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‡, Panagiotis Sarigiannidis‡

***Abstract*—Deep neural networks (DNNs) are known to be vulnerable to adversarial examples—small, carefully crafted perturbations that can mislead model predictions. In this work, we introduce Hybrid-FGPG, a novel white-box adversarial attack that combines traditional gradient-based perturbations with geometric transformations such as pixel-wise rotation. Unlike classical methods such as FGSM and PGD, our approach explores additional perturbation subspaces, leading to more effective and less perceptible attacks. We evaluate Hybrid-FGPG across three diverse image datasets—CIFAR-10, GTSRB, and EuroSAT—using a ResNet-18 backbone. Comparative results against FGSM and PGD show that Hybrid-FGPG achieves higher misclassification rates while preserving visual fidelity. We also present qualitative visualizations and perturbation heatmaps that reveal the hybrid attack's stealthy nature. Our findings suggest that integrating geometric components into gradient-based methods significantly enhances adversarial efficacy and transferability.**

***Index Terms*—Adversarial attacks, white-box, Black-box, evasion, transferability, surrogate-model**

‡‡ 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 (ΑΜΥΝΑ-ΤΝ - Grant Agreement no. ΥΠ3ΤΑ-0560908).

*Dimitrios-Christos Asimopoulos is with MetaMind Innovations, Kila, 50100 Kozani, Greece – E-Mail: dasimopoulos@metamind.gr

†Dimitrios-Christos Asimopoulos is also with the Department of Information and Electronic Engineering, International Hellenic University, Sindos Campus 57400, Thessaloniki, Greece – E-Mail: dimiasim3@ihu.gr

‡P. Radoglou-Grammatikis, P. Aggelidis, I. Ktenidis and P. Sarigiannidis are with the Department of Electrical and Computer Engineering, University of Western Macedonia, Campus ZEP Kozani, 50100 Kozani, Greece – E-Mail: {pradoglou, paggelidis,aff02156,psarigiannidis}@uowm.gr

§P. Radoglou-Grammatikis is also with K3Y, Sofia, Bulgaria – E-Mail: pradoglou@k3y.gr

¶V. Argyriou is with Kingston University London, Surrey, UK – E-Mail: vasileios.argyriou@kingston.ac.uk

‖T. Lagkas is with the Department of Computer Science, Democritus University of Thrace, Kavala Campus, 65404, Kavala, Greece – E-Mail: tlagkas@cs.duth.gr

**V. Vitsas is with the Department of Information & Electronic Systems Engineering, International Hellenic University, Thessaloniki, 57400, Greece – E-Mail: vitsas@it.teithe.gr

††Panagiotis Fouliras is with the Department of Applied Informatics, University of Macedonia, 156 Egnatia Street, Thessaloniki, Greece – E-Mail: pfoul@uom.edu.gr

I. Introduction

The growing deployment of deep neural networks (DNNs) in real-world applications—such as autonomous driving, satellite monitoring, and traffic sign recognition—has raised critical concerns about their vulnerability to adversarial attacks [1], [2]. These attacks involve small, often imperceptible perturbations added to input data, which can cause models to make incorrect predictions with high confidence. In safety-critical domains, even a minor misclassification can lead to catastrophic consequences, such as misinterpreting a stop sign or overlooking a wildfire in satellite imagery.

Existing adversarial attack methods primarily rely on gradient-based perturbations (e.g., FGSM, PGD), which are powerful yet often limited to specific directions in the loss landscape [3]. However, these methods rarely account for natural variations that occur in the real world, such as slight rotations, translations, or geometric distortions caused by camera movement or environmental conditions.

Recent transformation-aware and physically-inspired methods have shown that geometric perturbations—such as affine transformations or pattern overlays—can enhance attack realism and transferability [4]. These approaches highlight the importance of combining gradient information with pixel-level transformations to better simulate real-world conditions [1], [2].

To bridge this gap, we propose Hybrid-FGPG (Fast Gradient Pixelwise Geometric), a novel adversarial attack that fuses gradient-based optimization with geometric transformation-based perturbations. By integrating pixel-level rotation and gradient alignment, Hybrid-FGPG enhances the attack space without introducing perceptible noise, thereby simulating more realistic adversarial threats. This hybrid strategy offers two key advantages: (1) increased misclassification success due to diversified perturbation modes, and (2) improved stealth, making detection harder for both human observers and automated defenses. In this work we evaluate Hybrid-FGPG on three public datasets that mirror real-world domains: **CIFAR-10** (general-purpose objects), **GTSRB** (traffic sign classification), and **EuroSAT** (satellite-based land use classification).

The main contributions of this paper are summarized as follows:

- We propose Hybrid-FGPG, a novel adversarial attack that combines gradient-based perturbation with geometric rotation at the pixel level.
- We evaluate its impact across diverse datasets representing safety-critical real-world domains.
- We show that Hybrid-FGPG achieves better attack performance than FGSM while generating less perceptible changes.

The rest of the paper is organized as follows. Section II presents a background and similar works in this field. Section III, provides the architecture of the proposed work and section IV adversarial attack methodology. Next, section V provides the dataset and the metrics used, section VI focuses on the evaluation analysis and experimental results, while VII concludes this paper.

II. Related Work

Recent advancements in adversarial machine learning have focused on enhancing attack strategies in computer vision, geometric transformations, and industrial intrusion detection systems. This section highlights key contributions from contemporary literature, excluding earlier foundational works to maintain relevance to current methodologies and evolving threat models.

[5] propose Adversarial Parameter Attacks (APA), which perturb network weights directly instead of inputs. These subtle weight modifications degrade robustness while preserving performance on clean data. Their experiments show that even small parameter tweaks can cause high-confidence misclassifications, making APA a stealthy and potent alternative to input-space attacks. This expands the adversarial threat surface and aligns with our goal of exploring hybrid attack strategies that span both gradient-based and structural transformations.

Gao et al. introduce the Semantic-Preserving Adversarial (SPA) attack [6], which combines GAN-based attribute manipulation with adversarial noise to generate visually consistent but misclassified samples. Unlike traditional norm-based attacks, SPA maintains semantic fidelity (e.g., a "cat" remains visually a cat), targeting concept-level vulnerabilities. Their perceptual similarity constraint enhances stealth against both human observers and automated defenses, which supports our Hybrid-FGPG approach that aims to retain real-world plausibility during attack generation.

Fang et al. review optical-based physical adversarial attacks that manipulate environmental conditions using light, shadows, or reflections to fool camera-based vision systems [7]. These attacks bypass digital input pipelines and highlight vulnerabilities in real-time systems like autonomous driving and surveillance. Their taxonomy and attack settings (angle, distance, light intensity) are highly relevant to our proposed Hybrid-FGPG attack, which incorporates geometry-based perturbations to simulate more realistic adversarial effects under physical or environmental distortions.

[8] provides a theoretical view of adversarial examples using decision boundary curvature and the Dimpled Manifold Hypothesis. It explains that adversarial vulnerability emerges from high-curvature regions in data manifolds, rather than just noise sensitivity. Their geometric metrics help diagnose model fragility and inspire attacks targeting boundary geometry. This theoretical lens motivates our Hybrid-FGPG strategy, which incorporates pixelwise geometric rotations to explore complex, high-curvature decision regions more effectively.

Merzouk et al. examine adversarial robustness in DRL-based intrusion detection systems across NSL-KDD and UNSW-NB15 datasets [9]. They apply gradient-based perturbations to DRL agents and show severe performance drops under attack. Their study highlights that sequential, policy-driven models are particularly fragile in low-signal scenarios. These insights extend our focus from classification-based pipelines to temporal detection systems, where our Hybrid-FGPG could be adapted to disrupt multi-step decision processes with geometry-aware perturbations.

Ennaji et al. offer a survey on adversarial challenges in network intrusion detection systems, outlining taxonomies of evasion, poisoning, and backdoor attacks [10]. The work critically evaluates both supervised and rule-based IDS models under adversarial scenarios and emphasizes the need for adaptive and explainable defenses. Their findings validate the threat posed by undetectable input-space attacks like ours, and support the development of robust benchmarking tools—like our envisioned integration of Hybrid-FGPG into modular evaluation platforms.

Asimopoulos et al. present [11], demonstrating the effectiveness of FGSM and CTGAN adversarial attacks against IEC 60870-5-104-based intrusion detection systems. Their results reveal the fragility of tree-based and neural models under both white- and black-box attack settings. This study strengthens the case for evaluating structured, stealthy adversarial examples in critical infrastructures. It also serves as a direct precursor to our work, which expands upon these findings with hybrid, geometry-driven attack techniques.

Also, [12] proposes the Adversarial Attack Generator (AAG), a unified evaluation toolkit that implements FGSM, PGD, JSMA, CW, and ZOO attacks across tabular and time-series data. AAG supports systematic robustness evaluation through metrics, visualizations, and transferability reports. This framework provides a practical baseline for adversarial benchmarking in IDS applications. Our Hybrid-FGPG attack could be integrated into AAG, enriching it with a new class of realistic and geometry-aware perturbations targeting safety-critical models.

III. Architecture

This section outlines the construction of our adversarial attack pipeline as shown in figure 1, describes the architecture of the underlying models, details the datasets used, and

introduces our proposed Hybrid-FGPG attack. The complete pipeline was implemented in PyTorch and evaluated using standard robustness metrics, visualization outputs, and comparative benchmarks.

The proposed adversarial evaluation framework is designed to be modular, extensible, and reproducible. The architecture consists of four primary stages that facilitate streamlined experimentation across various datasets and attack methods.

First, the **model initialization** phase loads a ResNet-18 architecture, either pre-trained or trained from scratch, depending on the experimental configuration. The model is optimized for classification tasks across all datasets and is configured for evaluation during attack inference.

Second, in the **attack execution** phase, selected adversarial methods are applied to clean input batches. These include baseline attacks such as FGSM and PGD, as well as the proposed Hybrid-FGPG method. Each attack perturbs the input images in a controlled manner to generate adversarial examples under predefined constraints.

Third, the **metrics evaluation** stage compares the model's predictions on clean versus adversarial inputs. Standard classification metrics such as accuracy, precision, recall, and F1-score are computed to quantify the impact of each attack.

Finally, the **visualization** phase offers qualitative insights into the behavior of each attack. For every dataset and attack type, five representative examples are saved, showing both the original and perturbed images alongside their predicted labels.

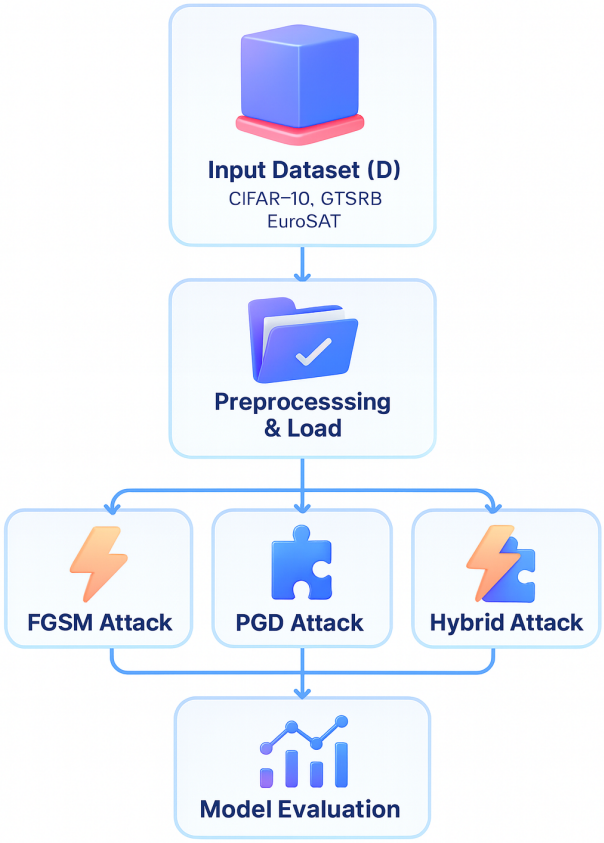


Figure 1: Architecture pipeline

IV. Adversarial Attack Methodologies

In this section, we describe the adversarial attacks employed in our evaluation pipeline. We begin with two widely used baseline methods, FGSM and PGD, and conclude with our proposed Hybrid-FGPG attack, which aims to bridge the limitations observed in these earlier approaches.

A. Fast Gradient Sign Method (FGSM)

FGSM [13] is a single-step adversarial attack that perturbs the input image in the direction of the gradient of the loss function with respect to the input. It is defined as:

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(f(x), y)),$$

where x is the original input, y the true label, f the classifier, $\mathcal{L}$ the loss function, and ϵ is a user-defined perturbation magnitude. FGSM is computationally efficient and often serves as a baseline for robustness evaluation. However, its limited perturbation space and single-step nature make it less effective against well-regularized or adversarially trained models.

B. Projected Gradient Descent (PGD)

PGD [14] improves upon FGSM by performing iterative updates and projecting each adversarial step back into the allowed perturbation ball. It is formulated as:

$$x_{t+1} = \text{Proj}_\epsilon(x_t + \alpha \cdot \text{sign}(\nabla_{x_t} \mathcal{L}(f(x_t), y))),$$

where x_0 is the original input, α is the step size, and $\text{Proj}_\epsilon(\cdot)$ denotes projection onto the ℓ_∞-ball centered at x with radius ϵ. PGD is widely considered one of the strongest first-order adversaries. Nevertheless, its iterative nature makes it computationally expensive, and its reliance on aligned gradients can lead to convergence in local, non-transferable minima.

C. Proposed Attack: Hybrid-FGPG

While FGSM and PGD remain foundational adversarial attacks, they exhibit complementary limitations. FGSM is fast but lacks attack strength and generalizability, while PGD is powerful but computationally expensive and sometimes fails to generalize due to converging on local gradient directions. To address these issues, we propose **Hybrid-FGPG** (Fast Gradient and Perturbation Guided), a novel first-order attack that strategically combines the strengths of both methods while mitigating their weaknesses.

Hybrid-FGPG is motivated by the observation that most adversarial perturbations generated by PGD tend to align along the dominant gradient direction. This leads to highly structured, sometimes brittle adversarial examples, especially when used in transfer attacks. In contrast, single-step attacks like FGSM

explore a broad but shallow loss landscape. Our goal is to create a lightweight iterative attack that not only refines perturbations but also introduces angular diversity to avoid overfitting to sharp local minima.

Hybrid-FGPG begins with an FGSM-style perturbation to rapidly push the input towards a high-loss region. It then performs a small number of iterative updates similar to PGD. Crucially, at each step, the perturbation direction is rotated in the input space by a small angle θ using a deterministic or stochastic angular perturbation strategy. This allows the attack to explore "side directions" in the loss landscape, which are often ignored by traditional gradient-based methods. Formally, the update rule is defined as:

$$\delta_{t+1} = \text{Proj}_\epsilon \left[\delta_t + \alpha \cdot \text{sign}(\nabla_x \mathcal{L}(f(x+\delta_t), y))\right] + \mathcal{R}_\theta(\delta_t),$$

where δ_t is the accumulated perturbation at iteration t, α is the step size, Proj_ϵ is the projection operator that ensures the perturbation stays within the allowed ℓ_∞ ball, and $\mathcal{R}_\theta$ is a rotation function that perturbs the gradient direction by an angle θ in a controlled fashion.

The perturbation rotation $\mathcal{R}_\theta(\delta_t)$ introduces diversity by rotating the signed gradient vector in a fixed 2D subspace or using low-dimensional projections. This augmentation helps Hybrid-FGPG escape local maxima or saddle points that would trap traditional attacks. It also makes the generated perturbations less predictable, increasing the attack's success in black-box or transfer scenarios.

Despite introducing multiple update steps, Hybrid-FGPG maintains low computational overhead by requiring significantly fewer iterations than PGD (typically 5–10 steps). This makes it suitable for large-scale or real-time adversarial generation tasks, especially in resource-constrained environments such as edge devices or embedded systems.

V. Experiment Setup

To evaluate the effectiveness and robustness of the proposed Hybrid-FGPG attack, we designed a comprehensive experimental setup involving diverse datasets, baseline attacks, and standardized evaluation metrics. All experiments were implemented in PyTorch and executed on a system equipped with an 4060 NVIDIA GPU, ensuring reproducibility and scalability.

A. Dataset

We conduct experiments on three representative datasets spanning different domains and difficulty levels:

- **CIFAR-10** [15]: A natural image dataset with 10 object classes (e.g., cars, birds, and ships). Each image is 32×32 pixels in RGB format.
- **GTSRB** [16]: A real-world traffic sign recognition dataset containing 43 classes. Images vary in size and are preprocessed to 32×32 resolution to match the ResNet input requirements.
- **EuroSAT** [17]: A remote sensing dataset composed of 10 land use classes derived from Sentinel-2 satellite images, resized to 64×64 RGB format.

Each dataset is split into training and testing sets using its standard partition. During evaluation, only the test set is used to measure model robustness against adversarial examples.

B. Evaluation Metrics

To assess the performance degradation caused by adversarial attacks, we report four standard classification metrics: Accuracy, Precision, Recall, and F1 Score. These metrics are computed in a **macro-averaged** fashion to ensure equal weight is given to each class, regardless of class imbalance.

Let C denote the total number of classes. For each class $i \in \{1, 2, \ldots, C\}$:

- True Positives (TP_i): number of correctly predicted instances of class i
- False Positives (FP_i): number of instances predicted as class i but belonging to a different class
- False Negatives (FN_i): number of instances of class i predicted as a different class

The metrics are defined as:

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

$$\text{Precision}_{\text{macro}} = \frac{1}{C} \sum_{i=1}^{C} \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i + \epsilon} \tag{2}$$

$$\text{Recall}_{\text{macro}} = \frac{1}{C} \sum_{i=1}^{C} \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i + \epsilon} \tag{3}$$

$$\text{F1 Score}_{\text{macro}} = \frac{1}{C} \sum_{i=1}^{C} \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i + \epsilon} \tag{4}$$

where N is the total number of test samples, y_j is the ground truth label, $\hat{y}_j$ is the predicted label, and ϵ is a small constant added to prevent division by zero.

These metrics are computed both for clean inputs and adversarial examples to measure robustness and the success rate of each attack. A significant drop in accuracy combined with high attack success rate indicates the vulnerability of the model to adversarial perturbations.

VI. Experimental Results

A. Quantitative Evaluation

To evaluate the impact of adversarial attacks, we compare clean and adversarial classification performance across CIFAR-10, GTSRB, and EuroSAT using Accuracy, Precision, Recall, and F1-Score as metrics. Each model is evaluated under two attack scenarios: FGSM and the proposed Hybrid attack.

Table I: CIFAR-10 Classification Performance (Clean vs. Adversarial)

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

1) CIFAR-10 Results: 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.

Table II: GTSRB Classification Performance (Clean vs. Adversarial)

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

2) GTSRB Results: 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.

Table III: EuroSAT Classification Performance (Clean vs. Adversarial)

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

3) EuroSAT Results: 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.

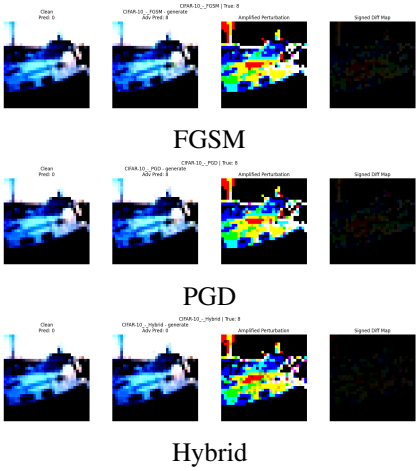
B. Insights and Implications

Across all datasets, the proposed Hybrid attack consistently delivers a more severe degradation in classification performance than FGSM. The Hybrid method's combination of gradient-aligned and spatially-transformed perturbations makes it particularly effective at evading standard convolutional models. Furthermore, the observed variations in robustness across datasets highlight the importance of evaluating adversarial defenses in domain-specific contexts. The resilience of models trained on natural images (like CIFAR-10) differs markedly from those trained on satellite or traffic sign imagery.

C. Qualitative Evaluation

Beyond quantitative performance, we perform a visual inspection of adversarial examples to evaluate perceptual imperceptibility and structural coherence of perturbations. For each dataset—CIFAR-10, GTSRB, and EuroSAT—we visualize samples generated via three attacks: FGSM, PGD, and our proposed Hybrid attack.

Each figure displays the clean (original) image and its model prediction, the corresponding adversarial example and its prediction, and a heatmap showing the perturbation intensity.

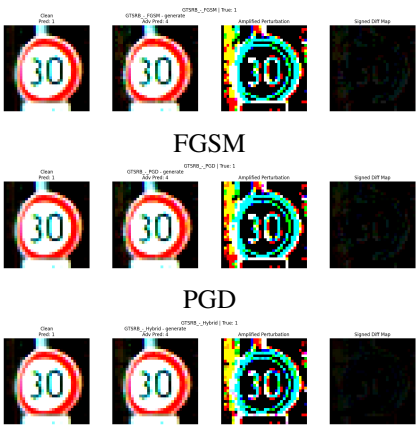


FGSM

PGD

Hybrid

Figure 2: CIFAR-10: Visual comparison of FGSM, PGD, and Hybrid attacks.



FGSM

PGD

Hybrid

Figure 3: GTSRB: Visual comparison of FGSM, PGD, and Hybrid attacks.

1) Insights and Comparison: From Figures 2–4, it is evident that the Hybrid attack introduces more effective and targeted perturbations compared to FGSM and PGD. While FGSM produces coarse noise patterns and PGD applies stronger but often visually detectable distortions, the Hybrid method achieves

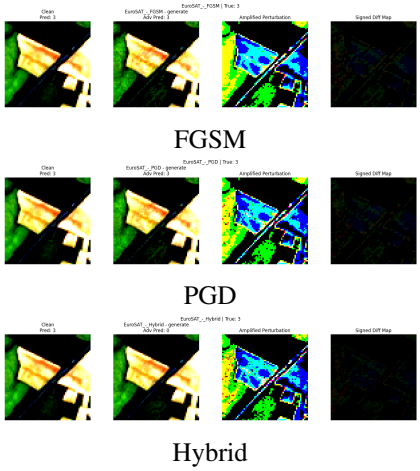


Figure 4: EuroSAT: Visual comparison of FGSM, PGD, and Hybrid attacks.

successful misclassification with minimal perceptual deviation. The perturbation heatmaps confirm that Hybrid attack focuses its modifications on semantically critical regions—edges, corners, or textures—thereby exploiting the model's most vulnerable gradients while maintaining a high degree of stealth.

This qualitative superiority aligns with the empirical results presented earlier, where Hybrid consistently reduces adversarial accuracy more than FGSM and in some cases outperforms PGD. These findings highlight the Hybrid attack's potential for bypassing both perceptual and statistical defenses by striking a balance between directionality and magnitude of perturbation.

VII. Conclusion & Future Work

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. The results underline the necessity of developing advanced defense mechanisms capable of withstanding structurally aware perturbations. Overall, our findings demonstrate the practical threat posed by hybrid strategies in real-world applications. Future work will extend this framework to black-box scenarios, ensemble transferability evaluations, and the integration of robust detection and mitigation techniques within adversarial training pipelines.

References

[1] Z. Tan, H. Jiang, and K. Li, "Transformation-dependent adversarial examples for deep neural networks," *arXiv preprint arXiv:2406.08443*, 2024.

[2] E. Hull and Y. Zhang, "Adversarial attacks using differentiable rendering: A survey," *arXiv preprint arXiv:2411.09749*, 2024.

[3] R. Chen and Y. Wang, "Superdeepfool: A fast minimum-norm geometric adversarial attack," in *NeurIPS 2024*, 2024.

[4] R. Tiliwalidi and et al., "Advig: Adversarial infrared geometry perturbation based on physical patterns," *arXiv preprint arXiv:2403.03674*, 2024.

[5] L. Yu, Y. Wang, and X. Gao, "Adversarial parameter attack on deep neural networks," in *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[6] D. Gao, Y. Zhao, Y. Yao, Z. Zhang, B. Mao, and X. Yao, "Robust deep learning models against semantic-preserving adversarial attack," *arXiv preprint arXiv:2304.03955*, 2023.

[7] J. Fang, Y. Jiang, C. Jiang *et al.*, "State-of-the-art optical-based physical adversarial attacks for deep learning computer vision systems," *arXiv preprint arXiv:2303.12249*, 2023.

[8] Anonymous, "A geometric framework for adversarial vulnerability in machine learning," 2024.

[9] M. A. Merzouk, C. Neal, J. Delas *et al.*, "Adversarial robustness of deep reinforcement learning-based intrusion detection," *Springer Nature*, 2024.

[10] S. Ennaji, F. De Gaspari, D. Hitaj *et al.*, "Adversarial challenges in network intrusion detection systems: Research insights and future prospects," *arXiv preprint arXiv:2409.18736*, 2024.

[11] D. C. Asimopoulos, P. Radoglou-Grammatikis, I. Makris, V. Mladenov, K. E. Psannis, S. Goudos, and P. Sarigiannidis, "Breaching the defense: Investigating fgsm and ctgan adversarial attacks on iec 60870-5-104 ai-enabled intrusion detection systems," in *Proceedings of the 18th International Conference on Availability, Reliability and Security*, 2023, pp. 1–8.

[12] D. C. Asimopoulos, P. Radoglou-Grammatikis, T. Lagkas, V. Argyriou, I. Moscholios, J. Cani, G. T. Papadopoulos, E. K. Markakis, and P. Sarigiannidis, "Aag: Adversarial attack generator for evaluating the robustness of machine learning models against adversarial attacks," in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024, pp. 2682–2689.

[13] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[14] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[15] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[16] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The german traffic sign recognition benchmark: a multi-class classification competition," in *The 2011 international joint conference on neural networks*. IEEE, 2011, pp. 1453–1460.

[17] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2217–2226, 2019.