV, three validity conditions—*Unconfoundedness*, *Exclusion Restriction*, and *Relevance*—must be satisfied, as established in previous studies [2, 4]. Accordingly, this section investigates whether the defined IV satisfies these three requirements in the adversarial fine-tuning environment of TRACE.

1.1. Unconfoundedness

Unconfoundedness requires that the IV Z be independent of the outcome error ψ; in other words, $\psi \perp Z$ must be satisfied. The outcome error is defined as the residual component of Y that is not predicted given treatment T. If Z is correlated with ψ, then rather than serving as a control for confounders, the IV could introduce additional confounding and distort causal inference. In TRACE, Z is defined as the backbone feature variation induced by adversarial patches. Z arises from physical properties of the patch, such as location, brightness, and angle. These variations are exogenously determined and structurally separated from the factors that determine the outcome error. While the outcome error ψ primarily arises in the localization and classification processes within the neck and head stages, Z is defined at the backbone feature level. Thus there is no correlation between ψ and Z. Consequently, in TRACE, Z is independent of ψ and the *Unconfoundedness* condition is satisfied.

1.2. Exclusion Restriction

Exclusion Restriction requires that the IV Z does not directly affect the outcome Y, but instead influences it only through the treatment variable T. Formally, $Z \perp Y \mid T, \psi$ must be satisfied. If Z were to affect directly on Y, then the IV would fail to control for unobserved confounders and would transmit spurious correlation. In TRACE, Z is not structurally passed directly from the backbone to the neck or head. Instead, Z must be merged with the clean feature F_{clean} to form either the counterfactual feature $CF = F_{clean} + g(Z)$ or the counterfactual causal feature $CC = F_{clean} + h(g(Z))$, through which it indirectly influences the outcome. As a result, the only structural pathway by which Z can affect Y is $Z \rightarrow T \rightarrow Y$. Therefore, the architecture of TRACE guarantees that Z cannot influence Y without passing through T, thereby satisfying the *Exclusion Restriction* condition.

1.3. Relevance

Relevance requires that the IV Z be statistically correlated with the treatment variable T; that is, $Cov(Z, T) \neq 0$ must be satisfied. If this condition is not satisfied, Z would be subject to instrument weakness. In TRACE, the treatment is defined as $T = F_{clean} + Z$, which structurally guarantees that Z directly contributes to the variation in T. Therefore, Z is strongly correlated with T and satisfies the *Relevance* condition.

2. Clean Accuracy

We measured the clean accuracy of the Base Model, which underwent no tuning, and two adversarial fine-tuned models, ImageAT and TRACE, using the MS COCO2017 validation set. Table 1 presents the clean accuracy results for the three models. Both ImageAT and TRACE showed lower performance than the Base Model across both architectures, Yolov5 and Yolov8. This decline reflects the trade-off introduced by adversarial fine-tuning, a well-known limitation for which various improvements have been studied, suggesting the possibility of further enhancement. A more detailed discussion of this issue is provided later in Sec. 6. When comparing ImageAT and TRACE, the results reveal slight differences between the two models. For Yolov5, the mAP50 values were identical, while TRACE outperformed ImageAT by a marginal 0.001 in mAP50-95. However, in terms of Precision and Recall, TRACE scored slightly lower than ImageAT. In contrast, for Yolov8 the performance gap between the two models was much more pronounced.

		Precision	Recall	mAP50	mAP50-95
Yolov5	Base Model	0.745	0.595	0.658	0.469
	ImageAT	0.663	0.506	0.538	0.364
	TRACE	0.659	0.496	0.538	0.365
Yolov8	Base Model	0.728	0.624	0.681	0.517
	ImageAT	0.634	0.506	0.540	0.389
	TRACE	0.651	0.544	0.580	0.417

Table 1. Clean accuracy comparison of ImageAT and TRACE.

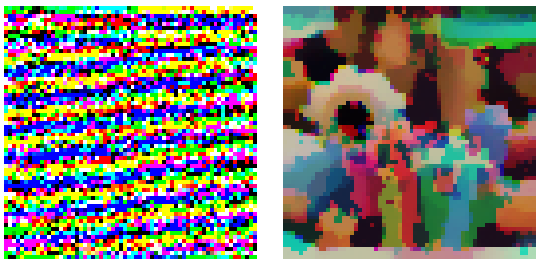
TRACE surpassed ImageAT by 0.04 in mAP50, and this advantage extended to mAP50-95, indicating more consistent performance across a wider range of IoU thresholds. We speculate that these performance differences between Yolov5 and Yolov8 stem from differences in backbone feature dependence. Yolov5 employ Cross-Stage Partial (CSP) Net [6], Yolov8 employ CSPDarknet [1], which intergrates CSP into Darknet architecture, so utilizes a more advanced backbone with improved feature extraction capabilities. As a result, Yolov8 not only achieves higher baseline performance but also exhibits greater dependence on backbone features. Unlike ImageAT, which attaches adversarial patches at the pixel level, TRACE injects causal features into the backbone during adversarial fine-tuning. The fact that TRACE consistently outperformed ImageAT across all metrics on Yolov8 suggests that the injected causal features contribute directly to correct predictions. Considering the increasing feature extraction capabilities of modern models, these results imply promising applicability of TRACE to a broader range of recent architectures.

3. Unseen Adversarial Patch

Fig. 1 shows the adversarial patches used during adversarial fine-tuning (Trained) and those used in the Unseen Attack experiment (NPSTV). The Trained patches, generated with the basic attack objective loss, primarily exhibit pixel-level noise patterns, whereas the NPSTV patches, produced with the additional incorporation of NPS Loss and TV Loss, take on more natural and visually coherent shapes. These two types of patches follow entirely different distributions and display distinct visual characteristics. Consequently, in the Unseen Attack results, ImageAT remains vulnerable to NPSTV patches because it has primarily learned the visual properties of the Trained patches. In contrast, TRACE demonstrates relatively stronger robustness by leveraging generalized causal feature–based representations, which enable it to resist patch distributions beyond those encountered during training.

4. Qualitative Evaluation

In the main paper, we conducted a qualitative evaluation to examine whether TRACE achieves generalization with concerning three major confounders—*location*, *rotation*, and *brightness*—and thereby contributes to securing generalized robustness. This Supplementary Materials provides a more detailed qualitative analysis and presents additional visualization cases for an in-depth comparison between the two models. For this purpose, 500 single-object images were extracted from the COCO validation set, and experiments were performed by applying both Trained patches used during adversarial fine-tuning and newly generated Adaptive patches. Under each condition, confidence scores for the objects were measured, and various visualization techniques, such as heatmaps for location changes, semi-circle plots for rotation, and line graphs for brightness variation, were employed to systematically compare the robustness of TRACE and ImageAT. All visualization results were arranged in a 2×2 grid to enable direct comparison between TRACE and ImageAT for a single original image. The top row presents TRACE (Trained, Adaptive), and the bottom row presents ImageAT (Trained, Adaptive), with colors indicating confidence scores.



(a) Trained patch. (b) NPSTV patch.

Figure 1. Generated patches for unseen attack.

4.1. Generalization on Location

To analyze the effect of adversarial patch location, we moved the patch within the bounding box at 16-pixel intervals and measured the corresponding confidence scores. The bounding box was divided into a 16×16 grid, and the weighted average score was computed by reflecting the overlap ratio of the patch within each grid cell. The results were converted into heatmaps and overlaid on the original images, with areas outside the bounding box set to zero and displayed in blue. The visualization results for location variation are shown in Figs. 2 and 3. Each heatmap represents the confidence score distribution with respect to patch location inside the bounding box, where blue denotes low scores and red denotes high scores. Overall, TRACE maintained a stable distribution of confidence scores across different patch locations and achieved higher average values. In contrast, ImageAT exhibited localized vulnerabilities where performance dropped sharply. For instance, in the bottom left of Fig. 2, ImageAT showed severe degradation when patches were placed near the upper center and corners of the object, whereas TRACE preserved stable dis-

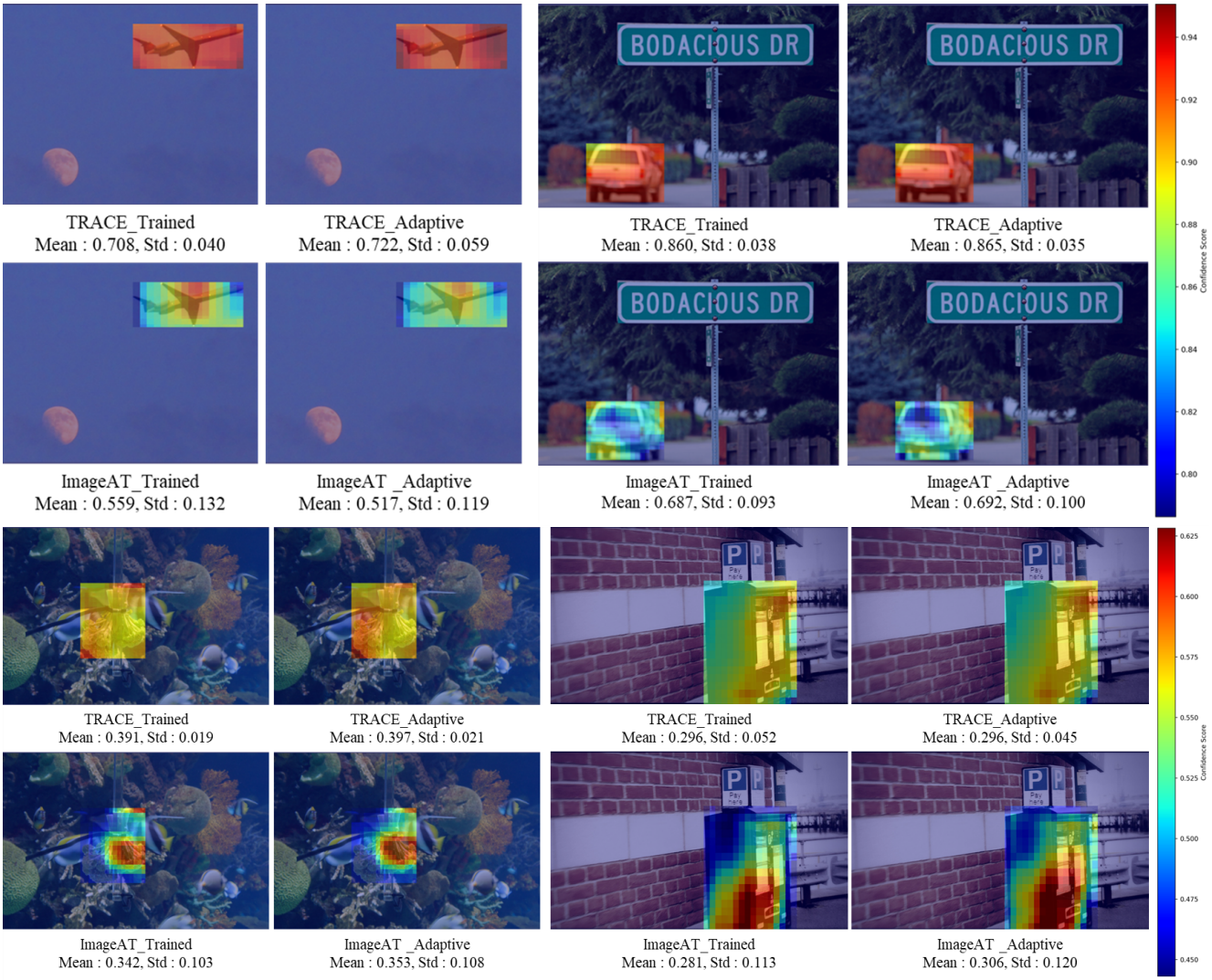


Figure 2. Visualization results of patch location generalization for TRACE and ImageAT.

tributions and high scores even at the same positions. Similarly, in the bottom left and bottom right of Fig. 2, ImageAT displayed large fluctuations between central and surrounding regions, while in the top right of Fig. 3, it revealed concentrated vulnerabilities in the lower bounding box area. By comparison, TRACE demonstrated a more uniform distribution of confidence scores across all such cases.

In conclusion, while ImageAT heavily relied on specific regions and exhibited localized weaknesses under patch location changes, TRACE successfully generalized over the location confounder, ensuring more consistent robustness and stable performance.

4.2. Generalization on Rotation

To analyze the effect of patch rotation, we rotated the patch around the center of the bounding box from –90° to 90° in 1° increments and measured the corresponding confidence scores. The distribution of scores across angles was then color-mapped and visualized using semicircle plots. The visualization results are shown in Fig. 4. Each semicircle plot represents the mean confidence score for each rotation angle within the –90° to 90° range, with colors indicating confidence score. As observed in the top-left and second-row-left plots of Fig. 4, ImageAT exhibited clear vulnerabilities outside the –30° to 30° range, with lower mean scores and substantially larger variance. This suggests that, because adversarial fine-tuning was limited to this angular range, the model failed to defend against patches rotated beyond it. Although confidence scores increased again outside the –60° to 60° range, this reflects a loss of attack effectiveness rather than true robustness of the model. In contrast, TRACE maintained stable distributions and com-

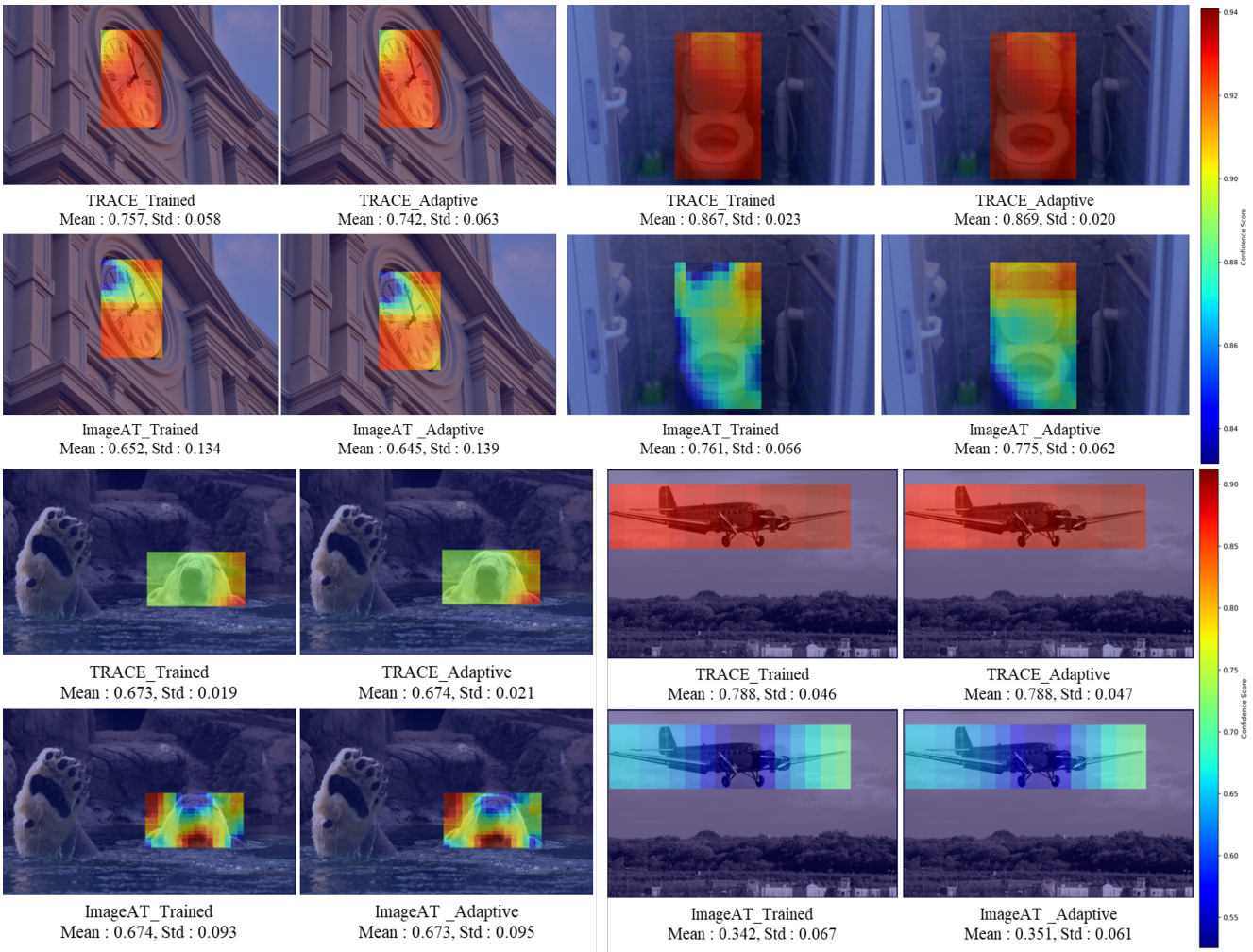


Figure 3. Visualization results of patch location generalization for TRACE and ImageAT.

parable performance across all angles, thereby demonstrating reliable generalization even under unseen conditions. A contrasting case is shown in the second-row-right plot of Fig. 4, where ImageAT maintained high scores within the $|30° \text{ to } 60°|$range but remained vulnerable elsewhere. This indicates that its robustness was inconsistently expressed, limited to certain angles where the patch attack was weaker. TRACE, by comparison, exhibited smaller variations and stable confidence scores across all angles. The fourth-row-right plot of Fig. 4 highlights these differences most clearly. TRACE preserved uniformly high confidence scores with low variance across the full rotation range, whereas ImageAT showed overall lower scores and irregular fluctuations, reflecting unstable performance. This implies that ImageAT may appear robust under specific rotation conditions, but fails to provide consistent robustness overall.

In conclusion, while ImageAT exhibited distinct vulnerabilities and unstable distributions at unseen rotation angles, TRACE maintained stable and consistent performance across all angles, confirming its superior generalization to rotational variations.

4.3. Generalization on Brightness

To analyze the effect of brightness variation, we adjusted patch brightness from –30% to 30% in 1% increments. For each brightness condition, confidence scores were measured, and their means and standard deviations were calculated. Individual images were visualized with line graphs showing score changes across brightness levels, while aggregated scores under each condition were used to derive an average curve for overall analysis. The visualization results are presented in Fig. 5. Each graph depicts the confidence scores as a linear plot over the –30% to 30% brightness range. Overall, TRACE maintained stable distributions and consistently high mean scores under brightness variation. In contrast, ImageAT occasionally showed increased scores

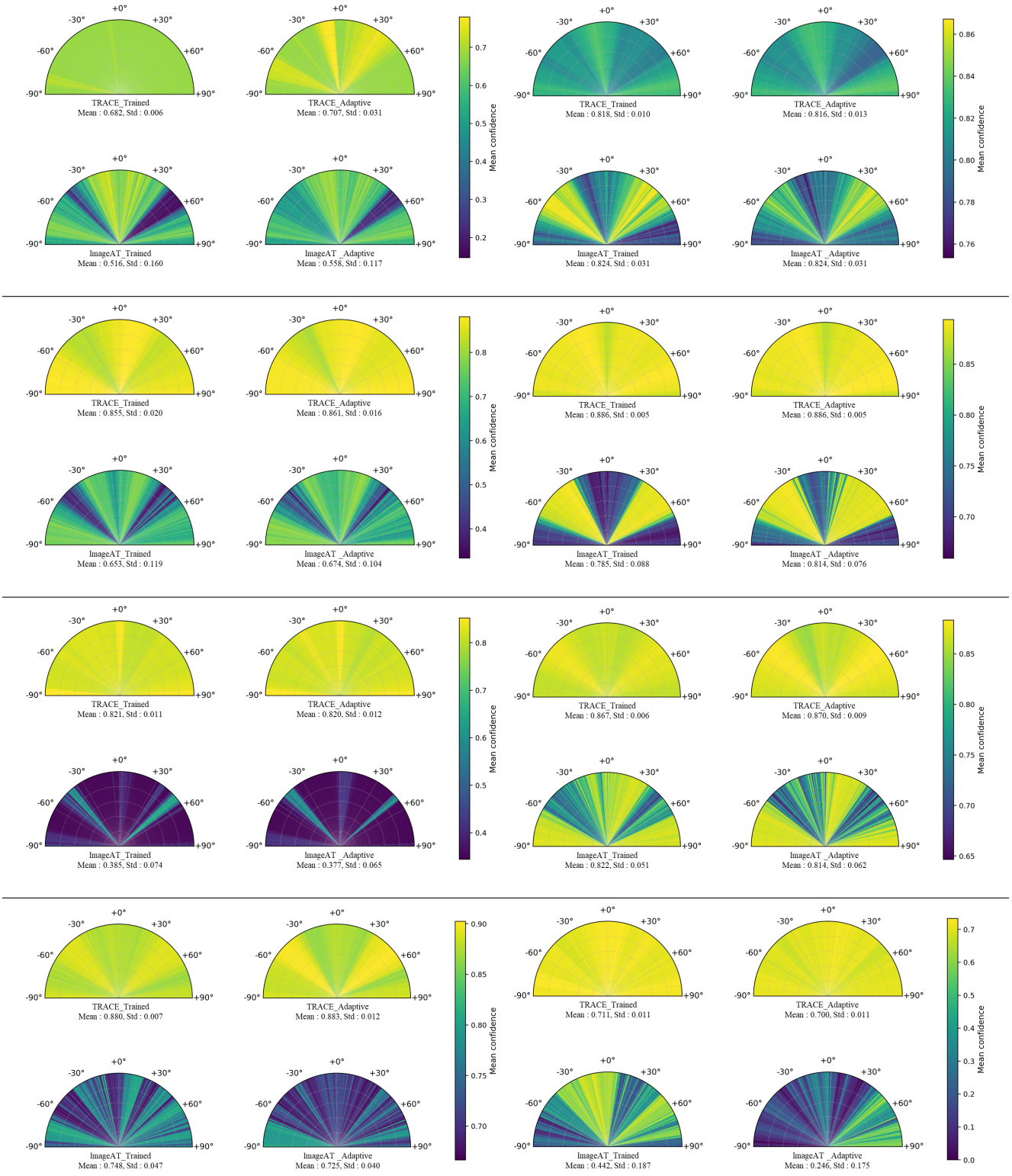


Figure 4. Visualization results of patch rotation generalization for TRACE and ImageAT.

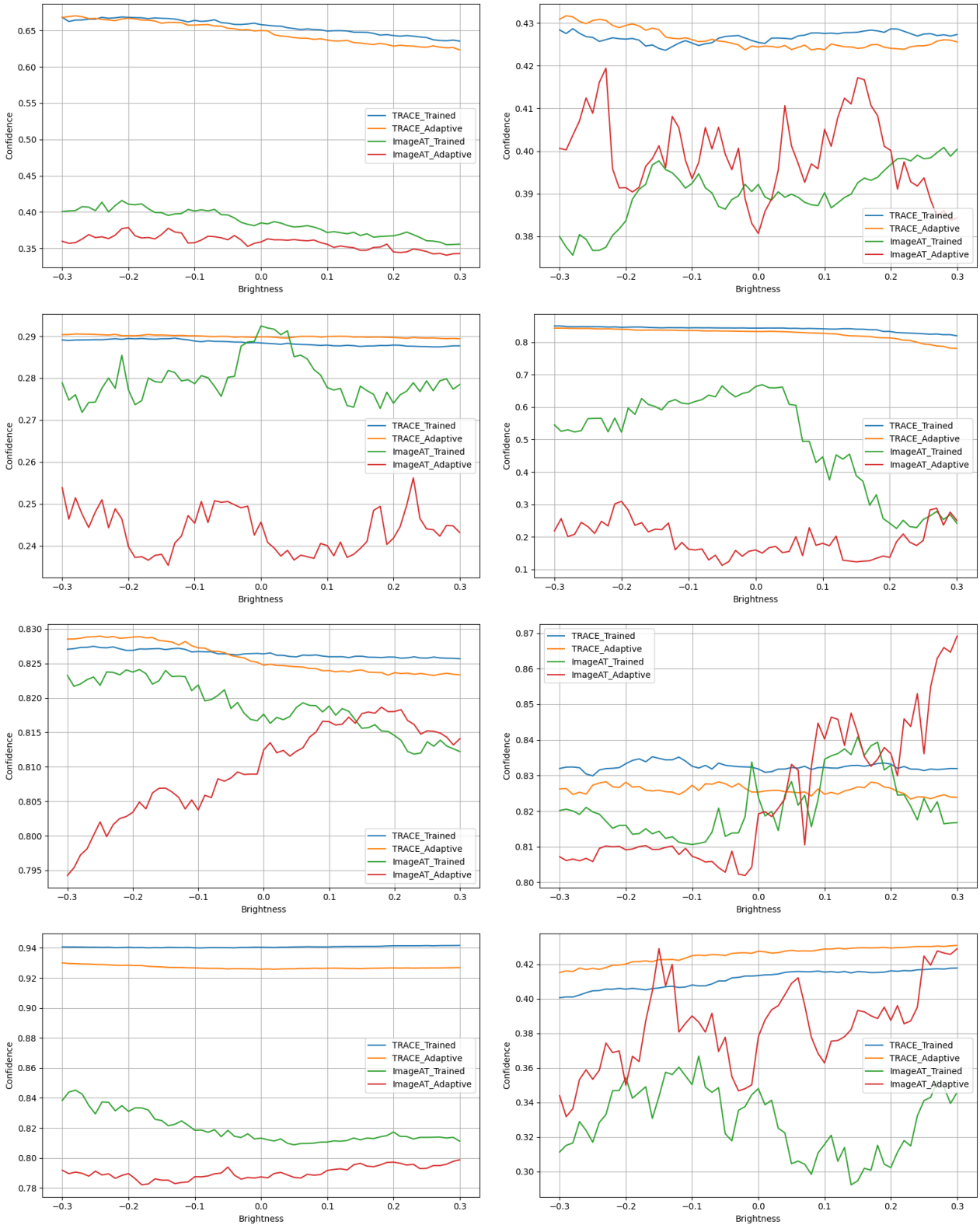


Figure 5. Visualization results of patch brightness generalization for TRACE and ImageAT.

in certain conditions but generally exhibited larger variance and lower mean scores. For instance, in the top-right of Fig. 5, TRACE demonstrated stable distributions with minimal variance across the entire range, whereas ImageAT recorded lower scores and displayed pronounced fluctuations depending on brightness. Similarly, in the second-row-left and third-row-right plots, ImageAT temporarily outperformed TRACE in specific brightness intervals, but the performance variation was substantial and overall levels remained lower. By comparison, TRACE consistently maintained higher confidence across all conditions.

In conclusion, ImageAT showed limited strengths in specific intervals but unstable behavior overall, whereas TRACE preserved robust and consistent performance across the full range of brightness variation.

When integrating the visualization analyses across all three confounders, TRACE consistently maintained stable and uniform confidence distributions, while ImageAT repeatedly revealed localized weaknesses and performance limited to specific conditions. In the location experiments, ImageAT exhibited excessive reliance on certain object regions, while in the rotation experiments, it showed pronounced vulnerability at angles outside the training range. Under brightness variation, ImageAT occasionally exceeded TRACE but suffered from high variance and lower average performance, reducing reliability. By contrast, TRACE achieved consistently higher mean scores with lower variance under all conditions, demonstrating stable generalization even for unseen scenarios. Therefore, TRACE provides more generalized robustness to diverse environmental variations compared to ImageAT.

5. TRACE in Real-world Scenario

To evaluate whether TRACE maintains robustness in physical environments, we constructed an unmanned store testbed and conducted physical experiments. The Supplementary Materials provide detailed descriptions of the physical setting and experimental procedures.

The testbed was built by connecting a webcam to NVIDIA's Xavier board, and the snack dataset [5], which consists of 21 categories, was used for model training. A Base Model was trained on this dataset, after which adversarial fine-tuning with ImageAT and TRACE was performed for comparison. All experiments were conducted under conditions closely resembling real-world scenarios. To ensure fairness, adversarial patches were applied at the same locations while maintaining consistent camera angles, lighting, and background conditions across all models. The patches were moved across objects within the camera's field of view to perform attacks, and the detection results and confidence variations of each model were measured. Real-time detection results were stored in both text logs and video recordings, with successful hiding attacks logged as "No detection." Quantitative evaluation was based on Attack Success Rate (ASR), defined as the proportion of frames labeled "No detection" within the last 100 frames of each test sequence. For detected objects, the confidence score was also recorded to assess not only whether detection occurred but also the stability of detection confidence. The results of the physical experiments conducted in the unmanned store testbed are shown in Fig. 6. Each figure illustrates the detection outcomes for the Base Model(No Attack, Attack), ImageAT, and TRACE (from left to right), with the numbers above bounding boxes indicating confidence scores. Fig. 6 shows that the Base Model was generally vulnerable to physical attacks, recording high ASR under hiding attack scenarios. In contrast, ImageAT and TRACE more frequently maintained detection compared to the Base Model. Notably, TRACE exhibited smaller decreases in confidence scores and consistently higher stability than ImageAT. Specifically, in the first three cases of Fig. 6, the Base Model failed to detect objects due to hiding attacks, while both ImageAT and TRACE maintained detection. However, ImageAT suffered large drops in confidence, whereas TRACE preserved higher scores. In the fourth case, both the Base Model and ImageAT were fully compromised by the attack, yet TRACE detected objects reliably. In the final case, all three models succeeded in detection, but the Base Model and ImageAT recorded low confidence scores, while TRACE maintained higher values, reflecting stable performance.

In conclusion, TRACE achieved higher detection performance and reliability than both the Base Model and ImageAT in physical environments, demonstrating strong robustness against hiding attacks. These results confirm that TRACE is not only effective in digital environments but also consistently robust in real-world scenarios. Furthermore, since TRACE operates without additional computational overhead or external modules during inference, it can be efficiently deployed even on resource-constrained edge devices such as Xavier. Beyond unmanned store scenarios, the characteristics of TRACE suggest strong potential for broader deployment in diverse real-world scenarios, providing a reliable foundation for adversarial robustness in practical applications.

6. Limitations and Future Work

In this section, we discuss the limitations of TRACE, potential solutions to address them, and future research directions. To enhance the inherent robustness of models against adversarial attacks, including adversarial patches, training with adversarial examples is essential. While such training improves robustness, it inevitably introduces a trade-off with clean accuracy. As described in Sec. 2, TRACE offers an advantage over ImageAT by achieving stronger robustness against adversarial patches with less degradation

Figure 6. Physical robustness experiments of Base Model(No Attack, Attack), ImageAT, TRACE (left to right).

of clean accuracy. However, compared to the untuned Base Model, both ImageAT and TRACE exhibited an approximately 10% reduction in performance due to the clean accuracy trade-off. Recent advances in adversarial fine-tuning methods [3, 7] have actively explored strategies to minimize damage to clean accuracy. Incorporating these approaches into TRACE could help mitigate the trade-offs observed in our works.

TRACE also involves a large number of hyperparameters, from patch generation to the adversarial fine-tuning process. In particular, parameters in the causal inversion step, perturbation budget ϵ, update step size α, and update number t, have a direct influence on performance. A comprehensive exploration of these parameters is therefore expected to further improve TRACE's effectiveness.

To validate the effectiveness of TRACE, we employed Yolov5 and Yolov8 models trained on the COCO dataset. COCO is one of the most representative datasets for object detection, containing 80 classes and a wide distribution of samples. Similarly, YOLO models are widely used in real-world applications due to their fast inference speed and high accuracy in real-time object detection. However, further experiments on more diverse models and datasets are required to verify the broader applicability of TRACE. Therefore, we plan to conduct additional experiments on a broader range of datasets and models to confirm the generalizability of TRACE.

References

[1] Guijin Han, RuiXuan Wang, MengChun Zhou, and Jun Li. Enhancing semantic and spatial information yolov8. Available at SSRN: https://ssrn.com/abstract=4797816, 2024. 2

[2] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. pages 1414–1423. PMLR, 2017. 1

[3] A. Jeddi, M. J. Shafiee, and A. Wong. A simple fine-tuning

is all you need: Towards robust deep learning via adversarial fine-tuning. *arXiv preprint arXiv:2012.13628*, 2020. 8

[4] Krikamol Muandet, Arash Mehrjou, Si Kai Lee, and Anant Raj. Dual instrumental variable regression. *NeurIPS*, 33: 2710–2721, 2020. 1

[5] Korea Nazarene University. Snack dataset. https://universe.roboflow.com/korea-nazarene-university/-d9kpq, 2023. visited on 2025-09-10. 7

[6] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *CVPRW*, pages 390–391, 2020. 2

[7] S. Wang, J. Zhang, Z. Yuan, and S. Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *CVPR*, pages 24502–24511, 2024. 8