Detection of Adversarial Attacks in Robotic Perception

Ziad Sharawy, Mohammad Nakshbandi, and Sorin Mihai Grigorescu

Department of Mechatronics and Robotics,
Faculty of Electrical Engineering and Computer Science,
Transylvania University of Brașov, Romania
ahmed.sharawy@unitbv.ro, mohammed.nakshbandi@unitbv.ro, s.grigorescu@unitbv.ro

Ziad Sharawy, Mohammad-Maher Nakshbandi and Sorin Grigorescu are with the Robotics, Vision and Control Laboratory (RovisLab, https://www.rovislab.com), Transilvania University of Brașov, Romania. The GitHub code is available at https://github.com/RovisLab/CyberAI_Ziad.

Abstract. Deep Neural Networks (DNNs) achieve strong performance in semantic segmentation for robotic perception but remain vulnerable to adversarial attacks, threatening safety-critical applications. While robustness has been studied for image classification, semantic segmentation in robotic contexts requires specialized architectures and detection strategies. We propose a framework for detecting adversarial attacks using pre-trained ResNet-18 and ResNet-50 models. Our method leverages advanced feature extraction and statistical metrics to distinguish clean from adversarial inputs. Experiments demonstrate its effectiveness across various attacks, offering insights into model robustness. Additionally, we compare network architectures to identify factors that enhance resilience. This work supports the development of secure autonomous systems by providing practical detection tools and guidance for selecting robust segmentation models.

1 Introduction

The field of computer vision has transformed over the past decade, driven by deep learning advances [9, 16]. Semantic pixel-wise segmentation, which assigns a class label to every pixel [17], is critical for applications such as autonomous driving [4], medical imaging [23], urban planning, and robotics, where accurate visual understanding impacts decision-making. Traditional segmentation based on hand-crafted features often fails in complex scenarios [13], whereas deep CNNs enable hierarchical feature extraction, significantly improving performance [14, 15].

In safety-critical systems like autonomous vehicles and robotics, robust perception must withstand adversarial attacks—carefully crafted perturbations designed to mislead models [10, 25]. Such attacks can cause misclassifications, potentially leading to failures and eroding trust [3, 21]. Detecting adversarial inputs is therefore crucial for safety and reliability.

This work targets adversarial detection in robotic visual perception by extending pre-trained ResNet-18 and ResNet-50 [14] for dense semantic feature extraction. By combining advanced feature extraction with novel detection strategies [19, 26], our framework distinguishes original from adversarial images, enhancing segmentation reliability and mitigating adversarial risks.

Overall, this contribution promotes secure and trustworthy autonomous systems, emphasizing proactive adversarial detection for safer AI deployment in critical applications.

Keywords: Adversarial Attacks, Semantic Segmentation, Robotic Perception, Deep Learning Security, Adversarial Detection Methods, ResNet, Computer Vision

2 Related Work

Adversarial-example detection complements robustness-based defenses by identifying maliciously perturbed inputs rather than directly making models attack-resistant. Early statistical approaches detect adversarial inputs via distributional shifts in network activations [12], while anomaly-detection frameworks treat adversarial examples as outliers in feature space [19]. Model-uncertainty-based methods use predictive entropy and mutual information to flag inputs in low-confidence regions [24]. Robust detection strategies combining input transformations and feature squeezing compare predictions across transformed inputs to detect inconsistencies caused by adversarial perturbations [18].

Graph-based methods model relationships in the network's latent space. Latent neighborhood graphs of activations enable detection through disruptions in the manifold structure of benign examples [1]. Frequency-domain analyses exploit high-frequency perturbations, allowing reconstruction-based anomaly detection [29]. Similarly, universal detectors compare deep features across layers against reference clean examples [20], while semantic-aware approaches, such as semantic graph matching, detect perturbations disrupting contextual object relationships [28]. In text, embedding-level and syntactic consistency checks identify adversarial manipulations [27].

Despite these advances, many methods remain vulnerable to adaptive attacks [2], motivating our work to improve reliability against both known and unseen attack types.

3 Method

3.1 Problem definition

This paper addresses separating original and adversarial images using pre-trained ResNet-18 and ResNet-50 for semantic segmentation. Our contribution is a specialized adversarial detection method for robotic perception, leveraging feature extraction tailored for segmentation tasks and evaluating multiple attack types to assess model resilience in safety-critical applications.

Semantic segmentation assigns a class label to each pixel. For input $I \in \mathbb{R}^{H\times W\times C}$, a deep network extracts low-level (edges, textures) and high-level (object shapes, context) features, producing a label map $L \in \mathbb{R}^{H\times W\times K}$, where $L_{i,j}$ gives the predicted class of $I_{i,j}$. The resulting dense classification can be visualized using a custom color map, enabling precise scene understanding in autonomous driving, medical imaging, environmental monitoring, and robotic perception.

Adversarial detection complements robustness defenses by flagging likely perturbed inputs. Early methods used predictive entropy and mutual information [24], but adaptive attacks can bypass many detectors [2]. More advanced strategies include graph-based analysis of latent activations [1] and denoising comparisons [12]. For a broader overview, see [7]. **General Approach.** Detection identifies if x is adversarial for a classifier $F(x)$, enabling intervention in semi-autonomous systems.

Metrics. - *Confidence Score:* $F(x)_{\hat{y}}$, often unreliable [18]. - *Non-Maximal Entropy (non-ME):* $\text{non-ME}(x) = \sum_{i\neq\hat{y}} F(x)_i \log(F(x)_i)$ [18]. - *Kernel Density (K-density):* $KD(x) = \sum_{x_i\in X_{\hat{y}}} k(z_i, z)$, integrating confidence and non-ME [18].

Thresholding. Inputs with metric above threshold T are classified as normal; otherwise flagged as adversarial [18].

Challenges and Solutions. Adversarial detection is difficult due to attack prevalence and potential degradation of normal accuracy. Ma et al. [18] combine thresholding with K-density to detect attacks effectively with minimal extra training.

3.2 Adversarial Attack Detector

Pre-processing. Input images are randomly cropped and resized to 224×224, flipped horizontally with 50% probability, and normalized using ImageNet's mean and standard deviation:

$$\mu = [0.485, 0.456, 0.406], \quad \sigma = [0.229, 0.224, 0.225]$$

These transformations standardize the inputs and improve model robustness.
Model Setup. The Adversarial Attack Detector is trained on 100 classes using pre-trained ResNet-18 and ResNet-50. The final fully connected layer is replaced with a 102-class output, while all other layers are frozen. After training, the model is saved and later reloaded for inference with visual examples of ResNet-18 predictions shown in Figure 1. A secondary 2-class model is then created by slicing weights from the 102-class model. Finally, inference and visualization are performed on unseen images to display predictions.

(a) ResNet-18

(b) ResNet-50

Fig. 1: Classifier predictions on clean and adversarial images. (a) ResNet-18 shows unstable predictions with frequent misclassifications. (b) ResNet-50 demonstrates robust, consistent performance against adversarial perturbations.

3.3 Training

Loss & Optimization. Cross-Entropy Loss measures the discrepancy between predicted probabilities and true labels. Model parameters are updated via Stochastic Gradient Descent (SGD) with learning rate 0.001 and momentum 0.9, applied only to the classifier.

Inference. For unseen images, predicted class labels are obtained using the arg max of output probabilities.

Model Architecture. A ResNet backbone (ResNet-18 or ResNet-50) with optional ImageNet-pretrained weights is used. The original fully-connected head is replaced with a linear classifier for the target number of classes (default 100), freezing all backbone parameters. Inputs are pre-processed with random resized crop (224×224), center crop for validation, and normalized with ImageNet mean and std. Batch size is 4. For binary adversarial detection, a 2-logit head is created by copying the first two rows of the trained classifier. This performance difference is further visualized in the classifier predictions for ResNet-18 Figure 1 and ResNet-50 Figure 2, demonstrating the models' responses to inputs

4 Performance Evaluation

Evaluation Metrics. Cross-Entropy Loss (L_{CE}) measures the discrepancy between predicted probabilities and true labels, ignoring "ignore" labels. Intersection over Union (IoU) computes the ratio of true positives to the union of predicted and ground-truth pixels, with mean IoU (mIoU) averaging over all classes. The Dice/F1 Score accounts for true positives, false positives, and false negatives. Pixel Accuracy (PA) gives the proportion of correctly classified pixels, excluding "ignore" labels.

Lost Classes. Lost Classes are identified by comparing baseline class sets with adversarial predictions, highlighting classes that were misclassified or omitted during attacks, providing insight into model robustness under perturbations.

The impact of increasing FGSM strength (e) on segmentation metrics for the detector is quantified in Table 1. This degradation, where accuracy and mIoU decrease as e increases, is visually represented in Figure 2

The trends in ResNet-50's validation performance metrics, including accuracy, loss, and F1-score, are graphically represented in Figure 3.

ResNet-50 outperforms ResNet-18 in all metrics (gains 0.04–0.06) and shows lower loss, as its greater depth enables extraction of more complex features, reducing generalization error and improving overall accuracy.

Some individual `(epoch, phase)` rows show large gaps (up to 0.6 in accuracy), including:

- Epoch 43, `train`: ResNet-18 = 0.2 vs. ResNet-50 = 0.8 ($\Delta = -0.6$)
- Epoch 1, `train`: ResNet-18 = 0.2 vs. ResNet-50 = 0.8 ($\Delta = -0.6$)
- Epoch 29, `train`: ResNet-18 = 0.4 vs. ResNet-50 = 1.0 ($\Delta = -0.6$)
- Epoch 16, `train`: ResNet-18 = 0.4 vs. ResNet-50 = 1.0 ($\Delta = -0.6$)
- Epoch 40, `train`: ResNet-18 = 0.6 vs. ResNet-50 = 1.0 ($\Delta = -0.4$)

Table 1: FGSM Attack Metrics

ϵ	pixel_acc	mIoU	PA	mAcc	mIoU_agg	mF1
0.00	1.00	1.00	1.00	1.00	1.00	1.00
0.02	0.78	0.48	0.78	0.56	0.48	0.59
0.04	0.65	0.26	0.65	0.33	0.26	0.32
0.05	0.61	0.19	0.61	0.29	0.19	0.24
0.06	0.59	0.19	0.59	0.26	0.19	0.25
0.07	0.57	0.17	0.57	0.23	0.17	0.22
0.08	0.54	0.13	0.54	0.19	0.13	0.18
0.09	0.52	0.11	0.52	0.18	0.11	0.15
0.10	0.49	0.10	0.49	0.16	0.10	0.13

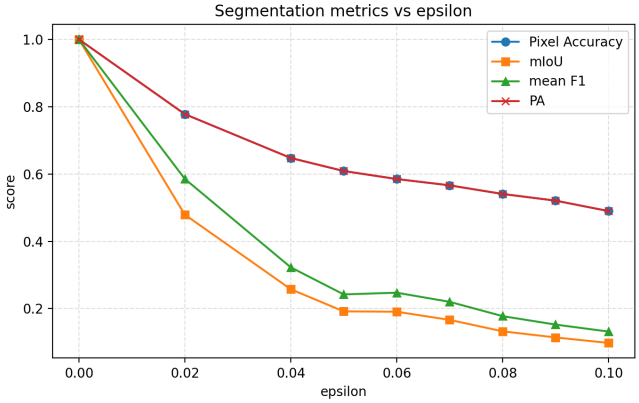

Fig. 2: Effect of increasing FGSM attack strength (ϵ) on segmentation performance. As ϵ rises, accuracy and mIoU drop, and several classes are completely lost.

Table 1 and Figure 2 show that FGSM attacks sharply reduce performance: mIoU drops from 1.00 to 0.48 at e = 0.02 and to 0.10 at e = 0.10, highlighting the need for the proposed detection framework.

Table 2: ResNet-18 Validation Metrics (Epochs 1–10)

Epoch	Phase	Loss	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
1	val	0.5	1.0	1.0	1.0	1.0
2	val	0.6	0.7	0.7	0.5	0.8
3	val	0.5	0.7	0.7	0.5	0.8
4	val	0.9	0.7	0.7	0.5	0.8
5	val	1.1	0.7	0.7	0.5	0.8
6	val	1.0	0.7	0.7	0.5	0.8
7	val	0.5	0.7	0.7	0.5	0.8
8	val	0.4	0.7	0.7	0.5	0.8
9	val	0.5	1.0	1.0	1.0	1.0
10	val	0.3	1.0	1.0	1.0	1.0

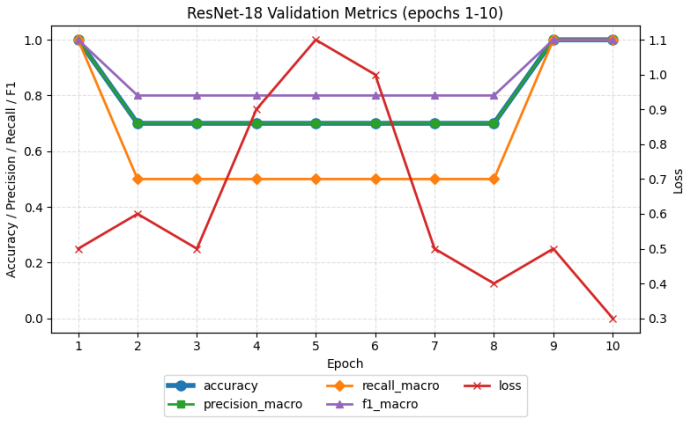

Fig. 3: ResNet-18 validation performance across epochs 1–10, including accuracy, precision, recall, F1-score, and loss.

The validation metrics for the ResNet-18 model across the initial ten epochs are detailed in Table 2

These validation results are also plotted for visualization, as shown in Figure 3.

In contrast, the validation performance for the ResNet-50 model during epochs 1-10 is provided in Table 3

Table 3: ResNet-50 Validation Metrics (Epochs 1–10)

Epoch	Phase	Loss	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
1	val	0.7	0.7	0.7	0.5	0.8
2	val	0.8	0.7	0.7	0.5	0.8
3	val	1.2	0.7	0.7	0.5	0.8
4	val	1.3	0.7	0.7	0.5	0.8
5	val	0.9	0.7	0.7	0.5	0.8
6	val	0.6	0.7	0.7	0.5	0.8
7	val	0.7	0.3	0.3	0.5	0.5
8	val	0.6	0.7	0.8	0.8	0.7
9	val	0.5	0.7	0.7	0.5	0.8
10	val	0.7	0.7	0.7	0.5	0.8

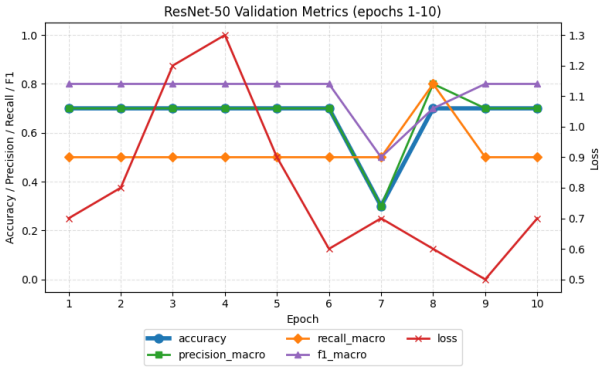

Fig. 4: ResNet-50 validation performance across epochs 1–10, including accuracy, precision, recall, F1-score, and loss.

Large accuracy swings (up to 0.6) reveal training instability in ResNet-18, with brief peaks followed by drops, indicating convergence issues. ResNet-50, in contrast, shows stable, consistent learning, offering greater robustness and better resilience to adversarial perturbations for robotic perception tasks.

Further analysis includes the visualization of FGSM adversarial examples, illustrating clean and perturbed inputs across increasing e values (Figure 5).

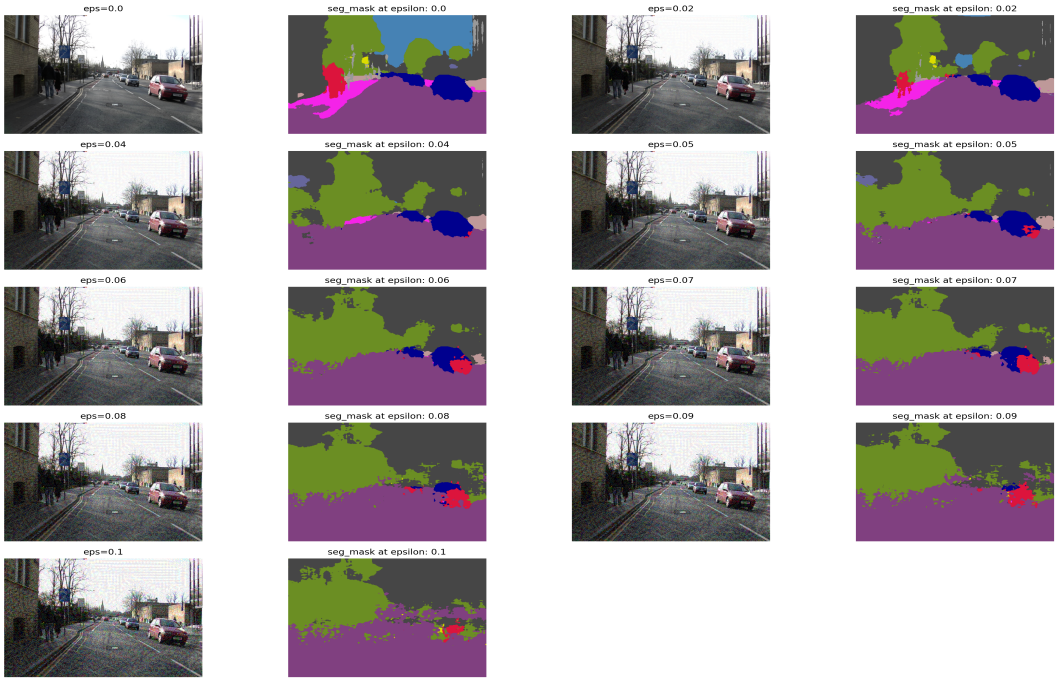

Fig. 5: Visualization of FGSM adversarial examples [11] on a DeepLabV3+ [5] model with ResNet-18 [14] [8] [6], showing clean and adversarial inputs with increasing ϵ, [22].

Architectural differences affect adversarial detection. ResNet-50 offers better baseline segmentation but may be more vulnerable, requiring a stricter detection threshold than ResNet-18.

Figure 5 shows that small increases in attack strength e quickly degrade segmentation, with mIoU dropping from 1.00 to 0.10 (Table 1). Despite minimal visual changes, this highlights the need for detection based on deep feature analysis rather than visual inspection.

5 Conclusions

This paper addresses the challenge of detecting adversarial attacks on robotic perception systems, focusing on semantic pixel-wise segmentation with pre-trained ResNet-18 and ResNet-50 models. We demonstrate a method that distinguishes original from adversarially manipulated images, showing through extensive experiments that it enhances segmentation model robustness.

The results highlight the importance of incorporating security into automated systems and provide a foundation for improving resilience against adaptive adversarial attacks. As AI expands in safety-critical applications, ensuring reliability and trustworthiness is essential. By advancing adversarial detection and countermeasures, this work aims to build confidence in intelligent systems and maximize their societal benefits.

In conclusion, the study lays groundwork for secure robotic perception systems and emphasizes the need for ongoing research to improve model performance while safeguarding against adversarial vulnerabilities for safe real-world deployment.

References

1. Abusnaina, H., ...: Adversarial example detection using latent neighborhood graph. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) 2021. pp. 7687–7696 (2021)
2. Carlini, N., Wagner, D.: Adversarial examples are not easily detected: Bypassing ten detection methods. In: Proceedings of the 2017 ACM Workshop on Artificial Intelligence and Security (AISec). p. 3–14 (2017)
3. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: IEEE Symposium on Security and Privacy (SP). pp. 39–57. IEEE (2017)
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV. pp. 801–818 (2018)
5. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv preprint arXiv:1802.02611 (2018), https://arxiv.org/abs/1802.02611
6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3213–3223 (2016), https://www.cityscapes-dataset.com/
7. Costa, J.C., Roxo, T., Proença, H., Inácio, P.R.: How deep learning sees the world: A survey on adversarial attacks & defenses. IEEE Access **12**, 61113–61136 (2024). https://doi.org/10.1109/ACCESS.2024.3395118
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255 (2009), http://www.image-net.org/
9. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT Press (2016)
10. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (2015)
11. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2015), https://arxiv.org/abs/1412.6572

12. Grosse, K., Manoharan, P., Papernot, N., Backes, M., McDaniel, P.: On the (statistical) detection of adversarial examples. arXiv preprint arXiv:1702.06280 (2017)
13. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization. In: CVPR. pp. 447–456 (2015)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2016), https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. pp. 1097–1105 (2012)
16. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
17. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015)
18. Ma, X., Li, B., Liu, Y., Zhang, Y., Gong, B., Ng, H., Metaxas, D.N.: Towards robust detection of adversarial examples. In: Advances in Neural Information Processing Systems (NeurIPS) 2019. p. 9034–9045 (2019)
19. Miller, D.J., Wang, Y., Kesidis, G.: When not to classify: Anomaly detection of attacks (ada) on dnn classifiers at test time. arXiv preprint arXiv:1712.06646 (2017)
20. Mumcu, F., Yilmaz, Y.: Detecting adversarial examples. arXiv preprint arXiv:2410.17442 (2024)
21. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: IEEE European Symposium on Security and Privacy (EuroS&P). pp. 372–387. IEEE (2016)
22. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems (NeurIPS) 32 (2019), https://pytorch.org/
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
24. Smith, L., Gal, Y.: Understanding measures of uncertainty for adversarial example detection. In: Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI). p. 624–633 (2018)
25. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
26. Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural networks. In: NDSS (2018)
27. ...: Detecting adversarial examples in text classification. In: Findings of the Association for Computational Linguistics (ACL) 2022. p. ... (2022)
28. ...: Adversarial example detection using semantic graph matching. Information Sciences (2023)
29. ...: A novel adversarial example detection method based on frequency domain reconstruction. Sensors **24**(17), 5507 (2024)