

Hybrid-FGPG: A Novel Fusion of Gradient and Geometric Transformations for Adversarial Vulnerability Evaluation

Dimitrios Christos Asimopoulos, Panagiotis Radoglou-
Grammatikis, Vasileios Argyriou, Thomas Lagkas, Pantelis
Angelidis, Vasileios Vitsas, Panagiotis Fouliras, Ioannis Ktenidis,
and **Panagiotis Sarigiannidis**

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework AMYNA-TN - Grant Agreement no. ΥΠ3ΤΑ-0560908.

Authors & Contributors

Funded



Dimitrios-
Christos Asimopoulos



Thomas Lagkas



Dimitrios-Christos Asimopoulos,
Vasileios Vitsas



Vasileios Argyriou



Panagiotis Radoglou Grammatikis,
Pantelis Angelidis, Ioannis Ktenidis
Panagiotis Sarigiannidis



Ioannis Moscholios



Panagiotis Radoglou Grammatikis



Greece 2.0
NATIONAL RECOVERY AND RESILIENCE PLAN



Funded by the
European Union
NextGenerationEU



This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework AMYNA-TN - Grant Agreement no. ΥΠ3ΤΑ-0560908.

Presentation Structure

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework AMYNA-TN - Grant Agreement no. ΥΠ3ΤΑ-0560908.

IEEE GlobeCom// Taipei, Taiwan 8-12 December, 2025



Introduction

1

Introduction

Related Work

Problem
Statement &
Research gap

Contributions

Main Part

2

Architecture

Adversarial Attack Methodologies

Hybrid Attack

Evaluation & Results

Conclusions

3

Sum up important
key points

Future Work

Q/A

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework AMYNA-TN - Grant Agreement no. YΠ3TA-0560908.

Introduction, Relevant Work & Contributions

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework AMYNA-TN - Grant Agreement no. YΠ3TA-0560908.

IEEE GlobeCom// Taipei, Taiwan 8-12 December, 2025

Introduction



Artificial intelligence models are widely deployed across critical domains such as autonomous driving, traffic systems, and satellite monitoring.



Adversarial examples are small, intentionally crafted perturbations that deceive AI models into producing incorrect predictions. These attacks can remain visually imperceptible while causing serious misclassifications.

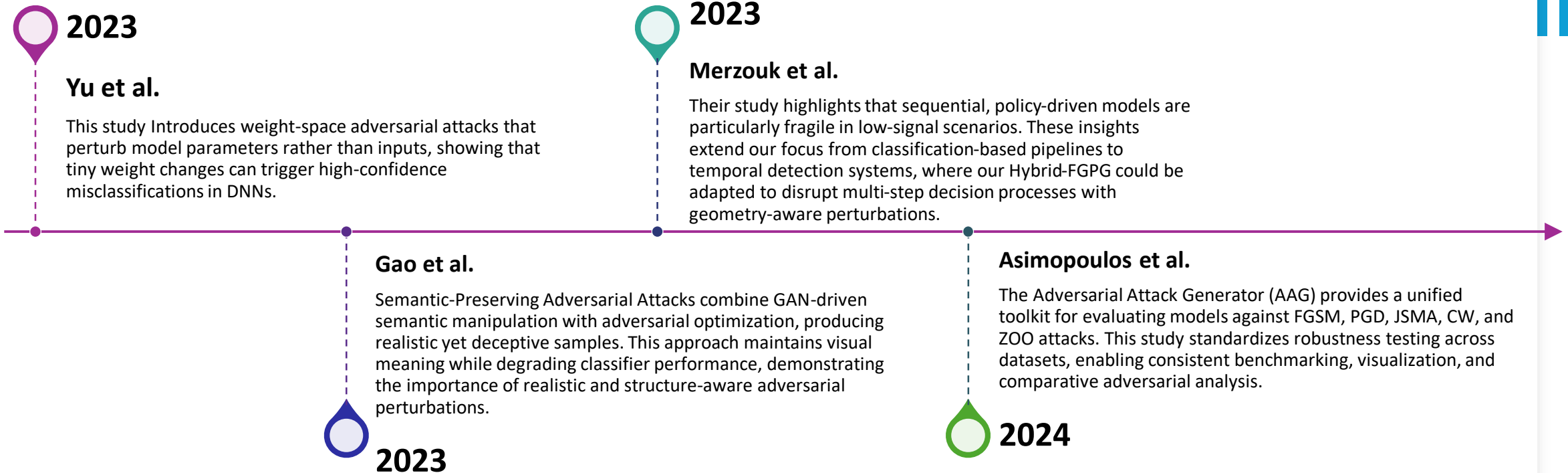


In safety-critical domains, even a single misclassification can lead to severe consequences—misinterpreting a stop sign, overlooking a wildfire, or mislabeling satellite imagery.



Traditional gradient-based methods like FGSM and PGD explore limited perturbation directions and rarely account for natural variations such as small rotations or geometric distortions. This gap motivates the development of more realistic and effective adversarial attacks.

Related Work



Problem Statement & Research Gap



Problem Statement

- Existing adversarial attacks rely primarily on **gradient direction**, limiting exploration of alternative perturbation paths.
- They fail to capture **natural variations** such as rotations, viewpoint changes, or sensor distortions.
- Generated perturbations often reduce accuracy, yet remain **unnatural, easily detectable**, or inconsistent with real-world conditions.



Research Gap

- No existing approach effectively explores multiple perturbation subspaces beyond simple gradient ascent.
- Lack of methods combining gradient information with geometric transformations to enhance adversarial strength.
- Need for attacks that achieve high misclassification while staying visually subtle, realistic, and difficult to detect.

Contributions

Hybrid-FGPG: A Novel Gradient + Geometric Adversarial Attack

- We propose Hybrid-FGPG, a new adversarial attack that fuses gradient-based perturbations with pixel-level geometric rotation, enabling exploration of side-directions in the loss landscape..

Modular Adversarial Evaluation Framework

- We design a modular, extensible evaluation pipeline supporting FGSM, PGD, and Hybrid attacks across CIFAR-10, GTSRB, and EuroSAT datasets.
- Cross domain Experimental Validation

Hybrid attacks vs Popular adversarial attacks

- We compare our proposed adversarial attack with traditional adversarial attacks such as FGSM and PGD

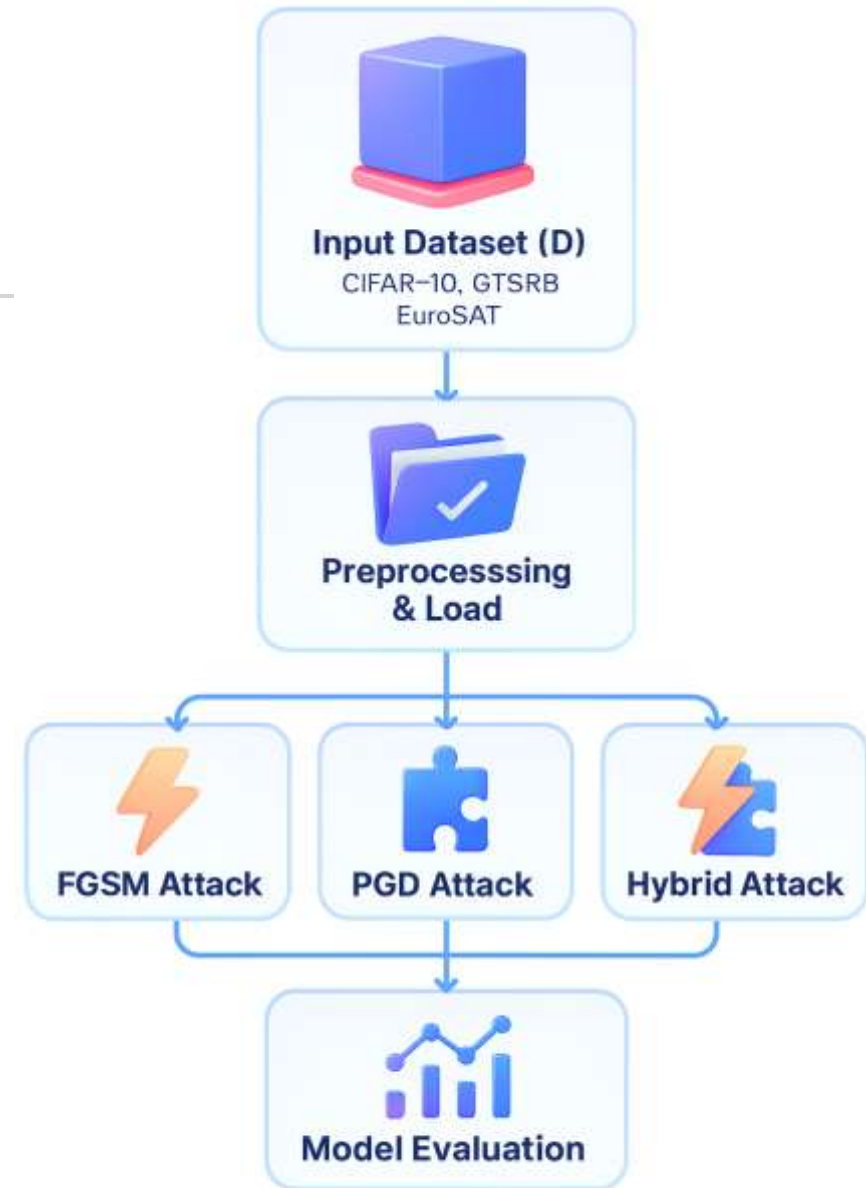
Adversarial Attack Generator Architectural Design

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework AMYNA-TN - Grant Agreement no. YΠ3TA-0560908.

IEEE GlobeCom// Taipei, Taiwan 8-12 December, 2025

Architecture

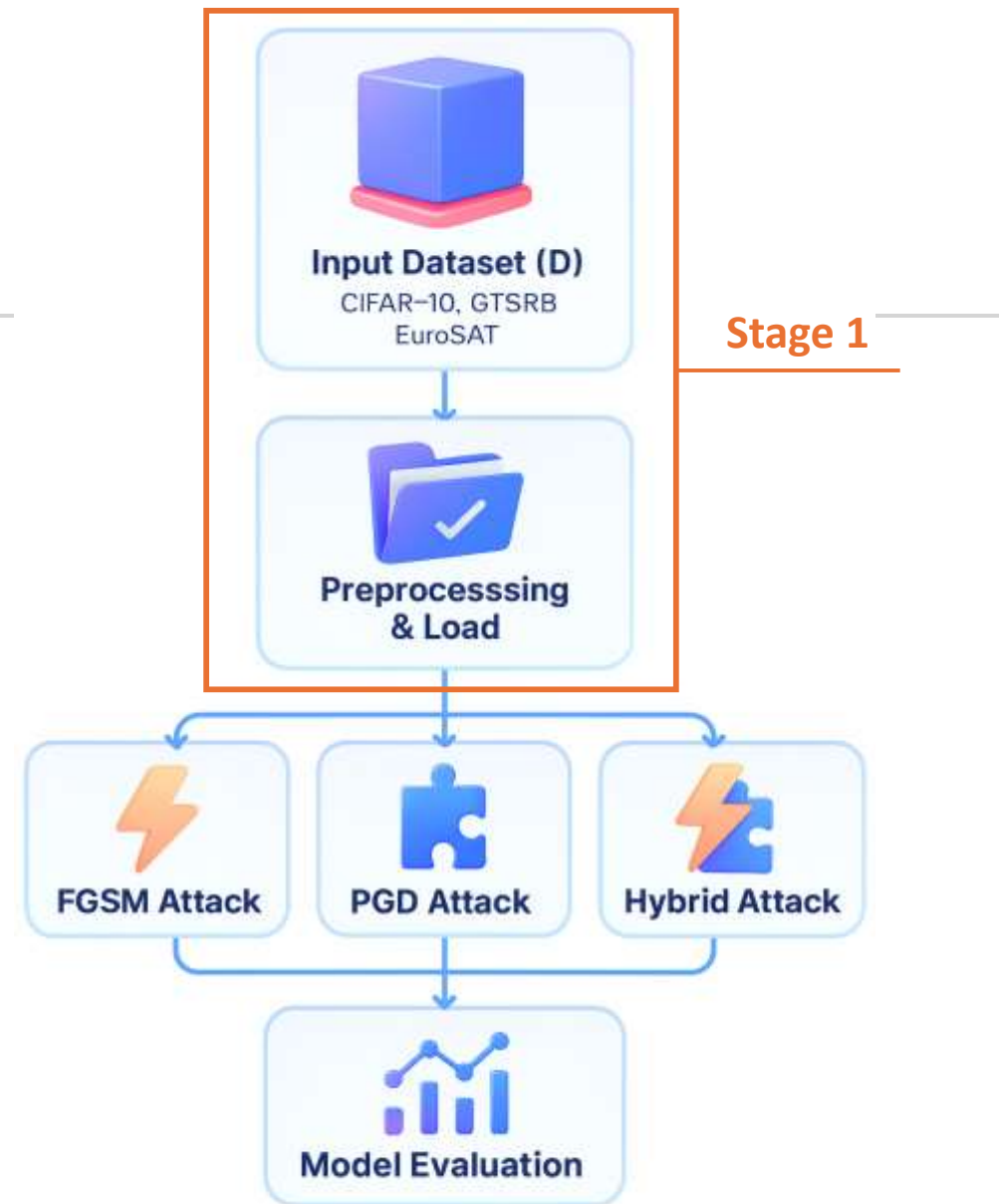
- **Modular, extensible, and reproducible design.**
- **Supports multiple datasets (CIFAR-10, GTSRB, EuroSAT).**
- **Enables streamlined experimentation across various attacks.**
- **Three core stages:**
 - a) Model Initialization
 - b) Attack Execution
 - c) Metrics Evaluation & Visualization



Model Initialization

Model Initialization Phase (Stage 1)

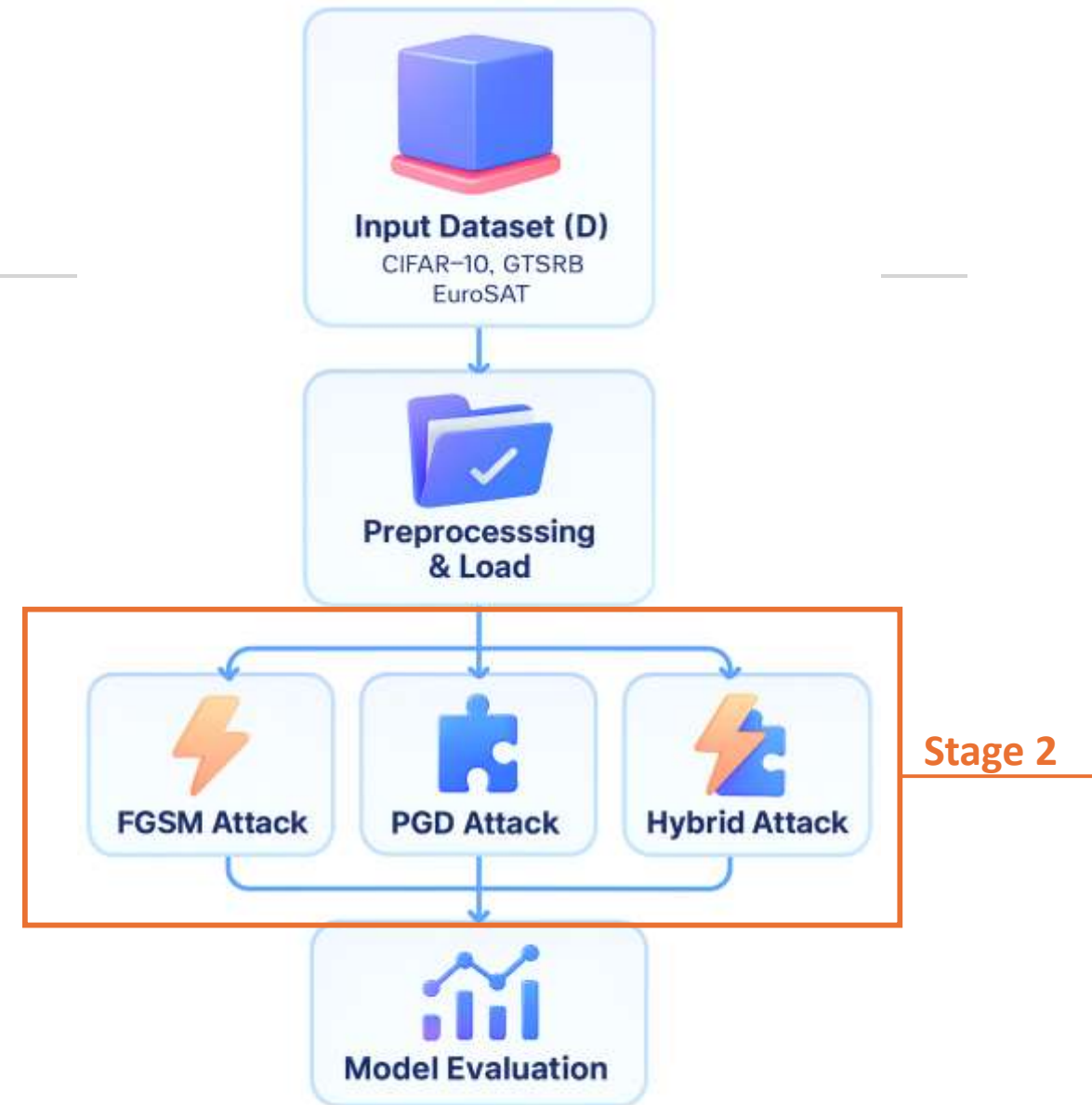
- Loads **ResNet-18** (pretrained or trained from scratch).
- Configured for cross-dataset image classification.
- Optimized for clean inference before attacks are applied.
- Ensures consistent baseline performance across experiments.



Attack Execution

Attack Execution Phase (stage 2)

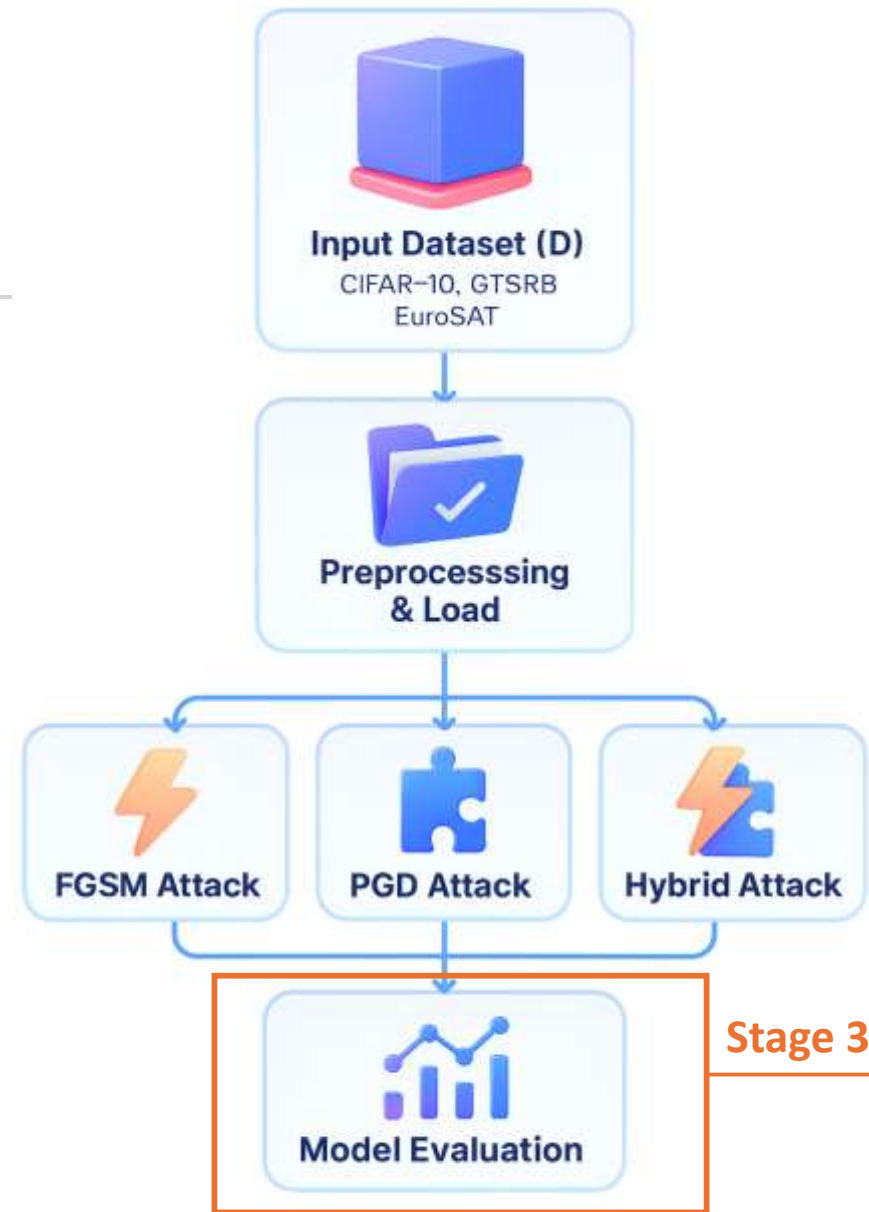
- Applies adversarial methods to clean image batches.
- Supported attacks:
 - **FGSM** (single-step gradient)
 - **PGD** (iterative gradient ascent)
 - **Hybrid-FGPG** (gradient + rotational perturbation)
- Each attack perturbs input within controlled constraints.
- Produces adversarial examples for evaluation.



Model Evaluation

Model Evaluation Phase (stage 3)

- Compares predictions on clean vs adversarial images.
- Computes standard metrics:
Accuracy, Precision, Recall, F1-Score
- Quantifies the effect of each attack on model robustness.
- Provides qualitative insight into each attack's behavior.
- Saves representative examples per dataset and attack type.
- Displays: Clean image, Perturbed image, Predicted labels (clean vs adversarial), Heatmaps



Adversarial Attacks

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework AMYNA-TN - Grant Agreement no. YΠ3TA-0560908.

IEEE GlobeCom// Taipei, Taiwan 8-12 December, 2025

Fast Gradient Sign Method (FGSM)

Overview

FGSM is a fundamental adversarial attack technique that exploits vulnerabilities in machine learning models by adding targeted noise to input data.

- FGSM generates adversarial examples by adding a small, crafted perturbation to the input.
- Perturbation is calculated by taking a step in the direction of the gradient sign of the loss function, maximizing prediction error.
- Efficient and widely used for evaluating the robustness of machine learning models.

Characteristics

- Fast and computationally inexpensive.
- Highly effective for small ϵ .
- Limited exploration of perturbation space (single direction).

FGSM Process



1. Input Data (e.g., Image)



2. Loss Function & Gradient Calculation



3. Sign of Gradient (Direction)



4. Adversarial Example (Original + $\epsilon \cdot \text{sign}(\nabla J)$)

$$\mathbf{x}_{\text{adv}} = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, \mathbf{y}))$$

$\text{adv_x} \rightarrow$ Adversarial data

$\mathbf{x} \rightarrow$ Original data

$\mathbf{y} \rightarrow$ Original input label

$\epsilon \rightarrow$ Multiplier to ensure the perturbations are small

$\theta \rightarrow$ Model parameters

$J \rightarrow$ Loss function. In our case, the CrossEntropy function

Project Gradient Descent (PGD)

Overview

PGD is an iterative, white-box attack methods that apply controlled, small perturbations to input data, aiming to mislead deep learning models into misclassification.

- Iteratively applies perturbations in the gradient direction to maximize the model's prediction error.
- PGD executes FGSM in small steps, repeatedly projecting perturbations to keep them undetectable.

Characteristics

- Stronger than FGSM
- Iterative, multi-step refinement
- Considered a "universal first-order adversary"
- Still restricted to gradient direction only

PGD Formula

FGSM

$$x_{t+1} = \Pi_{\epsilon}(x_t + \alpha \cdot \text{sign}(\nabla_x L(f(x_t), y)))$$

x_{t+1} → Updated adversarial example at iteration $t + 1$

x_t → Current adversarial example at iteration t

y → True class label of the input

α → Step size for each PGD update (small gradient step)

ϵ → Maximum allowed perturbation (radius of the ϵ -ball)

$\Pi_{\epsilon}(\cdot)$ → Projection operator that keeps the perturbation within the ϵ -ball

$\nabla_x L(f(x_t), y)$ → Gradient of loss w.r.t. the input

$f(\cdot)$ → Target neural network model

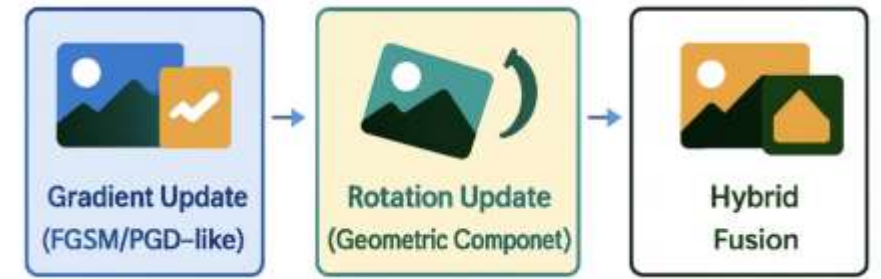
L → Loss function (CrossEntropy in our case)

Hybrid Adversarial Attack

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework AMYNA-TN - Grant Agreement no. YΠ3TA-0560908.

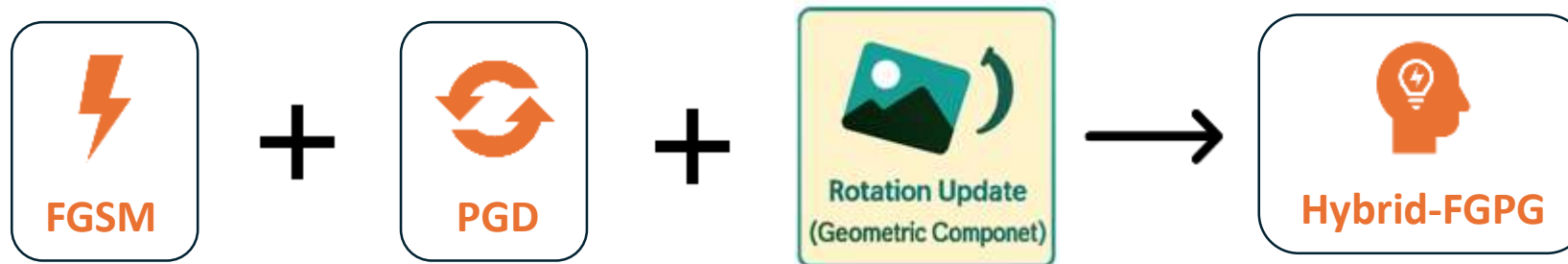
IEEE GlobeCom// Taipei, Taiwan 8-12 December, 2025

Hybrid Attack



A Novel Adversarial Attack Method

Combining Fast Gradient Sign Method and Projected Gradient Descent with perturbation guidance and rotation



This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework AMYNA-TN - Grant Agreement no. YΠ3TA-0560908.

Hybrid-FGPG: Architecture and Update Rule

Architecture Overview

-  **Fast Gradient Initialization:** Begins with FGSM-style perturbation to rapidly push input towards high-loss region
-  **Iterative Updates:** Performs a small number of updates similar to PGD (typically 5-10 steps)
-  **Perturbation Rotation:** At each step, rotates perturbation direction by angle θ using deterministic/stochastic strategy
-  **Angular Diversity:** Explores "side directions" in loss landscape, avoiding overfitting to sharp local minima

Key Advantages

- Lightweight compared to standard PGD
- Less predictable perturbations
- Better transferability

Update Rule

$$\delta_{t+1} = \text{Proje}[\delta_t + \alpha \cdot \text{sign}(\nabla_x L(f(x + \delta_t), y))] + R_\theta(\delta_t)$$

δ_t : Accumulated perturbation at iteration

α : Step size

Proje: Projects perturbation within ℓ_∞ ball

R_θ : Rotation function perturbing by angle θ

Rotation Mechanism

 Perturbation rotation introduces diversity by rotating signed gradient in a fixed 2D subspace

 Uses low-dimensional projections to escape local maxima

Evaluation & Results

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework AMYNA-TN - Grant Agreement no. YΠ3TA-0560908.

IEEE GlobeCom// Taipei, Taiwan 8-12 December, 2025

Experimental Setup

System Configuration

Implementation: PyTorch framework for all experiments

Hardware: NVIDIA 4060 GPU for computation

Evaluation Protocol: Standard dataset partitions used

Testing: Evaluation performed exclusively on test sets

Evaluation Focus

Attack effectiveness across varied domains
Performance comparison with baseline attacks
Robustness against different model architectures
Computational efficiency assessment

Datasets



CIFAR-10

Classes: 10 object categories

Format: RGB images

Resolution: 32×32 pixels

Domain: Natural images



GTSRB

Classes: 43 traffic signs

Format: RGB images

Resolution: Resized to 32×32 pixels

Domain: Traffic recognition



EuroSAT

Classes: 10 land use types

Format: RGB images

Resolution: 64×64 pixels

Domain: Satellite imagery

Evaluation Metrics

Accuracy

$$\frac{1}{N} \sum_{j=1}^N \mathbb{I}(y_j = \hat{y}_j)$$

Precision

$$\frac{1}{C} \sum_{i=1}^C \frac{TP_i}{TP_i + FP_i + \epsilon}$$

Recall

$$\frac{1}{C} \sum_{i=1}^C \frac{TP_i}{TP_i + FN_i + \epsilon}$$

F1 Score

$$\frac{1}{C} \sum_{i=1}^C \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i + \epsilon}$$

N: Total number of test samples

y_j: ground truth label

ŷ_j: predicted label

ε: small constant to prevent division by zero

TP → True Positives

TN → True Negatives

FP → False Positives

FN → False Negatives

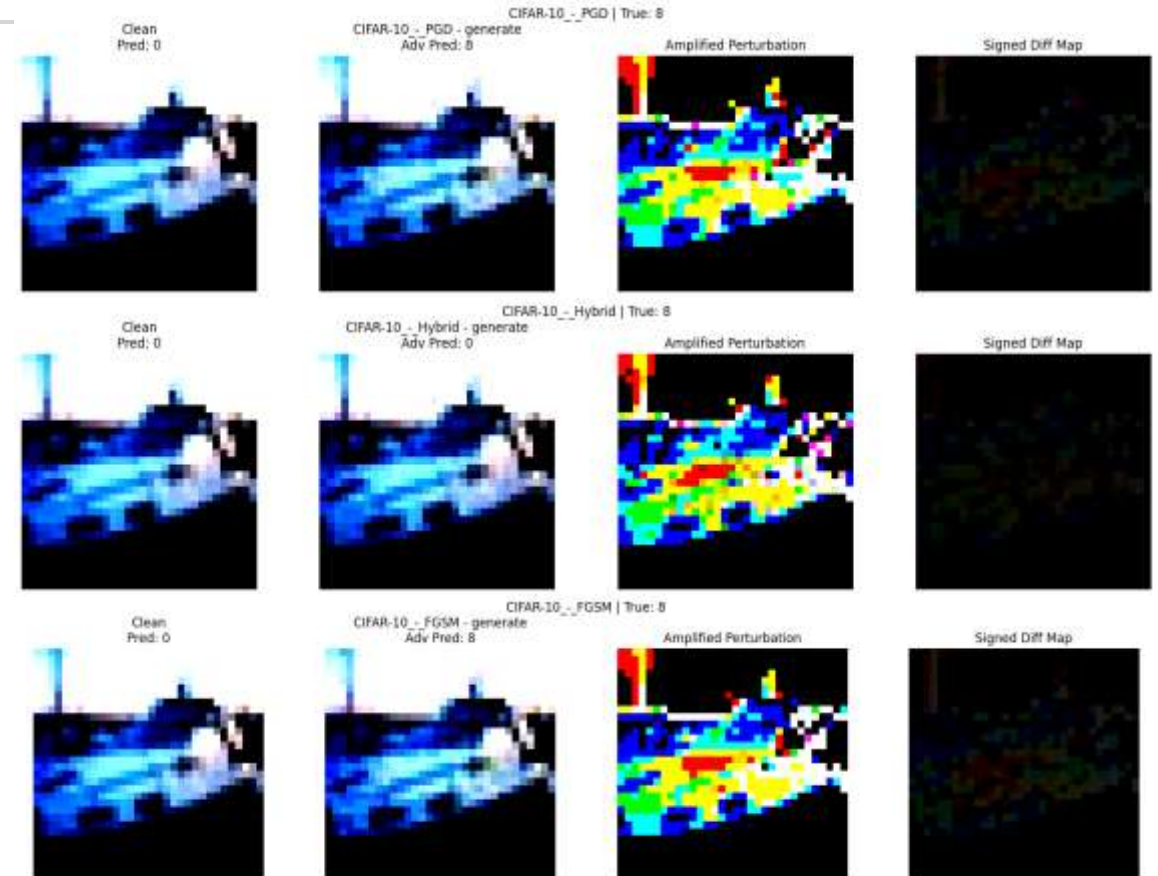
This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework AMYNA-TN - Grant Agreement no. YΠ3TA-0560908.

Experimental Results

CIFAR-10 Classification Performance

Scenario	Accuracy	Precision	Recall	F1-score
Clean	0.81	0.82	0.81	0.81
Adv – Hybrid	0.12	0.18	0.13	0.11
Adv- FGSM	0.62	0.63	0.61	0.61
Adv - PGD	0.58	0.59	0.57	0.57

The CIFAR-10 results show a dramatic performance drop under the Hybrid attack, with accuracy falling from 0.81 to 0.12, compared to 0.62 under FGSM. This highlights the Hybrid method's stronger attack potential on natural image datasets.



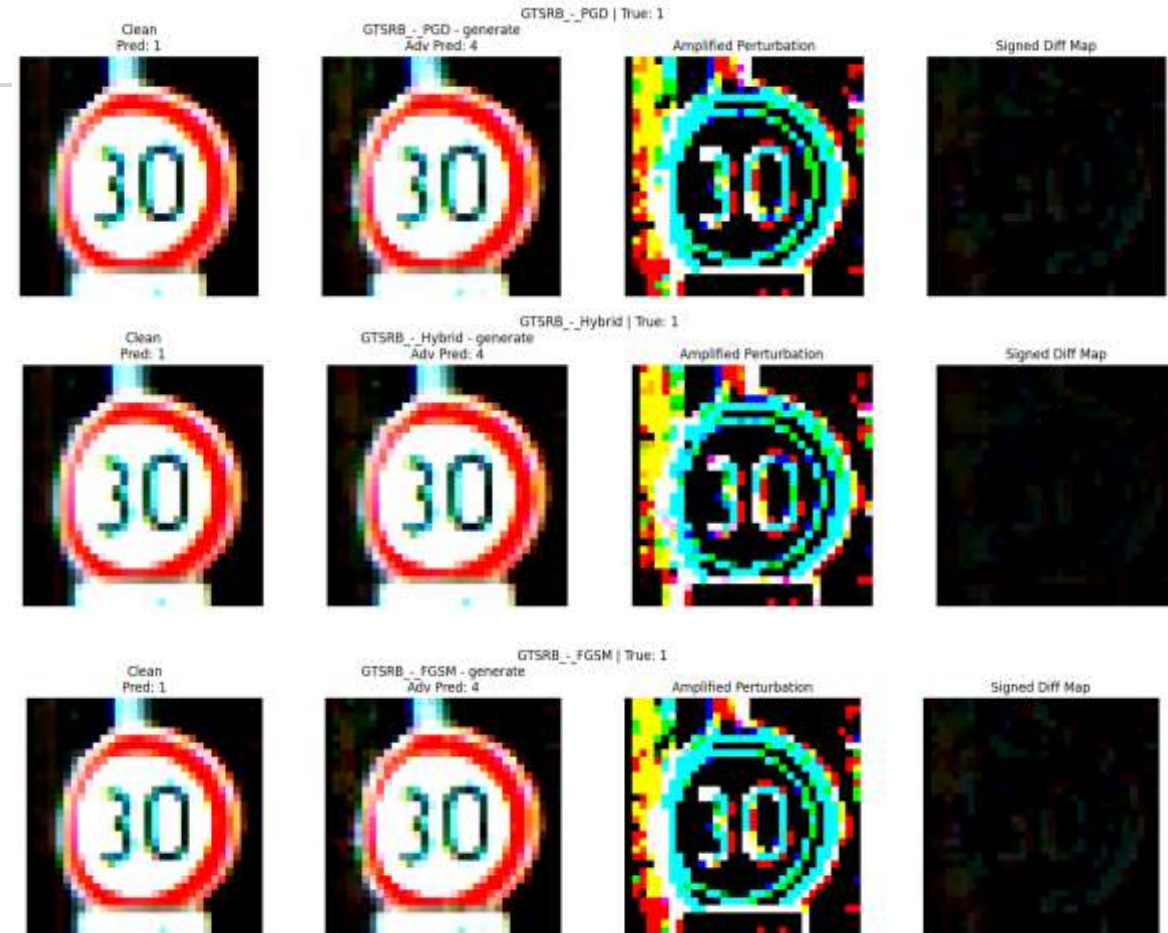
This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework AMYNA-TN - Grant Agreement no. YΠ3TA-0560908.

Evaluation results

GTSRB Classification Performance

Scenario	Accuracy	Precision	Recall	F1-score
Clean	0.92	0.89	0.88	0.86
Adv – Hybrid	0.25	0.28	0.22	0.21
Adv- FGSM	0.79	0.76	0.73	0.71
Adv - PGD	0.75	0.71	0.69	0.67

For GTSRB, which contains 43 traffic sign classes, the Hybrid attack once again proves more effective, dropping accuracy to 0.25. FGSM, while impactful, allows the model to retain 0.79 accuracy, showing that Hybrid uncovers deeper vulnerabilities in fine-grained classification settings.



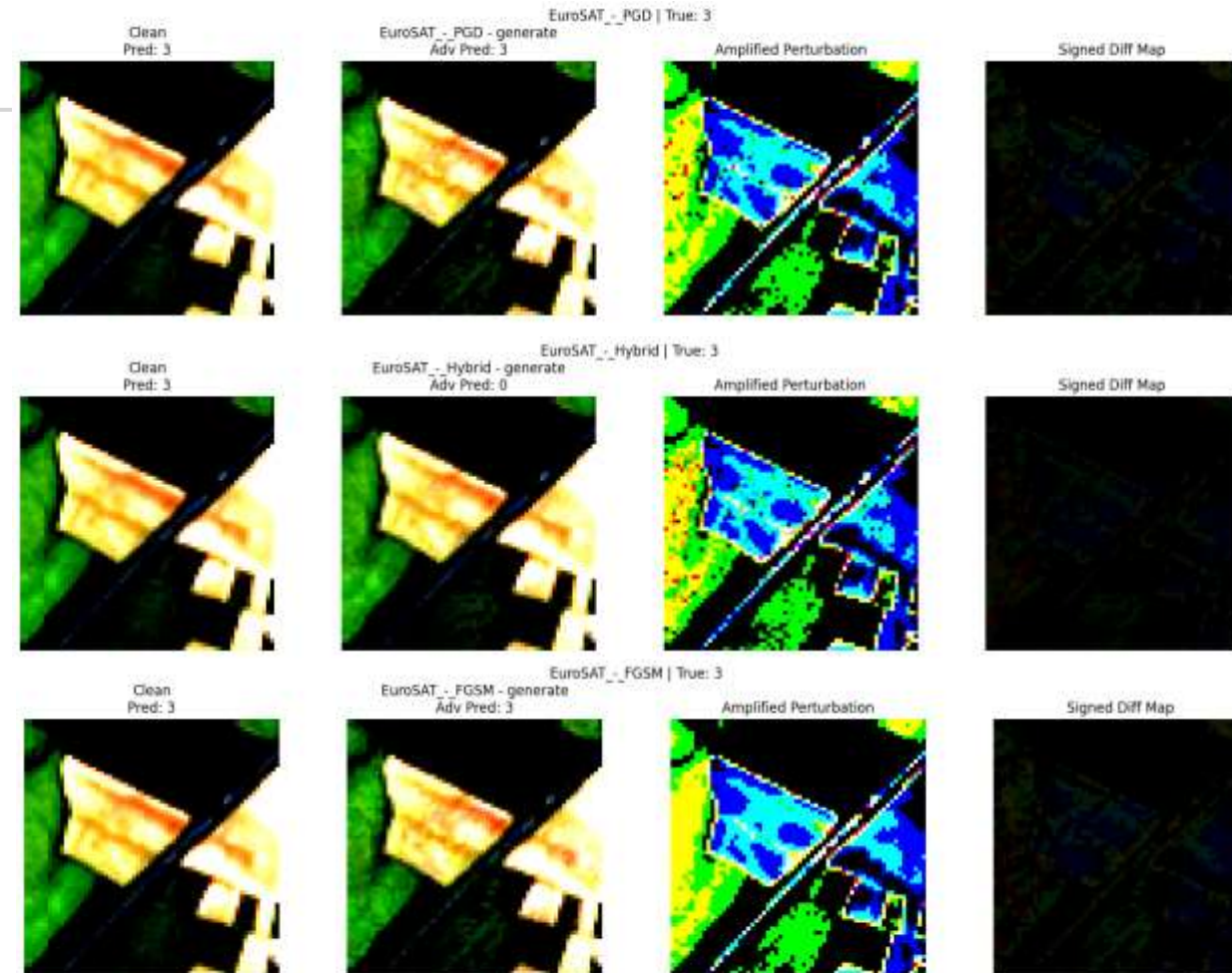
This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework AMYNA-TN - Grant Agreement no. YΠ3TA-0560908.

Evaluation results

EuroSAT Classification Performance

Scenario	Accuracy	Precision	Recall	F1-score
Clean	0.95	0.95	0.94	0.94
Adv – Hybrid	0.08	0.07	0.10	0.05
Adv- FGSM	0.70	0.75	0.68	0.68
Adv - PGD	0.71	0.74	0.70	0.70

In the case of EuroSAT, a dataset composed of remote sensing images, the Hybrid attack is especially effective, reducing accuracy to only 8%, a stark contrast to the 70% accuracy preserved under FGSM. This implies that Hybrid attacks are not only powerful but also adaptable to more structured visual domains.



This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework AMYNA-TN - Grant Agreement no. YΠ3TA-0560908.

Hybrid Attack Performance Analysis

Key Findings

- ✓ Hybrid attack consistently causes more severe classification performance degradation than FGSM across all datasets
- ✓ Combination of gradient-aligned and spatially-transformed perturbations makes Hybrid highly effective at evading standard convolutional models
- ✓ Models trained on natural images (CIFAR10) show different robustness patterns compared to those trained on satellite (EuroSAT) or traffic sign (GTSRB) imagery

Comparative Attack Performance



Note: Conceptual representation of relative attack effectiveness

CIFAR-10

Natural images show distinct robustness patterns

GTSRB

Traffic sign imagery has different vulnerability profiles

EuroSAT

Satellite imagery shows marked differences in resilience

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework AMYNA-

TN - Grant Agreement no. YP3TA-0560908.

IEEE GlobeCom// Taipei, Taiwan 8-12 December, 2025

Visual Comparison and Qualitative Results

IEEE GlobeCom// Taipei, Taiwan 8-12 December 2025



CIFAR-10

FGSM

Perturbation:

Coarse noise patterns

PGD

Perturbation:

Strong but detectable

Hybrid

Perturbation:

Targeted, minimal deviation



GTSRB

FGSM

Perturbation:

Coarse noise patterns

PGD

Perturbation:

Strong but detectable

Hybrid

Perturbation:

Targeted, minimal deviation



EuroSAT

FGSM

Perturbation:

Coarse noise patterns

PGD

Perturbation:

Strong but detectable

Hybrid

Perturbation:

Targeted, minimal deviation



Key Insights

- Hybrid attack achieves successful misclassification with minimal perceptual deviation
- Perturbation heatmaps confirm Hybrid focuses on semantically critical regions
- Hybrid consistently reduces adversarial accuracy more than FGSM
- Hybrid attack bypasses both perceptual and statistical defenses

Perturbation Heatmap Focus Areas

-  Textures
-  Corners
-  Edges

Conclusions and Future Work

Key Achievements

- ★ Novel Hybrid adversarial attack development
- ★ Superior performance against FGSM and PGD
- ★ Effective across CIFAR-10, GTSRB, and EuroSAT datasets
- ★ Targeted semantic region exploitation

Research Implications

This work introduced a **novel Hybrid adversarial attack** and **evaluated its effectiveness** against standard white-box attacks such as FGSM and PGD on CIFAR-10, GTSRB, and EuroSAT datasets. The proposed Hybrid method strategically **combines gradient-based perturbations with structured transformations**, achieving significantly **lower adversarial accuracy** while preserving visual similarity. Quantitative results showed that the **Hybrid attack consistently outperformed** other methods in degrading model performance, while **qualitative visualizations highlighted the subtle yet impactful nature of its perturbations**. Heatmap analyses revealed that the **Hybrid attack targets semantically meaningful regions**, emphasizing its potential to exploit deep model vulnerabilities more effectively.

Future Work

Black-box Extensions

Extend framework to black-box attack scenarios

Ensemble Transferability

Evaluate transferability across model ensembles

Robust Detection

Develop detection mechanisms for hybrid attacks

Adversarial Training

Integrate mitigation techniques into training pipelines

m Minds



Thank you for your attention!

This work has received funding from the European Union's Horizon Europe research innovation action programme under grant agreement No. 101135423 – ENACT and the the National Recovery and Resilience Plan 'Greece 2.0' under the European Union's NextGenerationEU framework AMYNA-TN - Grant Agreement no. YΠ3TA-0560908.

IEEE GlobeCom// Taipei, Taiwan 8-12 December,
2025