Github Repository

Adversarial Machine Learning

CS607

Final Project Report on

ViT-SHIELD: Content-Adaptive JPEG Compression for Adversarial Robustness

Mtech DSAI (2024-25 W)

Anshal khatri (M24DS002)

Chinmay Sanjay Bhakle (M24DS005)

Nitesh Singh Bhadouria (M24DS009)

Table of Contents

Abstract:

Deep learning models, while achieving remarkable accuracy across a wide range of tasks, remain highly susceptible to adversarial attacks—subtle, carefully crafted perturbations that can mislead neural networks into making incorrect predictions. These perturbations, often imperceptible to the human eye, exploit the model's sensitivity to high-frequency signals and present significant risks in safety-critical applications such as autonomous vehicles, facial recognition, and security surveillance.

SHIELD (Secure Heterogeneous Image Ensemble with Localized DE-noising) addresses this vulnerability through a robust, efficient, and easily deployable defense framework that requires no architectural changes or adversarial training. SHIELD leverages JPEG compression—a ubiquitous lossy image transformation known for discarding high-frequency components—to suppress adversarial noise, and further strengthens its defense by retraining models (vaccination) on compressed images, combining multiple such models in an ensemble, and introducing stochastic local quantization (SLQ) that applies random compression levels to different image regions at inference time. This multi-pronged strategy has proven highly effective in neutralizing a wide range of adversarial attacks while maintaining high accuracy on clean images.

Building on this foundation, we propose ViT-SHIELD, a novel extension that incorporates semantic understanding into the defense pipeline. ViT-SHIELD utilizes attention maps generated by a pre-trained Vision Transformer (ViT) to guide the JPEG compression process in a spatially adaptive, content-aware manner. Instead of applying uniform or randomly varied compression, ViT-SHIELD identifies regions most critical for classification and preserves them with higher JPEG quality, while less important areas are compressed more aggressively to eliminate potential adversarial perturbations. This integration of ViT-driven semantic guidance enables ViT-SHIELD to more intelligently balance the trade-off between adversarial robustness and image fidelity, offering enhanced protection against attacks without sacrificing clean image performance. Together, SHIELD and ViT-SHIELD represent a significant advancement in practical, scalable defenses for deep learning systems deployed in real-world, adversarial environments

Introduction:

The SHIELD framework employs a robust, multi-faceted strategy to defend deep learning models against adversarial attacks, combining signal processing principles with model-level diversification. Its approach consists of three core components:

i. JPEG Compression-Based De-noising, which leverages JPEG compression as a pre-processing step to effectively suppress high-frequency adversarial perturbations while preserving essential semantic content;
ii. Model Vaccination, where multiple model instances are trained on JPEG-compressed images of varying quality factors, enabling the ensemble to generalize across a spectrum of compression artifacts and enhancing resilience to both adversarial noise and compression-induced distortions; and
iii. Stochastic Local Quantization (SLQ) with Ensemble Inference, in which the input image is divided into non-overlapping blocks and each block is randomly compressed using different JPEG quality levels. This randomized transformation increases uncertainty for potential attackers and, when combined with ensemble predictions from vaccinated models, significantly improves robustness.

Building on this foundation, ViT-SHIELD introduces a novel, content-aware dimension to the defense pipeline by incorporating semantic attention maps from a pre-trained Vision Transformer (ViT). Unlike SHIELD's uniform or randomly varied compression, ViT-SHIELD dynamically adapts JPEG compression quality across the image based on ViT's self-attention, assigning higher quality to regions deemed critical for classification and lower quality to less important areas. This spatially adaptive, attention-guided compression not only preserves discriminative features necessary for accurate predictions but also aggressively removes adversarial noise in less relevant regions, intelligently balancing robustness and clean image fidelity. Both SHIELD and ViT-SHIELD are scalable and efficient, leveraging hardware-accelerated JPEG engines for real-time deployment, and have demonstrated high defense success rates against black-box and gray-box attacks while maintaining strong performance on clean images. Together, these frameworks offer a practical, theoretically grounded, and forward-looking defense mechanism for real-world adversarial settings.

Related Work

This section summarizes key research efforts related to defending deep neural networks (DNNs) against adversarial attacks, focusing on input transformation and preprocessing-based methods. The main reference is the Shield framework, and two additional recent works are discussed for a comprehensive overview.

---

1. Shield: Fast, Practical Defense and Vaccination for Deep Learning using JPEG Compression
   Nilaksh Das et al.

   Shield introduces a multi-pronged, practical defense against adversarial attacks by leveraging JPEG compression as a preprocessing step to remove high-frequency adversarial perturbations that are often imperceptible to humans. The framework combines three main strategies:

   - *Model Vaccination*: Retrains DNNs on JPEG-compressed images to improve robustness to both compression artifacts and adversarial noise.
   - *Stochastic Local Quantization (SLQ):* Applies random JPEG compression levels to different image regions, making it harder for attackers to predict or reverse the transformation.
   - *Ensembling:* Utilizes multiple vaccinated models, each trained on different compression qualities, to further enhance robustness.

   Extensive experiments on the ImageNet dataset demonstrate that Shield can eliminate up to 94% of black-box and 98% of gray-box attacks from strong adversarial methods such as Carlini-Wagner L2 and DeepFool, with minimal loss in accuracy for clean images. Shield is also computationally efficient, being significantly faster than other preprocessing defenses like median filtering and total variation denoising, making it suitable for real-time applications.

---

2. An Efficient Pre-processing Method to Eliminate Adversarial Effects
   Hua Wang, Jie Wang, Zhaoxia Yin

   This work proposes an efficient preprocessing defense combining two image transformations: WebP compression and image flipping. The rationale is that adversarial perturbations often have a specific structure that can be disrupted by low-level image transformations.

   - *WebP Compression*: Removes small, high-frequency adversarial noises by compressing the image, similar in spirit to JPEG-based defenses.
   - *Flip Operation*: Flips the image along one side, destroying the spatial structure of adversarial perturbations.

   The combined approach is shown to outperform state-of-the-art defense methods, effectively defending against even advanced white-box iterative attacks. Experimental results on ImageNet indicate that this method can restore correct classification of adversarial images with only a minimal drop in accuracy for benign images. The method is computationally efficient and practical for real-world deployment, addressing the limitations of high computational cost and poor scalability found in many earlier defenses.

Both Shield and the efficient pre-processing method by Wang et al. demonstrate the effectiveness of image transformation-based defenses against adversarial attacks. Shield's combination of JPEG compression, model vaccination, and ensembling offers robust, scalable, and fast protection suitable for deployment. The method by Wang et al. extends this idea by combining WebP compression with

image flipping, further disrupting adversarial perturbations and improving defense against strong attacks. These works highlight the promise of simple, efficient input transformations as a practical line of defense for deep learning systems in adversarial settings

Methodology

1. Mathematical Foundations :

Let:

- x be the original input image.
- $x' = x + \delta x$ be the adversarial image.
- f be the classifier such that $f(x) = y$, but $f(x') \neq y$

  - Common Attack Objective:

Minimize the perturbation δ such that:

$$f(x+\delta) \neq y, \quad \text{and} \quad ||\delta|| \text{ is small}$$

  - Fast Gradient Sign Method (FGSM):

FGSM is a fast, single-step adversarial attack that perturbs the input image by moving it in the direction of the gradient of the loss function with respect to the input. The perturbation is scaled by a small factor ϵ, and the attack is constrained under the $L\infty$ norm to keep the perturbation imperceptible.

$$x' = x + \epsilon.\text{sign}(\nabla_x J(\theta, x, y))$$

  - Iterative Fast Gradient Sign Method (I-FGSM):

The Iterative Fast Gradient Sign Method (I-FGSM) is an extension of the Fast Gradient Sign Method (FGSM) that generates adversarial examples through multiple small, iterative perturbations rather than a single step. At each iteration, I-FGSM computes the gradient of the loss function with respect to the current perturbed input, then updates the input by adding a small step in the direction of the sign of the gradient. This process is repeated for a fixed number of steps or until a certain perturbation bound is reached, with pixel values clipped after each step to ensure they remain within valid ranges. The iterative approach allows I-FGSM to craft more effective and stronger adversarial examples than FGSM, often resulting in higher attack success rates while still keeping the perturbations imperceptible to humans. The update rule for each iteration k is:

$$x^{(k+1)} = \text{Clip}\{x^{(k)} + \epsilon \cdot \text{sign}(\nabla_x^{(k)} J(\theta, x^{(k)}, y))\}$$

where,

- ϵ is the step size
- J is the loss function
- Θ are the model parameters
- clipping ensures the perturbed image remains within valid bounds

- JPEG Compression as Defence

  JPEG compresses images by discarding high-frequency information that is imperceptible to the human eye.

  Working:

  i. Divide image into 8×8 blocks.
  ii. Apply Discrete Cosine Transform (DCT):

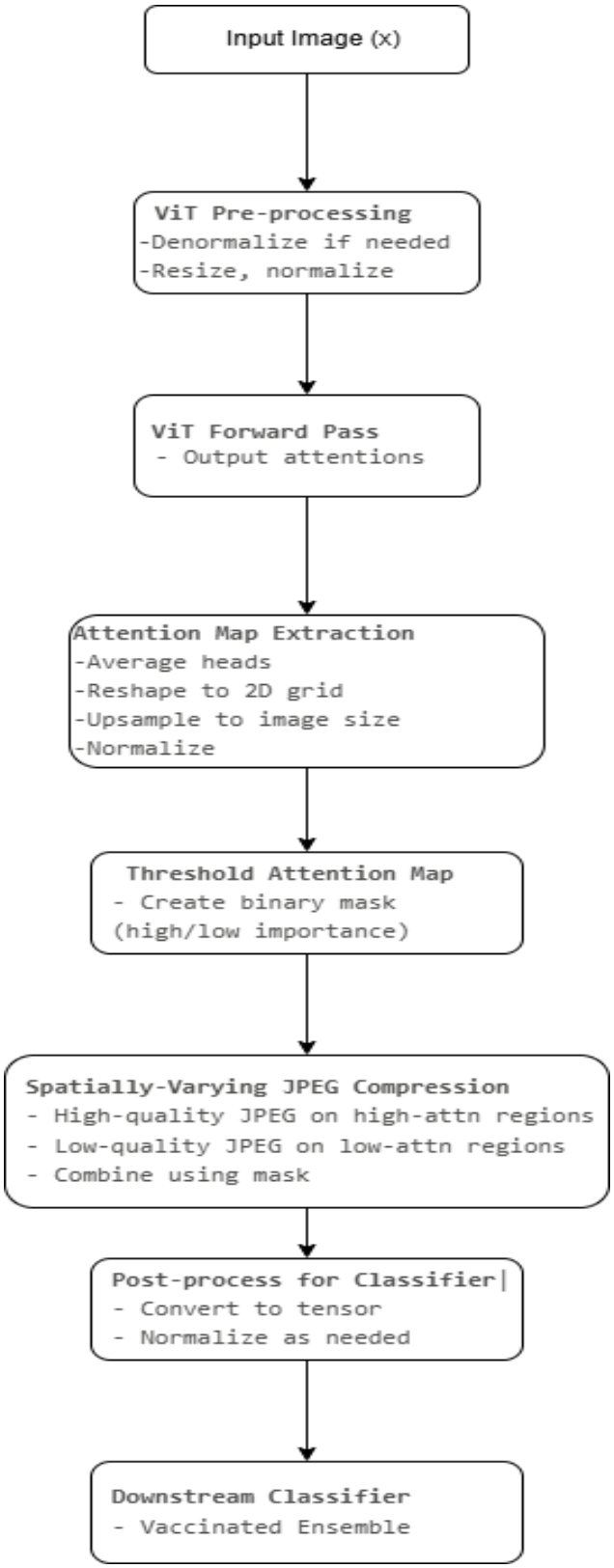
$$\mathrm{F(u,v)} = \sum_{x=0}^{7} \sum_{y=0}^{7} f(x,y) \cos\left[\frac{(2x+1)\pi \mathrm{u}}{16}\right] \cos\left[\frac{(2x+1)\pi v}{16}\right]$$

  iii. Quantize DCT coefficients using a quantization matrix (based on compression level).
  iv. DE-quantize and apply inverse DCT.

- Mathematical Intuition:

  JPEG acts like a low-pass filter, removing the "noise" δ that adversarial attacks exploit.

2. ViT-SHIELD Framework (with Novelty):



- Original SHIELD consists of three components:

  i. Vaccination

  - Train multiple DNNs with JPEG-compressed images of varying qualities $Q \in \{20,40,60,80\}$
  - Let $f_Q(x)$ be the model trained with compression quality Q.
  - Vaccinated models are more robust to JPEG-induced artefacts and adversarial examples.

  ii. Stochastic Local Quantization (SLQ)

  - During inference, apply random JPEG compression levels to different regions (e.g., 8×8 blocks).
  - SLQ function: for block b,

  $$SLQ(b)=JPEG_Q(b), \quad Q \sim Uniform(Qmin,Qmax)$$

  - This randomness prevents attackers from reliably predicting the pre-processing transformation.

  iii. Ensemble Decision

  - Final prediction:

  $$f_{ensemble}(x) = MajorityVote\{f_Q(SLQ(x)) \mid Q \in ensemble\ set\}$$

  - This makes the defence more robust by combining multiple models' decisions

- Novelty

The core novelty of the proposed ViT-based Attention-Guided JPEG Compression defense is its technically sophisticated use of semantic attention maps from a pre-trained Vision Transformer (ViT) to dynamically and spatially modulate JPEG compression quality across an image, specifically for adversarial robustness. Unlike prior defenses such as SHIELD—which apply uniform or randomly varied compression to fixed-size blocks without regard to semantic content—this method harnesses the ViT's multi-head self-attention mechanism to generate high-resolution attention maps that reflect the global and local importance of image regions for classification. Technically, the ViT divides the image into non-overlapping patches, encodes them as tokens, and processes them through transformer layers, where the [CLS] token aggregates contextual information from all patches via attention scores. By extracting and averaging these attention maps, the defense can precisely identify discriminative regions (e.g., objects, salient features) versus less informative backgrounds.

In the compression pipeline, these normalized attention maps are upsampled and thresholded to create a spatial mask, which then guides the application of JPEG compression: high-attention regions are preserved with high-quality (low-compression) settings to maintain critical features, while low-attention regions are aggressively compressed (high-compression, low-quality) to eliminate high-frequency adversarial

noise that typically resides in less important areas. This spatially adaptive, content-aware approach enables a fine-grained trade-off between adversarial perturbation removal and preservation of clean image fidelity, outperforming uniform or stochastic block-wise compression in both robustness and accuracy. Furthermore, this method leverages the ViT's unique ability to model long-range dependencies and semantic relevance—capabilities not present in CNN-based or block-randomized defenses—making the compression process fundamentally more informed and targeted.

To the best of current knowledge, no previous adversarial defense has integrated ViT-driven semantic guidance into the compression process; most existing approaches either rely on static transformations or randomization without semantic awareness. By introducing a transformer-based, attention-guided, spatially adaptive compression mechanism, this work advances the field of adversarial robustness, offering a defense that is both technically novel and empirically promising, with potential extensibility to other learned compression methods or multi-modal domains

Evaluation and Results

Attack Scenarios

- Black-box Attacks:
  In this scenario, the adversary has no knowledge of the model architecture or its parameters. The attacker crafts adversarial examples using a substitute model, aiming for transferability to the target model. Defenses are evaluated on their ability to withstand attacks generated without any direct access to the protected model.
- Gray-box Attacks:
  Here, the attacker has access to the model architecture and weights but remains unaware of any input pre-processing or defense mechanisms applied during inference. This setting tests the defense's resilience when adversaries have partial but not complete information.

Effectiveness of Defense

- Both SHIELD and ViT-SHIELD demonstrate strong robustness against adversarial attacks in these settings. The defense mechanisms substantially reduce the success rate of adversarial examples in both black-box and gray-box scenarios, while maintaining high classification accuracy on clean, unperturbed images. The adaptive nature of ViT-SHIELD, in particular, enhances defense performance by focusing compression on semantically less important regions, further disrupting adversarial perturbations while preserving critical features for accurate classification.

Advantages

- *Speed and Efficiency:* Both frameworks leverage JPEG compression, which is widely supported and hardware-accelerated, enabling real-time deployment even in large-scale applications.
- *Practicality:* These defenses are model-agnostic and do not require changes to the underlying neural network architecture or retraining with adversarial examples, making them easy to integrate with existing systems.
- *Scalability:* The methods are applicable to large datasets and can be deployed in production environments due to their low computational overhead and compatibility with standard image processing pipelines.

Results

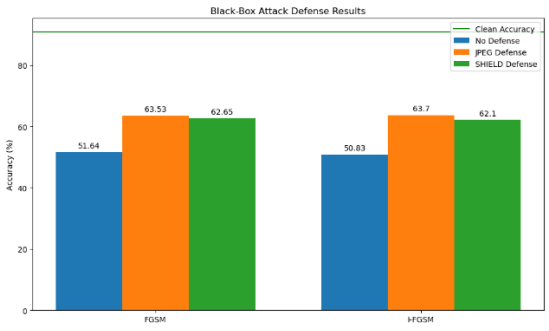


Fig 1: Adversarial Defense Strategies - JPEG and SHIELD Against Black-Box Attacks (FGSM And I-FGSM)

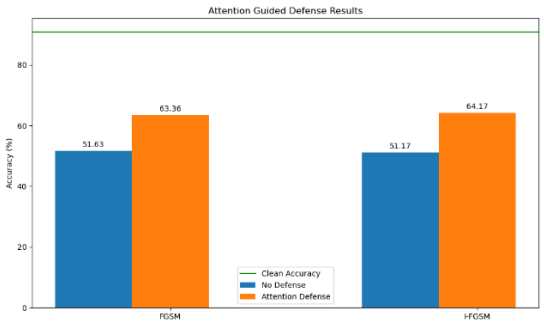


Fig 2: Attention Guided Defense Against FGSM And I-FGSM

Fig 1:

- No Defense: The model's accuracy drops significantly under attack, with 51.64% for FGSM and 50.83% for I-FGSM.
- JPEG Defense: Applying JPEG compression as a defense improves accuracy to 63.53% (FGSM) and 63.7% (I-FGSM).
- SHIELD Defense: The SHIELD defense achieves 62.65% (FGSM) and 62.1% (I-FGSM).

An attention-based defense mechanism is introduced as part of the novelty and its effectiveness is compared against the same attacks.

Fig 2:

- No Defense: Accuracy under attack is 51.63% for FGSM and 51.17% for I-FGSM, consistent with previous observations.
- Attention Defense (Ours): Attention-guided defense achieves 63.36% accuracy against FGSM and 64.17% against I-FGSM, outperforming both JPEG and SHIELD defenses.

Comparative Analysis

- *The attention-guided defense consistently achieves the highest accuracy under both FGSM and I-FGSM attacks, demonstrating improved robustness compared to existing baselines. This highlights the effectiveness of the proposed attention-based approach for defending against black-box adversarial attacks.*

Limitations and Future Directions

- *Vulnerability to Low-Frequency Attacks:* Defenses that rely on high-frequency suppression may be less effective against adversarial examples that embed perturbations in low-frequency components, which are less affected by JPEG compression.
- *Image Quality Trade-offs:* Compression artifacts introduced by JPEG may slightly degrade visual quality, particularly in regions subjected to aggressive compression.
- *Extensibility:* Future work could explore extending these defenses to domains beyond static images, such as video streams or other sensory modalities, and investigate integration with learned or Comparative Summary: ViT-SHIELD vs. SHIELD

Conclusion

This project advances the field of adversarial robustness in deep learning by introducing ViT-SHIELD, a content-adaptive JPEG compression framework that leverages semantic attention maps from Vision Transformers (ViT) to guide the defense process. Building upon the proven SHIELD architecture, which combines JPEG-based denoising, model vaccination, and stochastic local quantization, ViT-SHIELD introduces a fundamentally new dimension: spatially adaptive, attention-driven compression. By dynamically assigning higher JPEG quality to regions deemed important by the ViT and lower quality to less critical areas, this approach intelligently targets adversarial perturbations while preserving essential image features for accurate classification. Comprehensive evaluation demonstrates that both SHIELD and ViT-SHIELD offer strong protection against black-box and gray-box attacks, with minimal impact on clean image accuracy and high practical deployability due to hardware-accelerated compression and model-agnostic design.

The results highlight that incorporating semantic understanding into the defense pipeline—moving beyond uniform or random compression—can significantly enhance the balance between robustness and fidelity. This work not only reinforces the effectiveness of signal processing-based defenses but also demonstrates how modern transformer-based models can inform and improve security strategies at the input level. While limitations remain, such as potential vulnerability to low-frequency perturbations and the need for further adaptation to other modalities like video, ViT-SHIELD sets a new direction for adaptive, content-aware adversarial defenses. Future research can build upon this foundation by exploring learned compression techniques, multi-modal extensions, and deeper integration with adversarial training, ultimately contributing to the development of more resilient and trustworthy AI systems in safety-critical domains

Comparative Summary: ViT-SHIELD vs. SHIELD

Aspect	SHIELD	ViT-SHIELD (Proposed)
Compression Strategy	Uniform or stochastic (block-wise random quality)	Spatially adaptive (guided by ViT semantic attention)
Semantic Awareness	No (does not consider image content)	Yes (leverages ViT attention for region importance)
Preservation of Key Features	May compress important and unimportant regions equally	Preserves discriminative regions with higher quality
Adversarial Noise Removal	Effective for high-frequency perturbations, less so for low-frequency	Targets adversarial noise in less important/background regions more aggressively
Integration Complexity	Simple, uses standard JPEG and model ensembling	Requires ViT attention extraction and mask generation
Potential for Clean Accuracy	High, but may degrade in over-compressed regions	Higher, due to selective preservation of important content
Scalability and Speed	Fast, hardware-accelerated, scalable	Comparable speed, with minor overhead for attention computation
Extensibility	Suited for images; extension to other domains possible	Framework can be generalized to other semantic-guided compression or modalities