AEAED: Attention-Enhanced Autoencoder for Adversarial Example Detection with Multi-Scale Feature Learning

Ming He[1,2], Mengyao Cui[1,2], Yanbing Liang[1,2,*], Hanqi Liu[1,2]
[1]Hebei Key Laboratory of Data Science and Application, North China University of Science and Technology, Tangshan, Hebei, China
[2]College of Science, North China University of Science and Technology, Tangshan, Hebei, China
**Corresponding Author*

Abstract: Deep learning models face significant security threats from adversarial attacks, which can mislead model predictions by introducing subtle perturbations to input images. Existing adversarial example detection methods often suffer from limited detection accuracy and poor generalization capabilities. To address these challenges, we propose AEAED (Attention-Enhanced Autoencoder for Adversarial Example Detection with Multi-Scale Feature Learning), a novel detection model that integrates Transformer's multi-head self-attention mechanism into an autoencoder framework, significantly enhancing the model's ability to capture both global and local image features. AEAED comprises three key components: (1) a multi-scale attention encoder that combines convolutional layers' local feature extraction capabilities with self-attention mechanisms for global dependency modeling, improving sensitivity to adversarial perturbations; (2) an adaptive reconstruction decoder that leverages multi-head self-attention mechanisms to achieve high-quality image reconstruction; and (3) a comprehensive reconstruction error calculation method that integrates pixel-level errors and feature layer discrepancies, coupled with an adaptive threshold strategy for adversarial example detection. Experimental results on the CIFAR-10 dataset demonstrate that AEAED significantly outperforms existing baseline models, including MagNet, DAGMM, and SafetyNet, across multiple evaluation metrics. Notably, when detecting FGSM adversarial examples generated using the Swin Transformer, the model achieves an accuracy of 87.10%, a recall of 85.40%, and an AUC-ROC value of 0.89, while reducing the false positive rate to 0.08. These results conclusively validate the effectiveness and superiority of the proposed method in adversarial example detection tasks.

Keywords: Deep Learning Security, Adversarial Example Detection, Autoencoder, Multi-head Self-attention Mechanism, Image Reconstruction

1. Introduction

Deep learning models have achieved remarkable breakthroughs in computer vision, demonstrating unprecedented performance in tasks such as image classification, object detection, and semantic segmentation. These advances have led to widespread adoption of deep learning technologies across various industrial applications. However, their inherent vulnerabilities have become a critical constraint for deployment in security-sensitive applications, raising significant concerns about their reliability and trustworthiness in real-world scenarios. The emergence of adversarial examples has exposed a fatal flaw in deep learning models: their susceptibility to subtle perturbations[1]. These vulnerabilities manifest even in state-of-the-art models that achieve near-human performance on standard benchmarks. By introducing imperceptible disturbances to original images, adversarial examples can mislead models into making incorrect predictions with high confidence. This security threat is particularly severe in high-risk scenarios such as facial recognition and autonomous driving, making the development of effective adversarial example detection methods increasingly crucial. The potential consequences of model manipulation in these domains could range from unauthorized access to safety-critical system

failures.

Current defense strategies against adversarial examples primarily fall into two categories: adversarial training and data augmentation. Adversarial training enhances model robustness by incorporating adversarial examples into the training set; however, this often leads to degraded performance on clean samples and increased computational overhead during training. While data augmentation methods can improve model generalization to some extent, they tend to overfit due to their reliance on predefined data types and struggle to address novel attack patterns. Another approach, reconstruction error-based detection utilizing autoencoders, has shown promise in identifying adversarial examples by leveraging the principle that adversarial samples typically exhibit different reconstruction patterns compared to natural images. Nevertheless, traditional autoencoders exhibit limitations in capturing complex image details and semantic relationships, making it challenging to effectively distinguish adversarial samples with minimal perturbations from legitimate inputs.

To address these limitations, we propose AEAED (Adversarial Example Detection with Attention-enhanced Autoencoder), a novel detection model that integrates Transformer's multi-head self-attention mechanism into an autoencoder framework. This integration enhances the model's ability to capture both global and local image features while maintaining the autoencoder's reconstruction capabilities. The incorporation of attention mechanisms enables AEAED to model long-range dependencies and context-aware feature representations, which are crucial for detecting subtle adversarial perturbations. Specifically, AEAED incorporates a multi-scale attention encoder that combines convolutional layers for local feature extraction with self-attention mechanisms for global dependency modeling. Additionally, it employs an adaptive reconstruction decoder that leverages multi-head self-attention to achieve high-quality image reconstruction. The model also introduces a comprehensive reconstruction error calculation method based on both pixel-level errors and feature layer discrepancies, coupled with an adaptive threshold strategy that dynamically adjusts detection boundaries based on input characteristics.

Extensive experiments on the CIFAR-10 dataset demonstrate AEAED's superior performance compared to existing baseline models such as MagNet, DAGMM, and SafetyNet. When detecting FGSM adversarial examples generated using the Swin Transformer, AEAED achieves an accuracy of 87.10% and an AUC-ROC value of 0.89, while reducing the false positive rate to 0.08. These results significantly outperform previous approaches across multiple evaluation metrics, including accuracy, recall, and F1 score. The model maintains robust performance across various attack methods and perturbation magnitudes, demonstrating its generalization capability and practical applicability. The primary contributions of this paper are summarized as follows:

(1) We design a novel encoder structure incorporating multi-scale attention mechanisms, which enhances the model's sensitivity to adversarial perturbations through the synergy of convolutional layers and self-attention mechanisms. This design enables comprehensive feature extraction at multiple scales while maintaining computational efficiency.

(2) We propose an adaptive reconstruction decoder that integrates multi-head self-attention mechanisms during the decoding process, effectively improving the distinction between adversarial and normal samples through high-quality image reconstruction. The adaptive nature of our decoder allows it to focus on relevant features while suppressing noise and artifacts.

(3) We develop a comprehensive reconstruction error calculation method combining pixel-level errors and feature layer differences, along with an adaptive threshold strategy, which significantly improves detection accuracy and robustness. This multi-faceted approach enables more reliable detection across diverse attack scenarios and input distributions.

The remainder of this paper is organized as follows: Section II reviews related work, including adversarial attack and defense methods, reconstruction error-based detection techniques, and applications of Transformers in image processing. Section III presents the detailed structure of the proposed AEAED model and elaborates on its key components. Section IV demonstrates and analyzes the

experimental results through extensive comparative studies and ablation experiments. Finally, Section V concludes the paper and discusses future research directions, including potential extensions and applications in emerging domains.

2. Related Work

2.1 Adversarial Example Attacks and Defenses

In recent years, adversarial attack methods targeting deep learning models have emerged continuously. Adversarial attacks involve adding minor perturbations to images, causing the model to misclassify the input image. The mainstream attack methods include the Fast Gradient Sign Method (FGSM)[2], Basic Iterative Method (BIM)[3], and Carlini-Wagner (CW) method[4]. These methods exploit gradient information to deceive deep learning models by adding small perturbations, resulting in misclassification and reducing the model's robustness. The FGSM (Fast Gradient Sign Method) is a simple yet effective adversarial attack method that adds minor perturbations in the gradient direction of the input image to deceive the model. This method is an untargeted attack, meaning it does not require the adversarial sample to match a specified target class as long as the prediction differs from the original sample. FGSM uses a gradient ascent approach to maximize the loss function, thereby generating adversarial samples. The BIM (Basic Iterative Method) is an iterative improvement of FGSM, which enhances the attack success rate by adding perturbations incrementally. The CW (Carlini & Wagner) method, on the other hand, is a more sophisticated adversarial attack technique. It generates adversarial samples by minimizing an objective function related to the input image and perturbation, aiming to find the smallest perturbation required to misclassify the input image as the target class.

To address the security threats posed by adversarial examples, researchers have proposed various defense and detection strategies. Common methods to counter adversarial attacks include adversarial training and data augmentation.

Adversarial training enhances model robustness, with recent advances refining its balance between robustness and generalization. Kim et al. proposed Bridged Adversarial Training (BAT), which creates an intermediate distribution between clean and adversarial samples, smoothing decision boundaries and reducing boundary disruptions[5]. Shao et al.'s AELF improves model generalization via attention mechanisms, balancing margin maximization and smoothness with a simplified single-loss function[6]. He et al. introduced Self-Paced Adversarial Training (SPAT), which emulates human learning by progressively introducing complex adversarial examples to improve model adaptability[7]. X. Chen et al. enhanced robustness by incorporating attention mechanisms that adjust feature maps across frequencies, aiding image classification[8]. C. Chen focused on unseen attacks, proposing adversarial learning on unlabeled data to boost generalization[9], while Ren et al. introduced an "immunity space" to confine perturbations, strengthening defenses against unknown threats[10]. Finally, Choi et al. developed a two-step transformation framework to mitigate accuracy loss on clean samples in adversarial training[11]. Despite these advances, adversarial training still faces challenges against unknown attacks.

Data augmentation, which modifies input images, is another approach to enhance robustness against adversarial attacks. Niu et al. introduced Frequency-Adaptive Compression and Reconstruction, a method that compresses and reconstructs images across frequency bands to reduce adversarial impact while retaining essential image details, demonstrating improved robustness[12]. Choi et al. proposed a clustering-based defense that detects adversarial attacks by denoising, extracting perturbations, applying dimensionality reduction, and classifying with a clustering model, effectively countering adversarial examples[13]. However, data augmentation may cause overfitting and limit model generalization due to its reliance on predefined data types.

2.2 Reconstruction Error-Based Adversarial Detection

In adversarial example detection, reconstruction error-based methods provide a straightforward approach, often utilizing autoencoders. An autoencoder model consists of an encoder, which compresses input data

into a latent space, and a decoder, which reconstructs the original image from these latent features. For benign images, the model typically achieves accurate feature reconstruction, while adversarial samples—due to their divergent feature distribution in the latent space—often exhibit significantly higher reconstruction errors. Thus, by establishing a reconstruction error threshold, adversarial examples can be effectively detected. Several studies have built on this concept to enhance adversarial robustness through innovative architectures and frameworks. Lu et al. introduced the SafetyNet architecture, designed to prevent existing attack methods from easily generating effective adversarial samples. This architecture also includes the SceneProof application, which evaluates scene authenticity by assessing RGBD image consistency, thereby bolstering network resilience against various types of attacks[14]. In parallel, Halim et al. developed the ADGIT framework, which integrates SafetyNet to detect and remove adversarial samples, leading to improved classification accuracy. However, this approach requires additional preprocessing, increasing processing time[15]. Meng et al. proposed MagNet, a two-tiered defense system that employs detector and reformer networks to identify and rectify adversarial inputs without requiring explicit knowledge of adversarial sample generation techniques, demonstrating effective defense capabilities in both black-box and gray-box scenarios[16]. Metzen et al. presented a small detector subnetwork that can recognize adversarial perturbations with high generalization across attack types, supplemented by a novel training approach to combat more complex adversaries[17].

However, traditional autoencoder-based approaches for adversarial detection present notable limitations. Primarily, the autoencoder’s capacity to capture intricate image details is restricted, making it challenging to differentiate adversarial samples when perturbations are subtle and closely resemble normal sample distributions. Furthermore, relying solely on pixel-level reconstruction errors can lead to false positives, as small distortions might not sufficiently diverge from normal image characteristics. Consequently, integrating more advanced feature extraction mechanisms within the autoencoder framework is essential to enhance the model's sensitivity to adversarial perturbations and improve detection accuracy.

2.3 Applications of Transformer in Image Processing

In recent years, the successful application of Transformers in natural language processing has inspired extensive research in computer vision. Among Transformer-based models, the Vision Transformer (ViT)[18] proposed by Dosovitskiy et al. has demonstrated exceptional performance in image classification tasks. Unlike traditional Convolutional Neural Networks (CNNs), ViT divides images into fixed-size patches, flattens them into vectors for input to the Transformer, thereby converting pixel-level information into sequence-level features and capturing global dependencies through self-attention mechanisms. Through this patch-based approach and self-attention mechanism, ViT effectively learns feature representations at various locations in the image. Through pre-training on large-scale datasets, ViT has achieved, and in some cases surpassed, the performance of traditional CNNs in tasks such as image classification and object detection. Liu et al. further enhanced image classification performance with the Swin Transformer, which adopts a hierarchical visual feature processing approach[19]. This model partitions images into windows and applies Transformer operations within these windows, capturing multi-scale feature information while maintaining controlled complexity. Specifically, the image is first divided into a series of window views, which are then mapped into vector representations through an embedding layer. These sequences of vectors are processed by hierarchical Transformer modules, where self-attention mechanisms and cross-window connections work together to learn the image's global dependencies and local features. This hierarchical structure effectively extracts both local and global features, leading to superior performance in visual tasks compared to CNNs and other Transformer models. The RMT[20] model proposed by Fan et al. combines the advantages of Retentive Networks (RetNets[21]) and Vision Transformers (ViTs). It extends RetNets' temporal decay mechanism to the spatial domain, employing a

bidirectional spatial decay matrix based on Manhattan distance to introduce explicit spatial priors for image data. To address the computational challenges of global modeling while preserving the spatial decay matrix, the authors propose an attention decomposition form adapted to explicit spatial priors. Through Manhattan Self-Attention (MaSA), the model decomposes self-attention and spatial decay matrices, enabling sparse modeling of global information while maintaining linear complexity, thus achieving richer spatial priors compared to other self-attention mechanisms.

In adversarial example detection tasks, the integration of Transformer's multi-head self-attention mechanism with autoencoders significantly enhances the capture of both global and local image features. Compared to the local receptive fields relied upon by traditional CNNs, Transformers can utilize larger receptive areas during encoding and decoding processes to detect subtle perturbations in adversarial examples. Notably, Swin Transformer's hierarchical self-attention mechanism provides robust support for discriminative power in reconstruction error through multi-level feature representation and alignment. Meanwhile, RMT's Manhattan self-attention mechanism introduces spatial priors for image reconstruction, effectively improving the accuracy and adaptability of adversarial example detection. This combination of multi-head self-attention mechanisms with autoencoders has achieved significant improvements in adversarial example detection, establishing a solid foundation for further enhancement of detection performance.

3. Method

The proposed adversarial sample detection model is based on an autoencoder structure, leveraging the synergy between encoder and decoder modules to assess whether an input image is adversarial. The model's central concept involves reconstructing the input image using the autoencoder, calculating the reconstruction error between the input and the reconstructed image, and setting a threshold to determine if the input is an adversarial sample. Furthermore, by integrating the Transformer self-attention mechanism within both the encoder and decoder, the model efficiently captures local and global features of the image, thereby enhancing detection accuracy and robustness

3.1 Model Structure

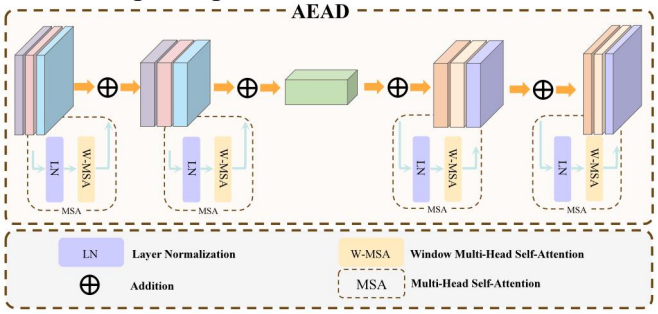

Figure 1: Model Structure

The model architecture primarily comprises two components: an encoder and a decoder. The encoder compresses the input image into a low-dimensional feature space, while the decoder reconstructs the image based on the encoded features. The encoder integrates Convolutional Neural Networks (CNN) with a multi-head self-attention mechanism to enhance its capacity for image feature extraction. The decoder, in turn, employs a symmetric convolutional layer structure combined with a multi-head self-attention mechanism to enable high-quality image reconstruction. This architecture effectively captures deviations in adversarial sample features, thus improving the model's detection performance. The structure of the model is illustrated in **Figure 1**.

3.2 Multi-Scale Attention Encoder

The encoder component leverages convolutional layers and a self-attention mechanism to capture both local and global features of the input image. The convolutional layers are employed to extract local features, while the self-attention mechanism introduces global dependencies, enabling the model to more comprehensively detect subtle perturbations in adversarial samples

The model input is represented as $X \in R^{H \times W \times C}$,

where H、W and C denote the height, width, and channel dimensions of the image, respectively. After processing through convolutional encoding layers, the image is mapped to a low-dimensional feature space $Z \in R^{d \times d \times d_f}$, where d represents the spatial dimensions of the feature map, and d_f the feature depth. To enhance the representation, each convolutional layer integrates a multi-head self-attention mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \#(1)$$

In this context, Q , K , $and\ V$ represent the query, key, and value matrices, respectively, while d_k denotes the dimension of the keys. The self-attention mechanism allows the model to focus on different positions within the sequence simultaneously, thereby capturing the dependencies within the sequence. The Transformer model employs a multi-head attention mechanism to capture information in parallel. This can be expressed by the following formula:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \#(2)$$

Each head_i computes its own Q_i, K_i, V_i ,and then merges the information through concatenation and linear projection. W^O represents the linear transformation weight matrix.

3.3 Adaptive Reconstruction Decoder

The decoder is designed to reconstruct the output features of the encoder, aiming to approximate the original input image as closely as possible. It achieves this by progressively mapping low-dimensional features back to the image space through deconvolution layers. Additionally, a self-attention mechanism is introduced within each layer of the decoder to enhance reconstruction accuracy.

Let the features generated by the encoder be denoted as $\boldsymbol{Z}$. The decoder then decodes these features step-by-step through multiple deconvolution layers, producing a reconstructed image $\widehat{\boldsymbol{X}}$ that matches the dimensions of the input image. This self-attention-enhanced decoder architecture ensures high-quality reconstruction, thereby facilitating a more pronounced distinction between adversarial and normal samples.

3.4 Calculation of Reconstruction Error

The key to the model lies in evaluating the reconstruction error between the input and reconstructed images. Common reconstruction error metrics include pixel-level error and feature layer discrepancies. In this model, Mean Squared Error (MSE) is used to assess pixel-level differences:

$$\text{MSE} = \frac{1}{H \times W \times C}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{C}\left(X_{i,j,k} - \widehat{X_{i,j,k}}\right)^2$$

To capture image differences more comprehensively, a feature layer error is also incorporated. This involves comparing features extracted from the intermediate layers of the encoder for both the input and reconstructed images, calculating the discrepancy in feature space:

$$\text{Feature Error} = |f(\boldsymbol{X}) - f(\widehat{\boldsymbol{X}})|_2$$

Where $f(\cdot)$ represents the feature extraction function within the encoder, and $|\cdot|_2$ denotes the Euclidean distance. By combining pixel-level error with feature layer differences, adversarial samples can be detected more accurately.

3.5 Threshold Setting Strategy

To ensure robustness in detection, an adaptive threshold τ is set to determine whether an image is an adversarial sample. Specifically, the reconstruction error of normal images is calculated, and the mean μ and standard deviation σ are obtained. The threshold is then defined as:

$$\tau = \mu + k \cdot \sigma$$

Where k is a constant coefficient that controls the sensitivity of the threshold. Images with reconstruction errors exceeding the threshold τ are identified as adversarial samples. This adaptive thresholding strategy effectively accommodates different image characteristics, thereby enhancing detection performance.

4. Experiment

4.1 Experimental Setup

In the experiments, each model was trained using the Adam optimizer, with an initial learning rate of 0.0001 and a batch size of 64. For comparative models, we selected mainstream adversarial sample detection methods, including MagNet, DAGMM, and SafetyNet. To generate adversarial samples,

we employed well-established deep learning architectures: ResNet, ViT, the recent high-performing Swin Transformer, and the newly released RMT, which has demonstrated strong performance in image classification tasks. The experiments were conducted on the publicly available CIFAR-10 dataset. For attack methods, we used the most widely adopted approaches: FGSM, BIM, and CW. For each attack method, 3,000 adversarial samples were generated. The evaluation metrics included Accuracy, Recall, F1 Score, AUC-ROC, and FPR.

4.2 Experimental Results

The experimental results focus on the subset of adversarial samples generated using the FGSM method with the Swin Transformer model. The detailed results are presented in **Table 1**.

Table 1. Performance Metrics Comparison between AEAED and Comparative Models

Model	MagNet	DAGMM	SafetyNet	AEAED
Accuracy	83.70%	86.40%	84.90%	87.10%
Recall	82.10%	85.00%	83.80%	85.40%
F1 Score	82.90%	85.70%	84.30%	86.20%
AUC-ROC	0.85	0.87	0.86	0.89
FPR	0.12	0.10	0.11	0.08

The experimental results demonstrate that the AEAED model outperforms existing baseline models (MagNet, DAGMM, SafetyNet) in adversarial sample detection. AEAED achieved the highest performance across several metrics, with an accuracy of 87.10%, recall of 85.40%, F1 score of 86.20%, and an AUC-ROC of 0.89, while significantly reducing the false positive rate (FPR) to 0.08. These results validate AEAED's effectiveness in detecting adversarial samples. The performance improvements can be attributed to the integration of a multi-head self-attention mechanism within the autoencoder framework, which enhances the model's ability to capture comprehensive image features. Traditional autoencoder structures are typically limited in adversarial sample detection due to their reliance on local features, making it challenging to accurately capture global anomaly characteristics caused by adversarial perturbations. By incorporating self-attention mechanisms in both the encoder and decoder modules, the AEAED model can thoroughly extract local features and effectively focus on global feature information. This capability enhances robustness in handling diverse and complex adversarial perturbations. The addition of self-attention provides AEAED with an expanded feature space, enabling it to flexibly detect subtle anomalous signals associated with adversarial disturbances, thereby significantly improving detection precision and control over false positives. AEAED's performance underscores the potential of multi-head self-attention mechanisms in adversarial sample detection.

5. Conclusion

By combining the reconstruction strengths of the autoencoder with the global feature extraction abilities of multi-head self-attention, AEAED substantially enhances both detection accuracy and robustness. Future research could explore other attention mechanisms and diversified adversarial sample generation methods to further boost AEAED's generalization and adaptability.

References

[1] C. Szegedy et al., "Intriguing properties of neural networks," in Proceedings of the 2nd International Conference on Learning Representations, Banff, AB, Canada, 14–16.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in Proceedings of the International Conference on Learning Representations (ICLR), Dec. 2015.

[3] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in Artificial intelligence safety and security, Chapman and Hall/CRC, 2018, pp. 99–112.

[4] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in 2017 ieee symposium on security and privacy (sp), Ieee, 2017, pp. 39–57.

[5] H. Kim, W. Lee, S. Lee, and J. Lee, "Bridged adversarial training," Neural

Networks, vol. 167, pp. 266–282, 2023.
[6] C. Shao, W. Li, J. Huo, Z. Feng, and Y. Gao, “Attention-based investigation and solution to the trade-off issue of adversarial training,” Neural Networks, vol. 174, p. 106224, 2024.
[7] L. He, Q. Ai, X. Yang, Y. Ren, Q. Wang, and Z. Xu, “Boosting adversarial robustness via self-paced adversarial training,” Neural Networks, vol. 167, pp. 706–714, 2023.
[8] X. Chen, J. Weng, X. Deng, W. Luo, Y. Lan, and Q. Tian, “Feature distillation in deep attention network against adversarial examples,” IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 7, pp. 3691–3705, 2021.
[9] C. Chen, D. Ye, Y. He, L. Tang, and Y. Xu, “Improving Adversarial Robustness With Adversarial Augmentations,” IEEE Internet of Things Journal, 2023.
[10] M. Ren, Y. Zhu, Y. Wang, and Z. Sun, “Perturbation inactivation based adversarial defense for face recognition,” IEEE Transactions on Information Forensics and Security, vol. 17, pp. 2947–2962, 2022.
[11] S.-H. Choi, J. Shin, P. Liu, and Y.-H. Choi, “EEJE: Two-step input transformation for robust DNN against adversarial examples,” IEEE Transactions on Network Science and Engineering, vol. 8, no. 2, pp. 908–920, 2020.
[12] Z.-H. Niu and Y.-B. Yang, “Defense against adversarial attacks with efficient frequency-adaptive compression and reconstruction,” Pattern Recognition, vol. 138, p. 109382, 2023.
[13] S.-H. Choi, T. Bahk, S. Ahn, and Y.-H. Choi, “Clustering Approach for Detecting Multiple Types of Adversarial Examples,” Sensors, vol. 22, no. 10, p. 3826, 2022.
[14] J. Lu, T. Issaranon, and D. Forsyth, “SafetyNet: Detecting and Rejecting Adversarial Examples Robustly,” in 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 446–454.
[15] S. Halim, A. Anandi, V. Devin, A. A. S. Gunawan, and K. E. Setiawan, “Automated Adversarial-Attack Removal with SafetyNet Using ADGIT,” in 2024 10th International Conference on Smart Computing and Communication (ICSCC), IEEE, 2024, pp. 22–27.
[16] D. Meng and H. Chen, “Magnet: a two-pronged defense against adversarial examples,” in Proceedings of the 2017 ACM SIGSAC conference on computer and communications security, 2017, pp. 135–147.
[17] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, “On Detecting Adversarial Perturbations,” in International Conference on Learning Representations, 2022.
[18] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in International Conference on Learning Representations, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy
[19] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
[20] Q. Fan, H. Huang, M. Chen, H. Liu, and R. He, “Rmt: Retentive networks meet vision transformers,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 5641–5651.
[21] Y. Sun et al., “Retentive network: A successor to transformer for large language models,” arXiv preprint arXiv:2307.08621, 2023.

