WACV 2026

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

Soongsil University

Wonho Lee, Jisu Lee, Hyunsik Na, Sohee Park, Daeseon Choi

WACV 2026
Tucson, Arizona • Mar 6 - 10, 2026

Soongsil University
AI Safety Center

Introduction

- **As AI becomes more widely used, its inherent security vulnerabilities have emerged as a critical concern**
  - Adversarial Patch Attacks pose a critical threat to object detection systems by causing severe mispredictions in both digital and physical environments

Liu, X., Yang, H., Liu, Z., Song, L., Li, H., & Chen, Y. (2018). Dpatch: An adversarial patch attack on object detectors. arXiv preprint arXiv:1806.02299.

Introduction

➢ Limitation of Existing Defenses

- Certified, Detection-based defense
  - High latency and Computational overhead during inference
- Conventional(Pixel-space) adversarial training
  - Overfits to specific patches
  - Lacks generalization to unseen or physical-world attacks

➢ The Root Cause

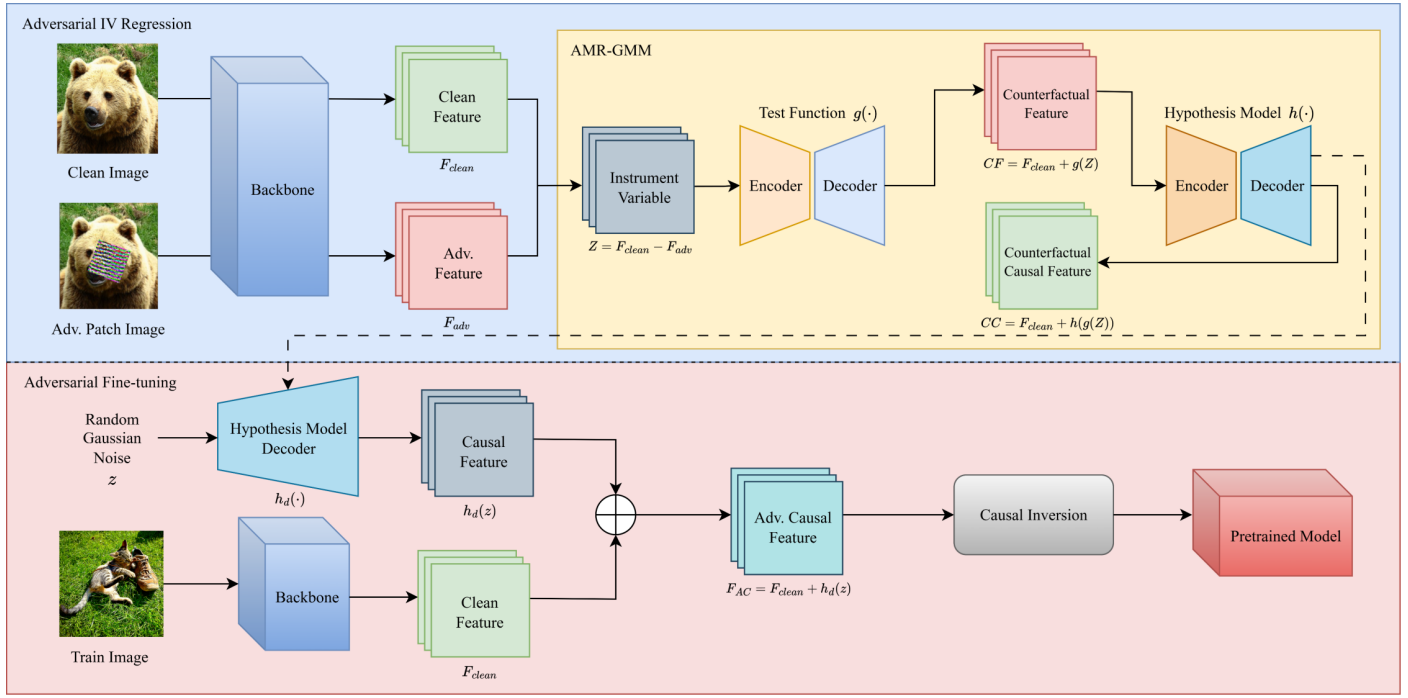
- Variations in patch (location, rotation, brightness, ⋯)

Confounders

- Confounders create spurious correlations that mislead the model
- It can be critical inherent vulnerabilities

Proposed Framework - TRACE

- **Definition : Tuning Robustness by Adversarial patch Confounder Elimination**
  - The first adversarial fine-tuning framework for object detection based on Instrumental Variable (IV) Regression
- **Objective**
  - Eliminate patch-related confounders
  - Guide the model toward causal features that sustain robust detection



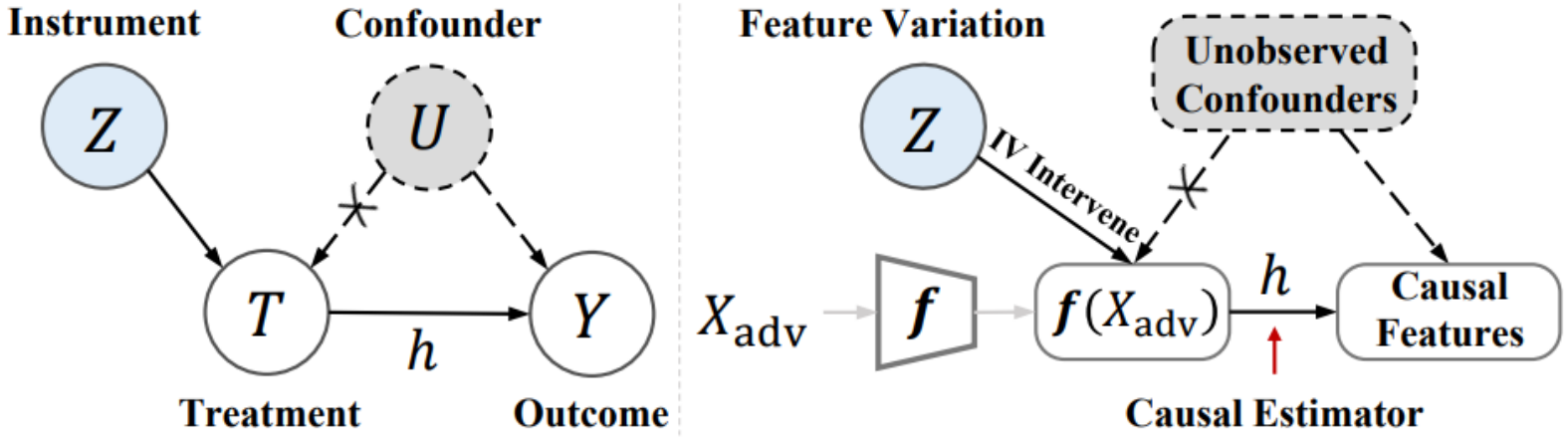
Motivation

➢ IV Regression

- Removes backdoor paths and estimate causal effects when confounders cannot be controlled

➢ IV Definition

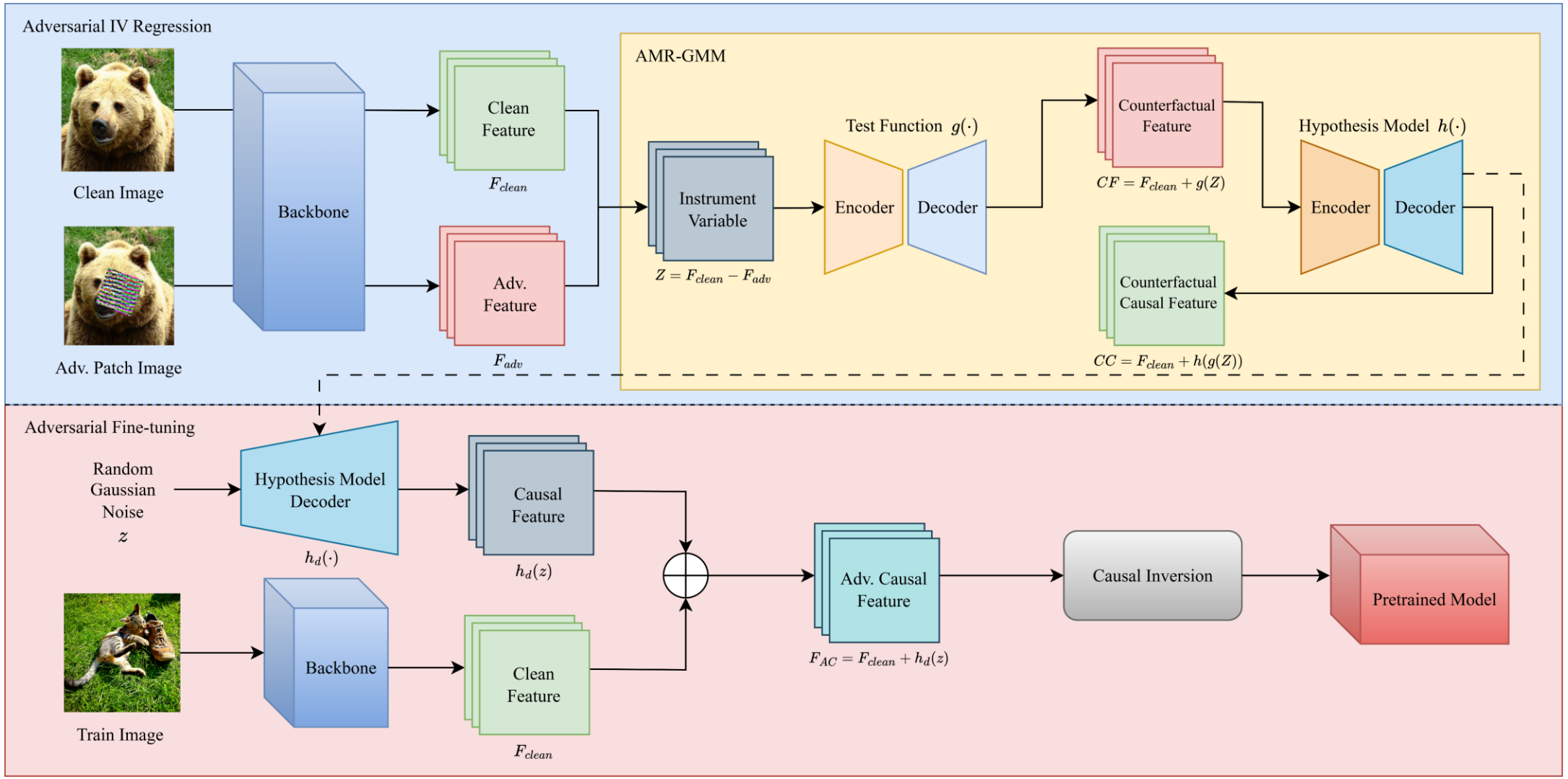
- Define instrument variable Z as feature variation from adversarial perturbation
  - $Z = F_{adv} - F_{natural}$



Kim, J., Lee, B. K., & Ro, Y. M. (2023). Demystifying causal features on adversarial examples and causal inoculation for robust network by adversarial instrumental variable regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12302-12312).

TRACE : Tuning Robustness by Adversarial patch Confounder Elimination

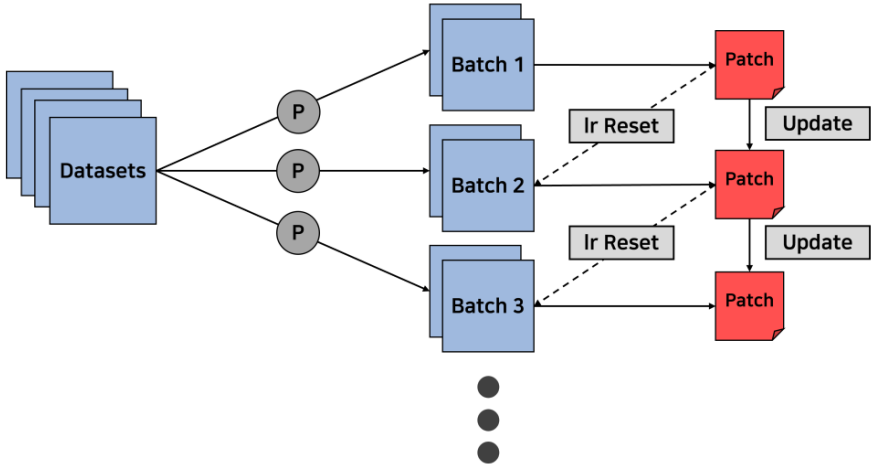
Soongsil University
AI Safety Center



TRACE : Tuning Robustness by Adversarial patch Confounder Elimination

➢ Adversarial IV Regression

- Generate Adversarial Patch
  - For strong generalization, faster generation, and larger output we adopt IPG
  - By sampling partial data with a Poisson Sampler, IPG reduces batch dependency and generate diverse adversarial patches
  - IPG can help TRACE to better cover a broader spectrum of vulnerabilities

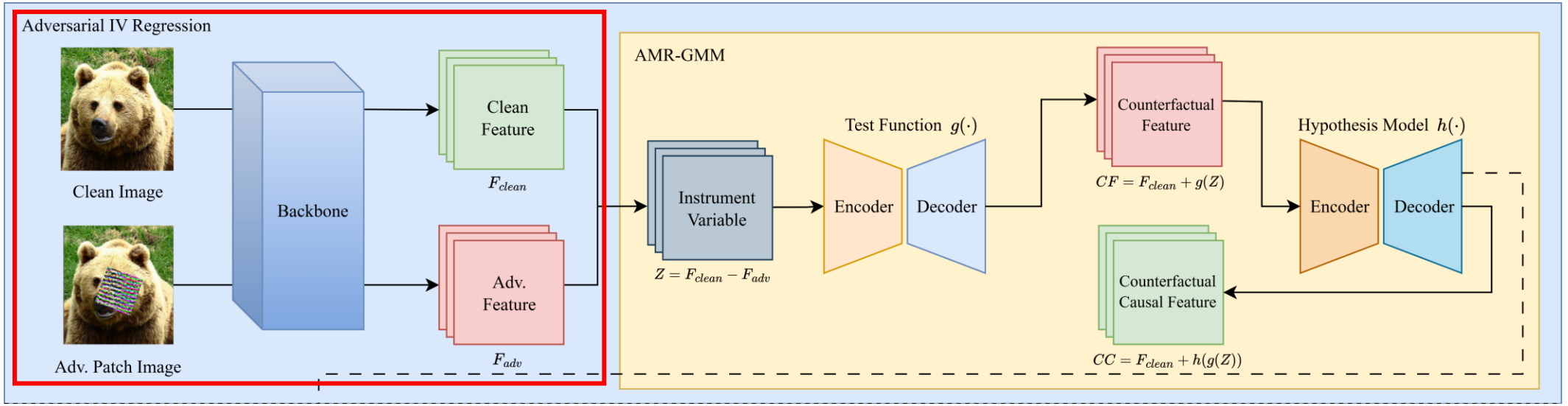


Lee, W., Na, H., Lee, J., & Choi, D. (2025). IPG: Incremental Patch Generation for Generalized Adversarial Patch Training. arXiv preprint arXiv:2508.10946.

TRACE : Tuning Robustness by Adversarial patch Confounder Elimination

➢ Adversarial IV Regression

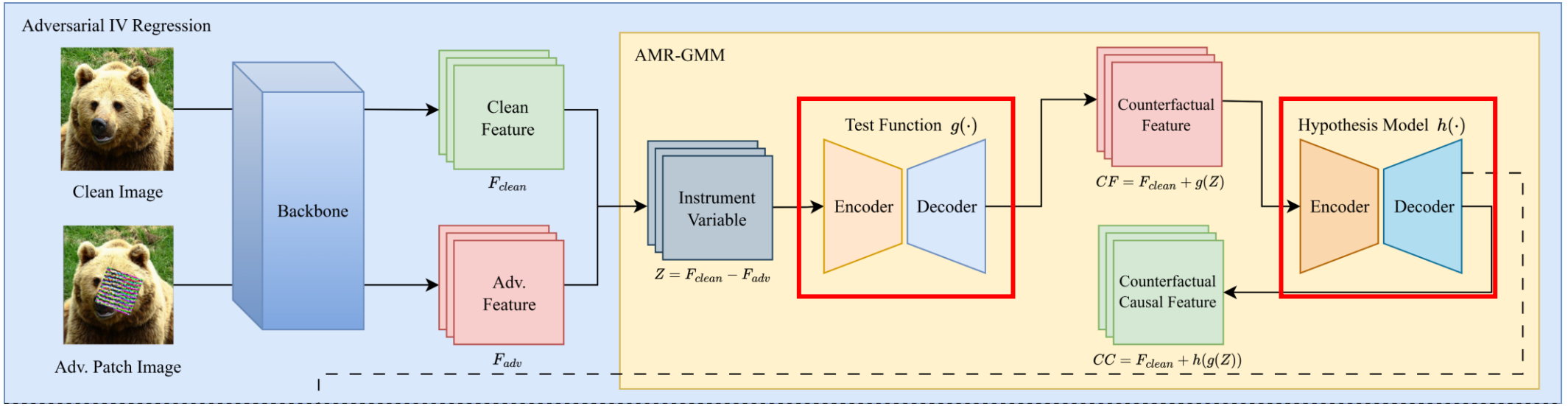
- Feature-space Processing
  - TRACE targets the backbone stage where pixel perturbations are first projecting into feature space



TRACE : Tuning Robustness by Adversarial patch Confounder Elimination

➢ Adversarial IV Regression

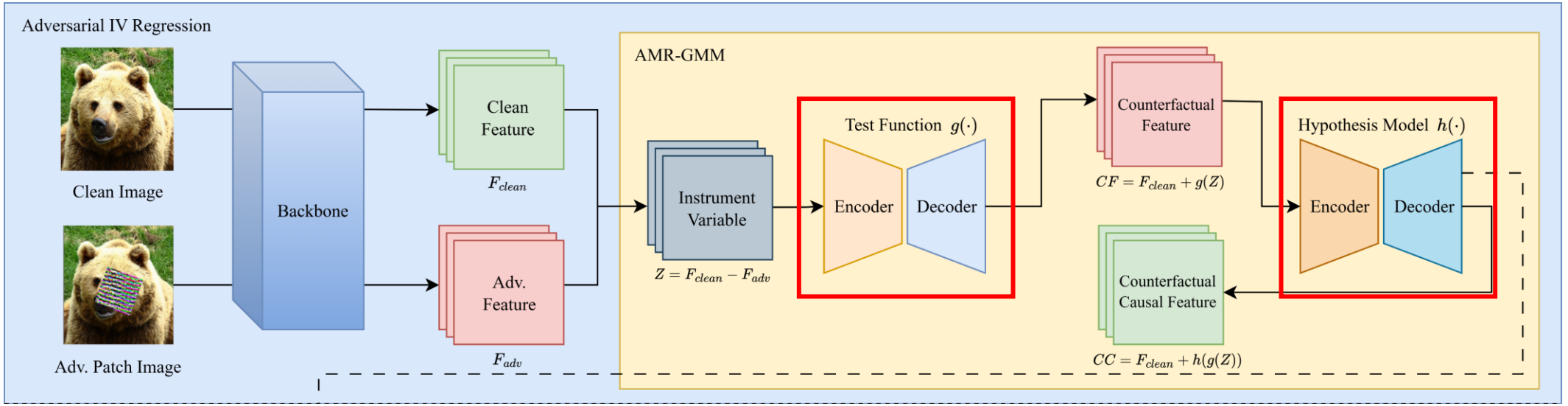
- VAE Structure
  - Employs a Variational Auto-Encoder to probabilistically generate patch-induced feature variations in latent space



TRACE : Tuning Robustness by Adversarial patch Confounder Elimination

➢ Adversarial IV Regression

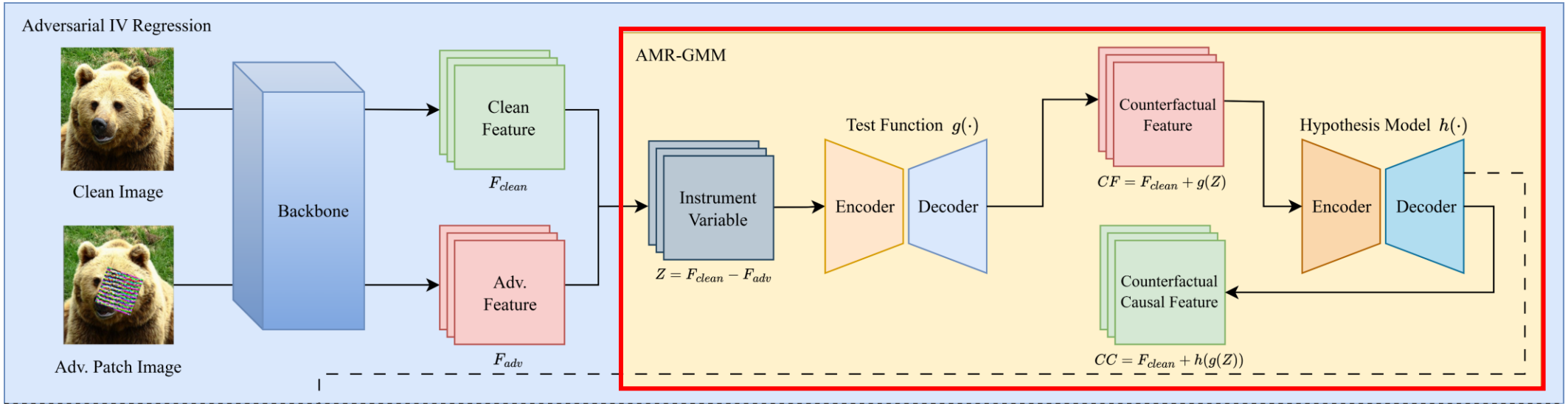
- Test Function $g(\cdot)$: generates worst-case counterfactuals features
- Hypothesis Model $h(\cdot)$: reconstructs stable causal feature from these counterfactuals that contribute to correct detection



TRACE : Tuning Robustness by Adversarial patch Confounder Elimination

➢ AMR-GMM Optimization

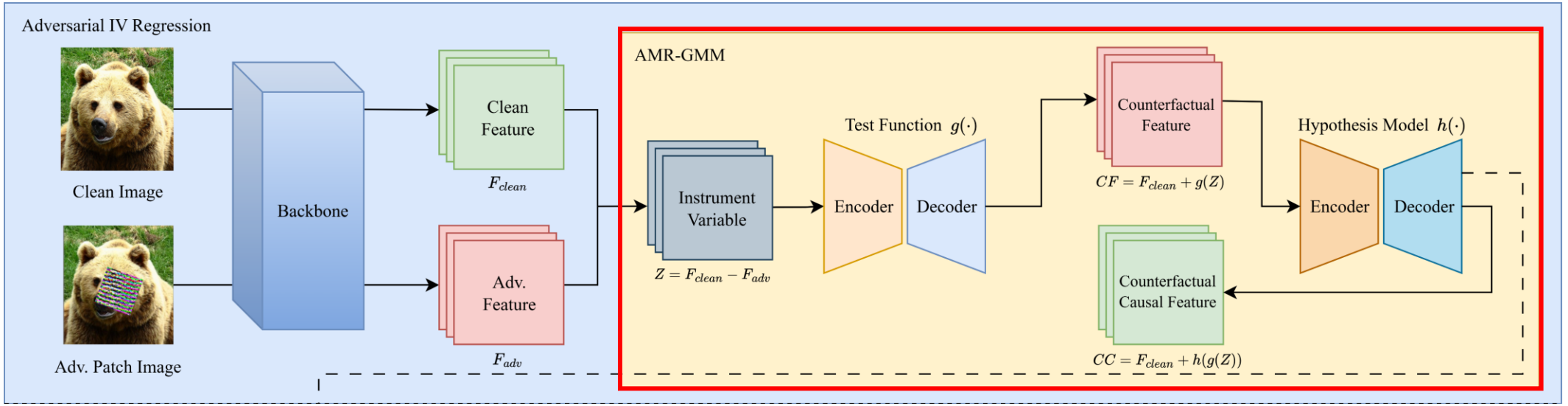
- Detection Objective
  - Unlike classification, TRACE uses a redesigned $objcls$ metric the product of objectness and class probability



TRACE : Tuning Robustness by Adversarial patch Confounder Elimination

➢ AMR-GMM Optimization

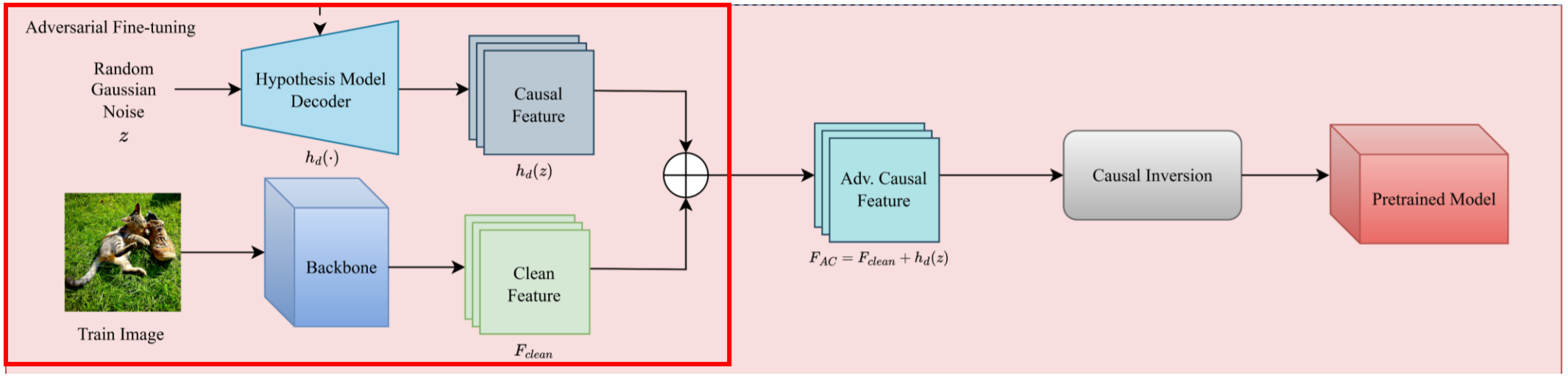
- A zero-sum game where the test function minimizes $objcls$ and the hypothesis model maximizes it to restore detection ability



TRACE : Tuning Robustness by Adversarial patch Confounder Elimination

➢ Adversarial Fine-tuning

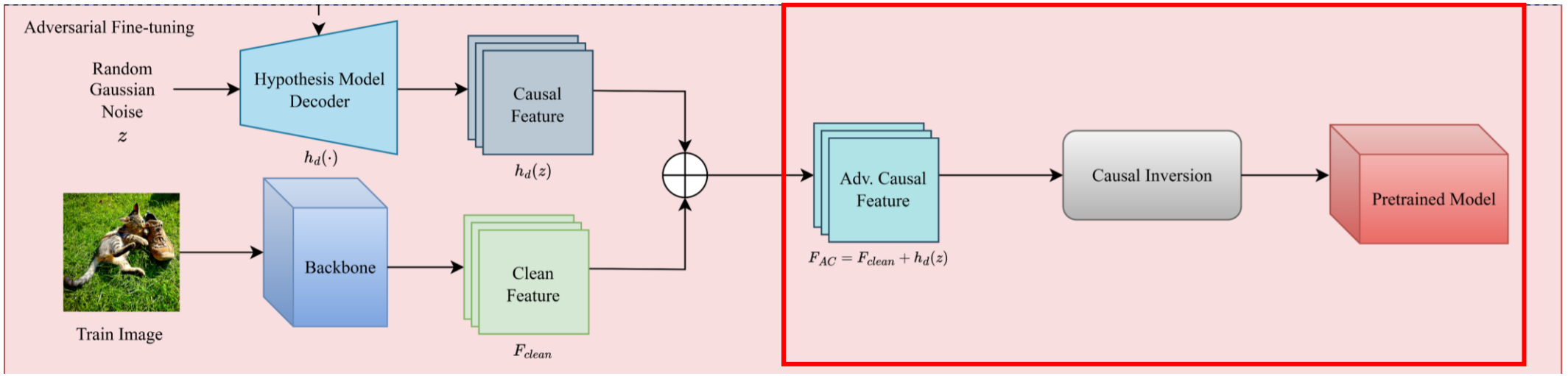
- The decoder of trained Hypothesis Model $h_d(\cdot)$
  - work as causal feature generator
- Adapt pretrained detectors efficiently with limited (10%) data



TRACE : Tuning Robustness by Adversarial patch Confounder Elimination

➢ Adversarial Fine-tuning

- Causal Inversion
  - Projects the learned Adversarial Causal Feature distribution back onto the input image
- Efficient Fine-tuning



Experiments

- **Model**
  - Yolov5
  - Yolov8
- **Datasets**
  - MS-COCO 2017
- **Baseline**
  - ImageAT
    - Conventional(Pixel-space) Adversarial Tuning with same patches

Experimental Results

➢ Generalization

- TRACE demonstrates a significantly smaller generalization gap than ImageAT under Trained and Adaptive patches

			Base Model	Trained		Adaptive	
				ImageAT	TRACE	ImageAT	TRACE
Yolov5	No Attack	*mAP*	1.0	1.0		1.0	
		mAR	1.0	1.0		1.0	
	Random Noise	*mAP*	0.348 ± 0.003	0.389 ± 0.004	0.288 ± 0.003	0.389 ± 0.004	0.288 ± 0.003
		mAR	0.408 ± 0.003	0.461 ± 0.003	0.348 ± 0.004	0.461 ± 0.003	0.348 ± 0.004
		ASR	0.525 ± 0.002	0.544 ± 0.002	0.545 ± 0.002	0.544 ± 0.002	0.545 ± 0.002
	Adversarial Patch	*mAP*	0.263 ± 0.014	0.379 ± 0.004	0.243 ± 0.010	0.176 ± 0.007	0.170 ± 0.006
		mAR	0.310 ± 0.017	0.455 ± 0.004	0.294 ± 0.011	0.230 ± 0.008	0.208 ± 0.006
		ASR	0.621 ± 0.013	0.553 ± 0.002	0.599 ± 0.011	0.764 ± 0.007	0.691 ± 0.006
Yolov8	No Attack	*mAP*	1.0	1.0		1.0	
		mAR	1.0	1.0		1.0	
	Random Noise	*mAP*	0.403 ± 0.003	0.407 ± 0.002	0.325 ± 0.003	0.407 ± 0.002	0.325 ± 0.003
		mAR	0.465 ± 0.003	0.480 ± 0.002	0.398 ± 0.003	0.480 ± 0.002	0.398 ± 0.003
		ASR	0.493 ± 0.002	0.544 ± 0.002	0.545 ± 0.002	0.544 ± 0.002	0.545 ± 0.002
	Adversarial Patch	*mAP*	0.342 ± 0.012	0.405 ± 0.002	0.283 ± 0.006	0.326 ± 0.027	0.260 ± 0.004
		mAR	0.403 ± 0.011	0.480 ± 0.002	0.353 ± 0.006	0.395 ± 0.028	0.320 ± 0.005
		ASR	0.550 ± 0.010	0.546 ± 0.002	0.596 ± 0.007	0.616 ± 0.027	0.630 ± 0.005

Table 1. Comparison of ImageAT and TRACE on Trained and Adaptive adversarial patches (mean ± std over 50 patches). Higher values of mAP and mAR indicate better performance, whereas lower ASR is preferable.

Experimental Results

➢ Stability

- Maintains lower variations in Attack Success Rate (ASR) across diverse patch distribution compared to ImageAT

			Base Model	Trained		Adaptive	
				ImageAT	TRACE	ImageAT	TRACE
Yolov5	No Attack	*mAP*	1.0	1.0		1.0	
		mAR	1.0	1.0		1.0	
	Random Noise	*mAP*	0.348 ± 0.003	0.389 ± 0.004	0.288 ± 0.003	0.389 ± 0.004	0.288 ± 0.003
		mAR	0.408 ± 0.003	0.461 ± 0.003	0.348 ± 0.004	0.461 ± 0.003	0.348 ± 0.004
		ASR	0.525 ± 0.002	0.544 ± 0.002	0.545 ± 0.002	0.544 ± 0.002	0.545 ± 0.002
	Adversarial Patch	*mAP*	0.263 ± 0.014	0.379 ± 0.004	0.243 ± 0.010	0.176 ± 0.007	0.170 ± 0.006
		mAR	0.310 ± 0.017	0.455 ± 0.004	0.294 ± 0.011	0.230 ± 0.008	0.208 ± 0.006
		ASR	0.621 ± 0.013	0.553 ± 0.002	0.599 ± 0.011	0.764 ± 0.007	0.691 ± 0.006
Yolov8	No Attack	*mAP*	1.0	1.0		1.0	
		mAR	1.0	1.0		1.0	
	Random Noise	*mAP*	0.403 ± 0.003	0.407 ± 0.002	0.325 ± 0.003	0.407 ± 0.002	0.325 ± 0.003
		mAR	0.465 ± 0.003	0.480 ± 0.002	0.398 ± 0.003	0.480 ± 0.002	0.398 ± 0.003
		ASR	0.493 ± 0.002	0.544 ± 0.002	0.545 ± 0.002	0.544 ± 0.002	0.545 ± 0.002
	Adversarial Patch	*mAP*	0.342 ± 0.012	0.405 ± 0.002	0.283 ± 0.006	0.326 ± 0.027	0.260 ± 0.004
		mAR	0.403 ± 0.011	0.480 ± 0.002	0.353 ± 0.006	0.395 ± 0.028	0.320 ± 0.005
		ASR	0.550 ± 0.010	0.546 ± 0.002	0.596 ± 0.007	0.616 ± 0.027	0.630 ± 0.005

Table 1. Comparison of ImageAT and TRACE on Trained and Adaptive adversarial patches (mean ± std over 50 patches). Higher values of mAP and mAR indicate better performance, whereas lower ASR is preferable.

Experimental Results

➢ Unseen Attack

- Also robust to Unseen Attack patches

	mAP	mAR	ASR
ImageAT	0.186 ± 0.009	0.236 ± 0.010	0.762 ± 0.009
TRACE	0.199 ± 0.007	0.243 ± 0.007	0.657 ± 0.007

Table 2. Unseen Attack(NPSTV) results of ImageAT and TRACE

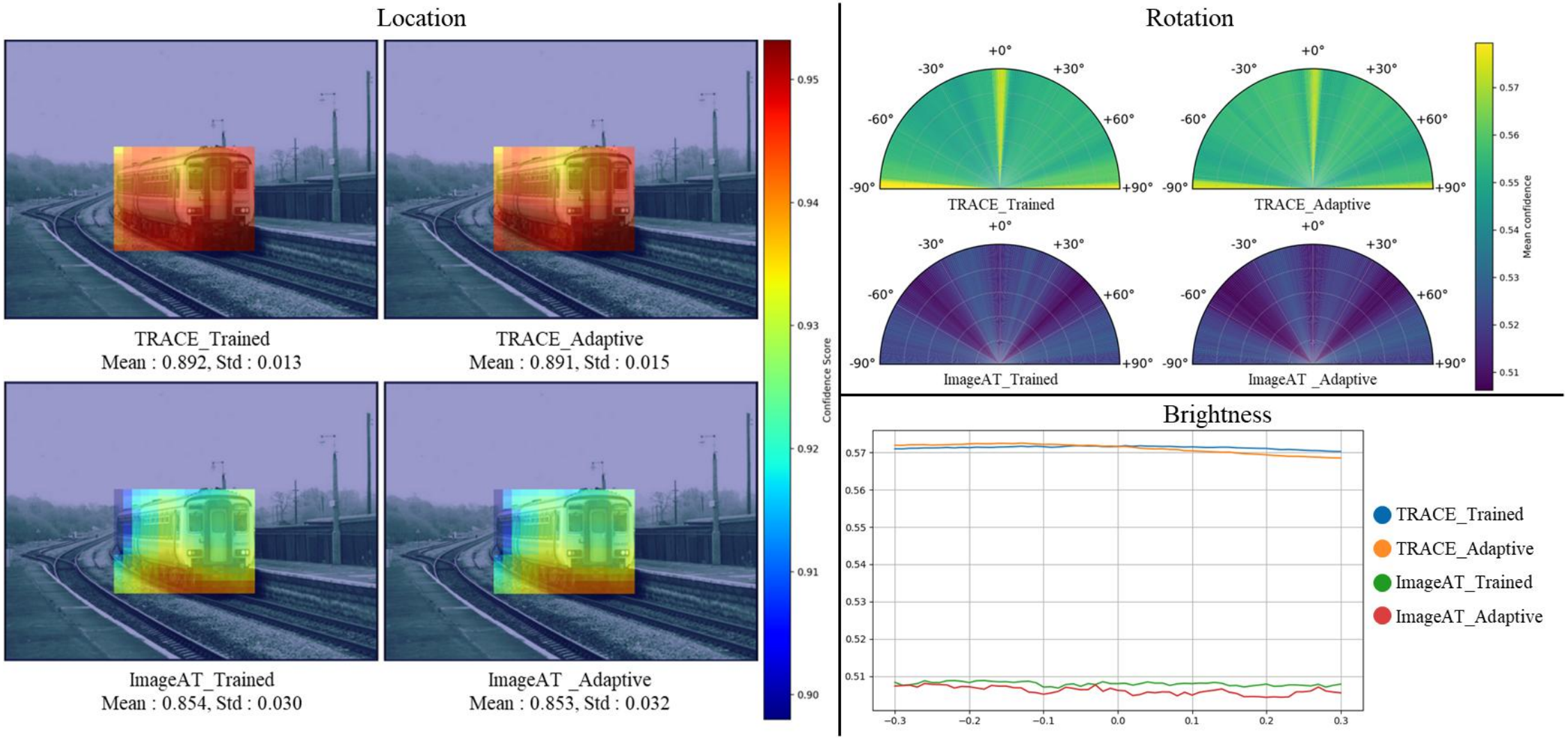
Experimental Results

- **Qualitative Evaluation**
  - Focusing on location, rotation, and brightness
- **Using 500 single-object images from the COCO validation set**

Confounder	Model		Mean ± Std
Location	ImageAT	Trained	0.513 ± 0.209
		Adaptive	0.512 ± 0.208
	TRACE	Trained	0.576 ± 0.195
		Adaptive	0.576 ± 0.197
Rotation	ImageAT	Trained	0.512 ± 0.222
		Adaptive	0.511 ± 0.222
	TRACE	Trained	0.538 ± 0.231
		Adaptive	0.538 ± 0.232
Brightness	ImageAT	Trained	0.508 ± 0.010
		Adaptive	0.506 ± 0.010
	TRACE	Trained	0.571 ± 0.008
		Adaptive	0.571 ± 0.007

Table 3. Generalization Evaluation about three confounders.

Experimental Results



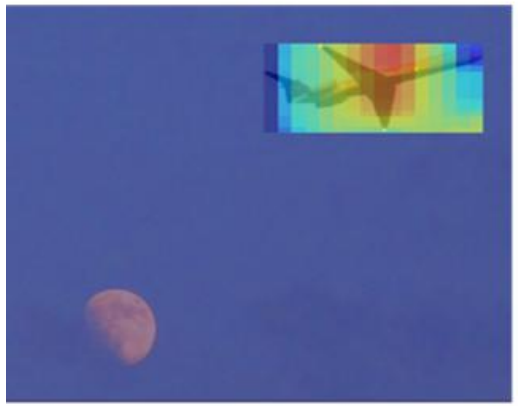
Experimental Results

TRACE_Trained
Mean : 0.708, Std : 0.040

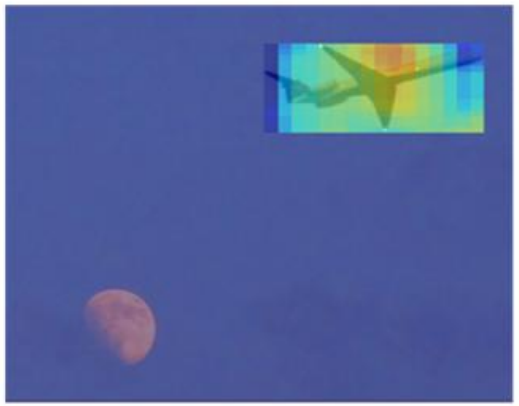
TRACE_Adaptive
Mean : 0.722, Std : 0.059

TRACE_Trained
Mean : 0.860, Std : 0.038

TRACE_Adaptive
Mean : 0.865, Std : 0.035



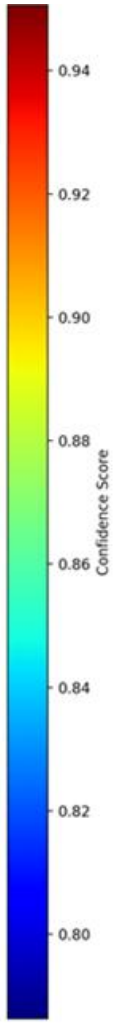
ImageAT_Trained
Mean : 0.559, Std : 0.132



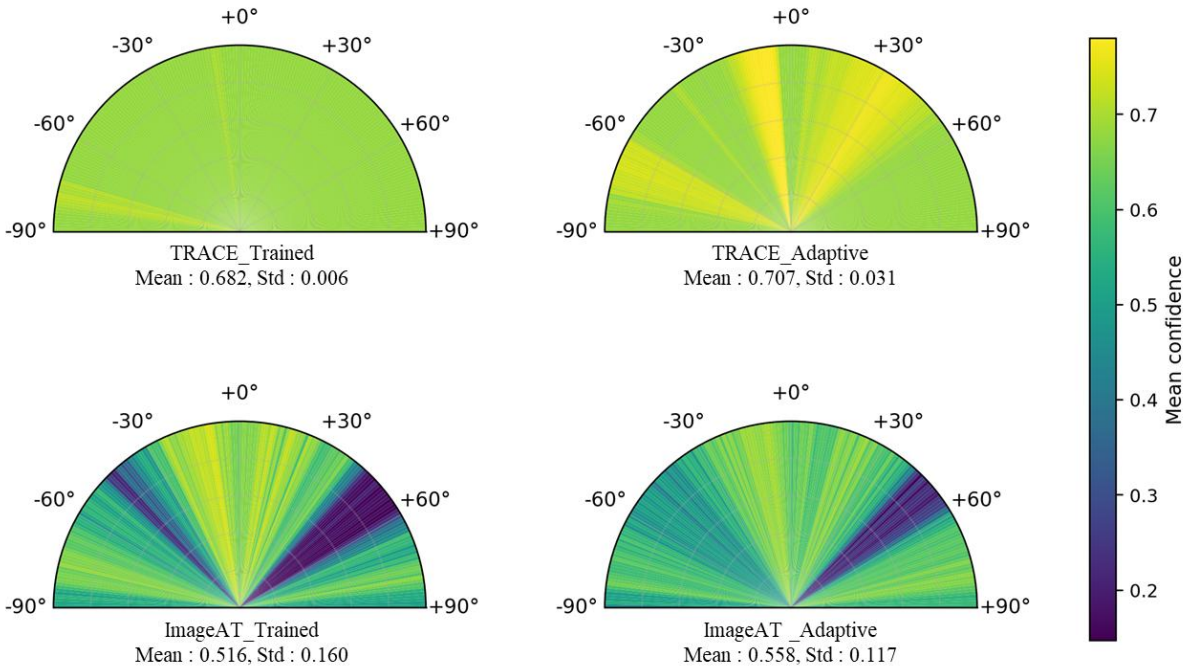
ImageAT _Adaptive
Mean : 0.517, Std : 0.119

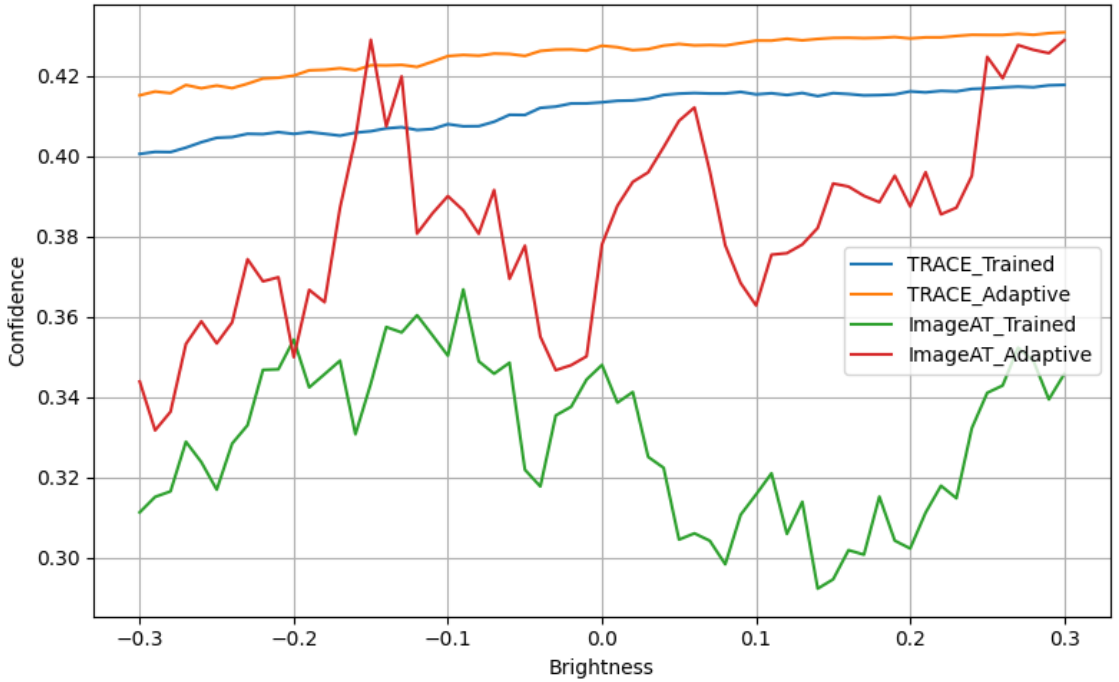
ImageAT_Trained
Mean : 0.687, Std : 0.093

ImageAT _Adaptive
Mean : 0.692, Std : 0.100



Experimental Results





Physical Test

➢ **Experiments on a custom unmanned-store testbed**

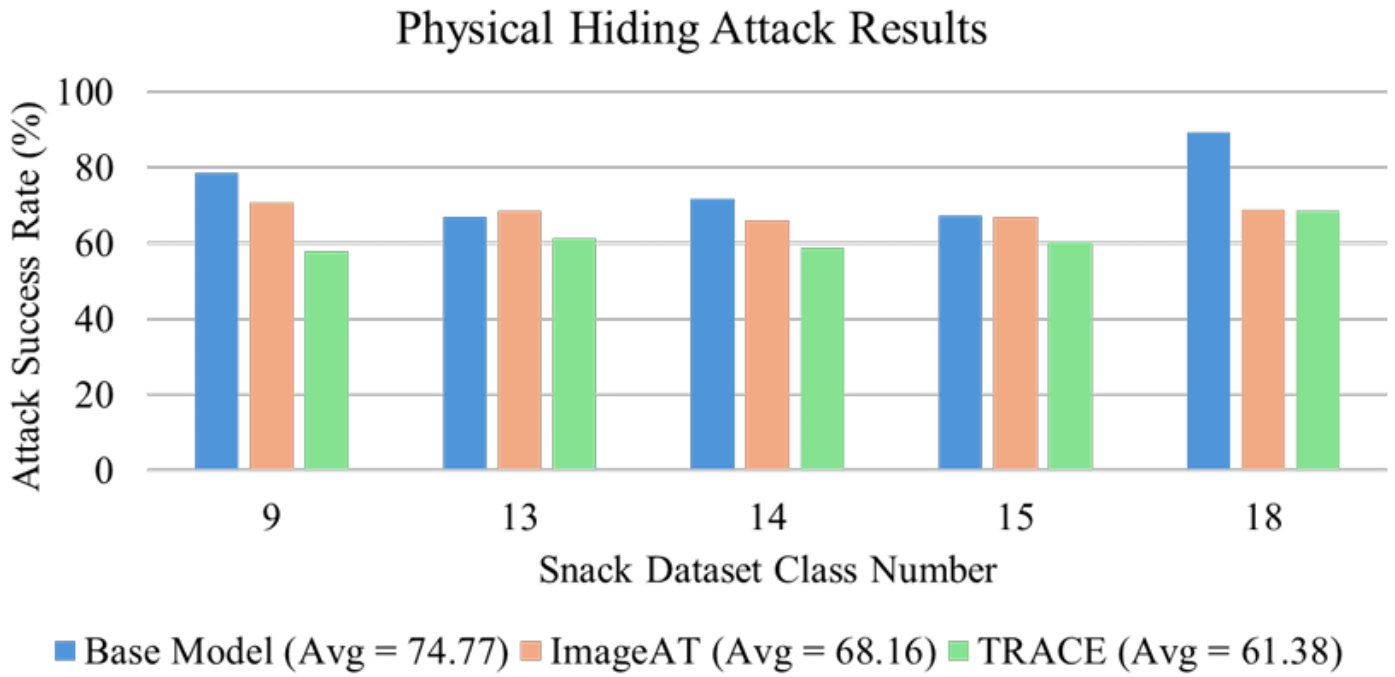
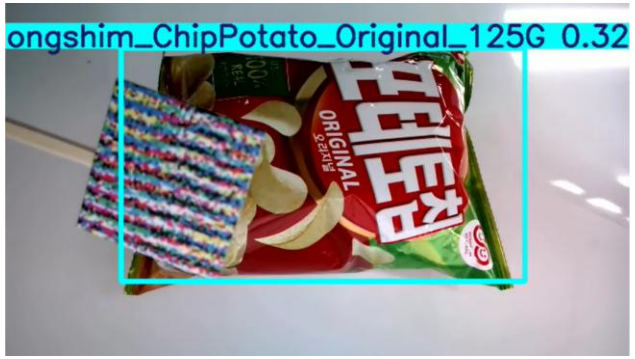


Figure 3. Physical environment results. Lower ASR indicates stronger robustness.

Physical Test

Soongsil University
AI Safety Center

➢ Experiments on a custom unmanned-store testbed

Base Model (No Attack)	Base Model (Attack)	ImageAT	TRACE
			
			

Conclusion & Contribution

- **Framework**
  - TRACE is the first IV Regression-based adversarial fine-tuning framework for object detection
- **Efficiency**
  - Leverages knowledge in large pretrained models and requires minimal training data compared to training from scratch
- **Impact**
  - Provides generalized robustness beyond patch-specific defenses, enhancing reliability in real-world scenarios

Thank you!

Soongsil University
AI Safety Center

➢ For More Information

- Main Paper & Supplementary Materials

- Contact
  - Wonho Lee
  - Soongsil University, Republic of Korea
  - Email : hoho0907@soongsil.ac.kr