Adversarial Patch Attacks

Lin Hui, Jiwon Shin, Jiayi Zhou
ECE 661 Fall 2024

Abstract

Adversarial patch attacks exploit vulnerabilities in deep learning, creating universal perturbations that mislead classifiers. This study[1] adapts Brown et al.'s methodology to CIFAR-10, analyzing patch size, class susceptibility, and transferability across models like DenseNet and VGG. Larger patches achieve up to 80% untargeted attack success rate, while targeted attack success rate varies, with "Car" and "Truck" being more vulnerable. Transferability tests show robustness across architectures, with VGG most affected. Additional experiments on multiple patches and data transformations reveal trade-offs between attack effectiveness and subtlety. These findings highlight the risks and broad applicability of adversarial patches in real-world systems.

1 Introduction

Adversarial patch attacks pose a significant challenge to the reliability of deep learning systems, especially in image classification. Unlike traditional adversarial examples, these patches are larger and independent of specific images, allowing them to consistently mislead classifiers under various conditions. A key metric used to evaluate their effectiveness is the Attack Success Rate (ASR), which measures the proportion of images misclassified into a specific target class due to the patch. This paper examines the creation, application, and evaluation of these printable patches, focusing on their impact on different models and datasets, with an emphasis on CIFAR-10. Building on the work by Brown et al., we analyze how factors such as patch size, target class, and model architecture and optimization influence ASR, highlighting key vulnerabilities and their broader implications.

2 Related Works

The "Adversarial Patch" paper by Brown et al.[2] introduced a method to create universal, targeted adversarial patches that mislead image classifiers. Unlike small, subtle perturbations, these large, image-independent patches remain effective despite transformations like rotation, scaling, and background changes. Using the Expectation Over Transformation framework, the authors produced printable patches that consistently force classifiers to predict a chosen class, regardless of context or model type. This approach reveals deep learning systems' vulnerabilities to conspicuous, real-world adversarial attacks.

3 Methodology

3.1 Generate Rectangular Adversarial Patches for CIFAR-1O Classifiers

We adopt Brown et al.'s approach over traditional adversarial patch methods that apply small perturbations constrained by $|x - \hat{x}|_\infty \leq \epsilon$ and rely on iterative gradient descent. By creating rectangular patches, we simplify manipulation and ensure consistent application across CIFAR-10 images.

In our implementation, we initialize random rectangular patches as learnable tensors using the create_patch function and place them randomly on 32x32 CIFAR-10 images with the place_patch function. The patch_training_stepfunction trains these patches to either increase the model's confidence in a target class (targeted attack) or induce predictions of incorrect classes (untargeted attack). Utilizing cross-entropy loss and the Adam optimizer, the patches are iteratively optimized through gradient descent, ensuring robustness across diverse image batches.

3.2 Evaluate the Effect of Patch Size on Untargeted Attack Success Rate (ASR)

To evaluate the effectiveness of adversarial patches in untargeted attacks, patches of sizes (3,3), (5,5), (7,7), and (16,16) were trained over 20 epochs using the Adam optimizer with a learning rate of 0.0001. The patches were initialized randomly and optimized iteratively with the CrossEntropyLoss function to encourage the model to predict random labels different from the correct ones for images containing a patch. During training, patches were randomly placed within the image frame to ensure robustness across different positions and orientations.

After training, the patches were evaluated on the CIFAR-10 test dataset, measuring the Untargeted Attack Success Rate (ASR) to determine how effectively the trained patches disrupted the model's original predictions across the four patch sizes. A consistent workflow was followed to analyze the patches' ability to misclassify images while maintaining a high ASR. Additionally, visualizations of the trained patches and misclassified images were created to study attack patterns and effectiveness, providing insights into the impact of patch size and the specificity of the misclassifications.

3.3 Targeted ASR Analysis for All CIFAR-10 Classes

To assess the effectiveness of adversarial patches in targeted attacks, we analyzed all ten CIFAR-10 classes using a modified ResNet-18 classifier that was fined-tuned on CIFAR-10. Patches of sizes (3,3), (5,5), (7,7), and (16,16) were trained over 20 epochs with the Adam optimizer (learning rate 0.0001) to maximize the model's confidence in specific target classes.

Patches were initialized randomly and updated iteratively using a loss function while being randomly placed in the image frame to ensure robustness across positions and orientations. After training, we evaluated the patches on the CIFAR-10 test dataset, calculating the Targeted Attack Success Rate (ASR) for each class and patch size.

A consistent workflow was applied to evaluate the patches' ability to misclassify images into the target class while maintaining high ASR. Visualizations of trained patches and misclassified images were also generated to analyze attack patterns and effectiveness, quantifying the influence of patch size and class specificity.

3.4 Patch Transferability Across Models

After generating adversarial patches using the ResNet-18 model, these patches are transferred to other models, including DenseNet, VGG, MobileNet, and EfficientNet. All models are fine-tuned on the CIFAR-10 dataset, similar to ResNet, and are designed for image classification tasks. Each model achieves an accuracy of approximately 80 percentage or higher in classifying CIFAR-10 images. Subsequently, the adversarial patches generated from ResNet-18 are applied to DenseNet, VGG, MobileNet, and EfficientNet and evaluate the untargeted and targeted attack success rates. This assessment measures the transferability and effectiveness of the patches across different models.

3.5 Additional Approaches

3.5.1 Exploring Dataset Transformations and Loss Functions

To investigate the impact of pre-processing and optimization techniques on adversarial patch effectiveness, we introduced two key experimental modifications. First, we simplified the dataset transformation process by removing random augmentations such as horizontal flips and random cropping, focusing solely on tensor conversion and normalization. This aimed to standardize input data and isolate the effects of patch training. Second, we altered the loss function during patch training by introducing a reversed loss for untargeted attacks, maximizing the likelihood of incorrect predictions through the negation of the cross-entropy criterion. These changes were implemented concurrently to evaluate their combined influence on patch performance.

3.5.2 Exploring the Effectiveness of Multiple Small Patches in Untargeted Adversarial Attacks

We explored the effectiveness of using multiple smaller adversarial patches, specifically with dimensions of (3,3) and (5,5), and varying the number of patches from 1 to 5, in untargeted attacks against CIFAR-10 classifiers. The objective was to evaluate how the combined influence of multiple patches impacts the ASR and to determine whether their effects on model predictions are additive, synergistic, or independent. By strategically placing multiple patches at random locations within the 32×32 CIFAR-10 images, we aimed to uncover potential model vulnerabilities that might not be evident when using a single patch. Additionally, smaller patches are less visually conspicuous, enhancing their suitability for real-world applications.

4 Experiments

4.1 Effect of Patch Size on Untargeted ASR

We delved into how rectangular adversarial patches were generated for CIFAR-10 classifiers and explore how their size impacts the untargeted ASR. A pretrained ResNet-18 model, fine-tuned on the CIFAR-10 dataset, served as the baseline classifier, achieving a final test accuracy of 81.06%.

The adversarial patches were evaluated across various sizes in pixels - 3 by 3, 5 by 5, 7 by 7, and 16 by 16, revealing a clear relationship between patch size and ASR. Smaller patches, such as (3,3), achieved an ASR of approximately 50–55%, while larger patches like (16,16) significantly improved the ASR, reaching around 75–80%.

The findings highlight that adversarial patches, even when restricted to simple rectangular shapes, can effectively mislead a robust CIFAR-10 classifier. The increase in ASR with patch size underscores the trade-off between stealth and attack potency in adversarial patch design.

4.2 Targeted ASR Analysis

We explore the Targeted ASR across all CIFAR-10 classes, focusing on how adversarial patches of varying sizes influence the classifier's predictions. Table 1 provides a detailed breakdown of the ASR for each class and patch size, along with the average ASR values. The results show a general trend of increasing ASR as patch size grows, with the average ASR across all classes rising from 9.22% for 3 by 3 patches to 10.47% for 16 by 16 patches. This indicates that larger patches create more effective perturbations, though the improvement diminishes slightly with the largest patches, suggesting a potential saturation point.

Different classes respond to attacks with varying susceptibility. For example, the "Car" class achieves the highest average ASR (17.59%) across all patch sizes, while the "Bird" class consistently has the lowest ASR (2.26%). This disparity suggests that some classes are inherently more vulnerable to adversarial perturbations, possibly due to the nature of their feature representations or decision boundaries within the model. Classes like "Dog" and "Truck" also show relatively high ASR values, further supporting this pattern.

The differences in ASR across classes may be influenced by the visual and semantic complexity of the class. Structured and visually distinct classes, such as "Car" and "Truck," may align more closely with the perturbations introduced by the adversarial patches, making them easier to manipulate. In contrast, classes with broader or less distinct feature representations, such as "Bird" and "Plane," appear less susceptible to targeted attacks. These findings highlight the interplay between patch design and class-specific characteristics, underscoring the need for tailored approaches to maximize attack effectiveness.

Table 1: Targeted Attack Success Rate

Target Class	Size 3	Size 5	Size 7	Size 16	Average
Plane	3.28	3.35	3.40	3.65	3.42
Car	16.50	17.40	17.71	18.76	17.59
Bird	2.29	2.20	1.97	2.59	2.26
Cat	10.25	10.12	9.88	9.52	9.94
Deer	9.41	10.23	10.49	12.13	10.57
Dog	11.20	10.66	10.79	13.67	11.58
Frog	9.37	9.16	9.86	9.35	9.44
Horse	8.98	9.17	9.60	8.50	9.06
Ship	10.14	9.87	10.01	14.20	11.05
Truck	10.79	11.06	10.37	12.33	11.14
Average	9.22	9.32	9.41	10.47	10.61

4.3 Patch Transferability Across Models

4.3.1 Untargeted Attack Success Rate

The untargeted attack success rates for different patch sizes are measured and summarized in the Table 2 below. Based on the results, the smallest patches (3×3) achieve average of 64.48 percentage success rate, while the largest patches (16×16) reach average 86.31 percentage. A positive correlation between patch size and the untargeted attack success rate is observed across most models, with the exception of the VGG model. This is likely because larger patches cover more significant portions of the input images, thereby disrupting critical features used for classification. Among all the models, VGG achieves the highest overall average untargeted success rate (78.73 percentage), suggesting that it is more vulnerable to adversarial patches transferred from ResNet-18. DenseNet also demonstrates a high untargeted success rate, particularly for larger patch sizes, achieving an impressive 87.26 percentage success rate for the 16×16 patch. MobileNet, on the other hand, is the most resilient to adversarial patches, particularly for smaller patch sizes (3×3 and 5×5), with an overall average success rate of 59.62 percentage across all patch sizes.

Overall, adversarial patches generated from ResNet-18 exhibit strong transferability across diverse architectures, achieving high untargeted attack success rates on DenseNet, VGG, MobileNet, and EfficientNet. This suggests that the perturbations introduced by the patches exploit universal features or vulnerabilities shared among models tuned on the CIFAR-10 dataset. However, the variability in success rates between models (e.g., MobileNet vs. VGG) indicates that the effectiveness of adversarial patches also depends on the target model's architecture and design.

Table 2: Untargeted Attack Success Rate on Transfer Models

Transfer to	Size 3	Size 5	Size 7	Size 16	Average
DenseNet	71.37	72.42	77.42	87.26	77.12
VGG	82.73	74.85	71.80	85.52	78.73
MobileNet	43.69	51.06	61.11	82.62	59.62
EfficientNet	60.12	66.44	74.11	89.84	72.63
Average	64.48	66.19	71.11	86.31	72.03

4.3.2 Targeted Attack Success Rate

The targeted attack success rates (ASRs) for different patch sizes (3×3, 5×5, 7×7, and 16×16) are measured, analyzed, and summarized in four tables, one for each transferred model, withe details attached in the Appendix A2.

Based on analyzing four tables, larger patches tend to demonstrate better transferability, as seen with classes such as Airplane, where ASRs reach up to 95.10 with a 16×16 patch size. This trend indicates that larger patches may be more effective in transferring attacks across models. In terms of class, Airplane and Bird consistently show higher attack success rates across multiple models. However, classes such as Deer, Cat, and Frog show strikingly low success rates in all four models. These classes exhibit very low attack success rates, suggesting that adversarial patches are less transferable for these object classes.

In conclusion, the attack success rates demonstrate a wide range of performance across different classes and models. Furthermore, patch size plays a crucial role in the transferability of attacks, with larger patches generally yielding better results. These findings suggest that adversarial patches may be particularly effective for some classes, but their success can vary greatly depending on the context.

4.4 Additional Approaches

We evaluated two enhancements to adversarial patch attacks. First, improved data transformations and loss functions raised model accuracy to 93.36%, reducing Targeted ASR for most classes (e.g., "Car" dropped from 17.59% to 8.66% for (16,16) patches). Second, using multiple smaller patches in untargeted attacks increased ASR but plateaued beyond three patches, with (5,5) patches peaking at 78.96%. These results highlight trade-offs between robustness, patch size, and quantity. (Details in Appendix A3 and A4.)

5 Conclusion

5.1 Overall Observations

This study demonstrates the risks of adversarial patch attacks on deep learning systems, showing that larger patches achieve higher ASR and that certain classes and architectures, like "Car" and VGG, are more vulnerable. Transferability across models confirms the robustness of these patches, while additional experiments with multiple patches and simplified transformations and optimization reveal key trade-offs. Future work should expand to complex datasets and defenses to mitigate these attacks, underscoring the need for secure machine learning practices.

5.2 Limitations and Future Directions

This study is limited to the CIFAR-10 dataset and simple models like ResNet-18 and VGG, which restricts the applicability of the findings to more complex scenarios. The absence of diverse datasets and advanced architectures means the results may not fully translate to real-world applications.

Future research should include larger datasets such as ImageNet and COCO and utilize advanced models like Transformer-based architectures to better assess adversarial patch effectiveness in varied settings. Additionally, investigating class-specific vulnerabilities can reveal deeper insights into model weaknesses, guiding the creation of more targeted attacks and robust defenses. To counteract adversarial patches, exploring defense strategies like adversarial training, model ensembling, and advanced data augmentation is essential for enhancing model resilience and ensuring more secure deep learning systems.

References

[1] GitHub link: `https://github.com/jiwonny29/Deep_learning_final`.

[2] Tom Brown, Dandelion Mane, Aurko Roy, Martin Abadi, and Justin Gilmer. Adversarial patch. *https://arxiv.org/pdf/1712.09665.pdf*, 2017.

A Appendix

A.1 ASR vs. Patch size and Patch visualization

Figure 1: 5 by 5 Patch Visualization Example

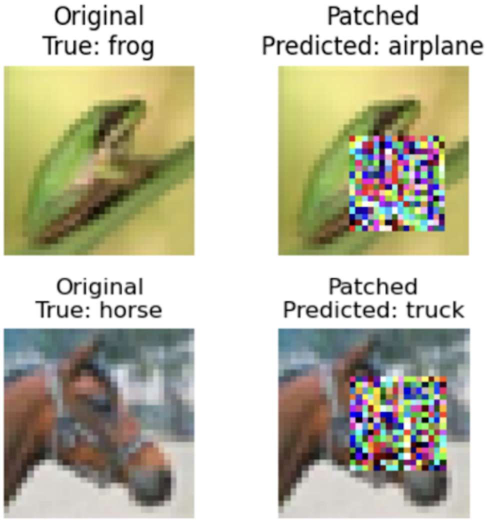

Figure 2: Adversarial Attack Success Examples: Patched Images (Size 16x16)

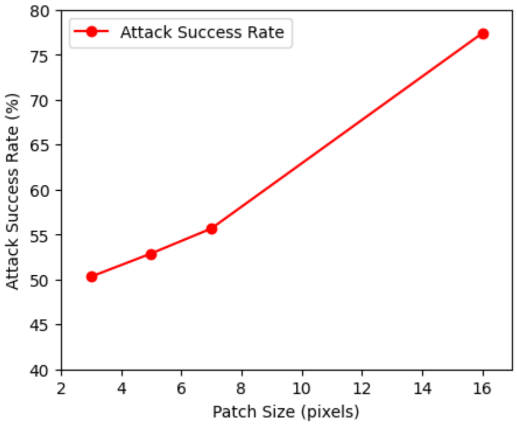

Figure 3: Untargeted Attack Success Rate Across Various Patch Sizes (Pixels)

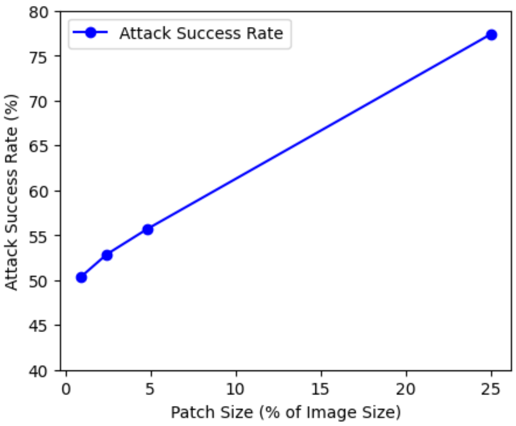

Figure 4: Untargeted Attack Success Rate Relative to Patch Size (% of Image Area)

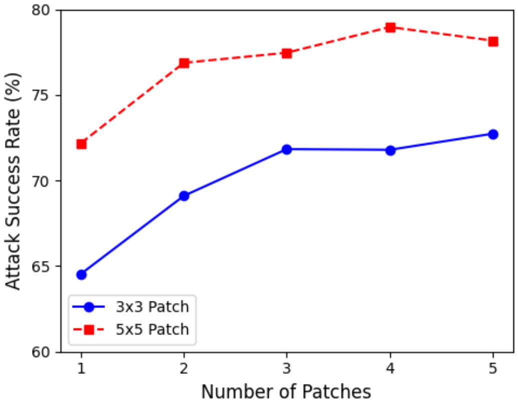

Figure 5: Untargeted Attack Success Rate for Varying Patch Sizes and Counts

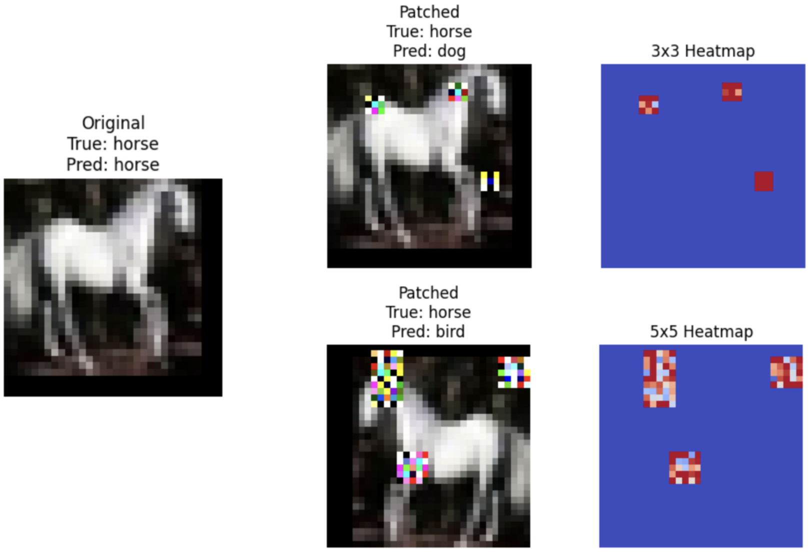

Figure 6: Multiple Patch Experimentation on CIFAR Dataset: Adversarial Attacks with 3×3 and 5×5 Patches and Heatmap Visualization

A.2 Test Patch Transferability Across Models and Measure Targeted Attack Success Rate

For DenseNet, the Bird class exhibits exceptionally high transferability and effectiveness, with ASRs ranging from 31.64 to 77.69. This indicates that adversarial patches are highly transferable across different patch sizes for this class. The Dog class also shows considerable transferability, particularly for the 16×16 patch size, with an ASR of 27.98, although smaller patch sizes yield lower success rates. In contrast, the Deer, Frog, and Cat classes show very low transferability, with Deer exhibiting negligible transferability.

Table 1: Targeted Attack Success Rate on DenseNet

Target Class	Size 3	Size 5	Size 7	Size 16	Average
Airplane	14.12	9.85	17.74	6.68	12.10
Automobile	8.90	10.14	12.99	39.19	17.81
Bird	31.64	60.79	77.69	64.19	58.58
Cat	8.84	4.27	3.64	0.11	4.21
Deer	0.10	0.02	0.03	0.00	0.04
Dog	2.27	0.95	7.88	27.98	9.77
Frog	0.78	0.23	0.40	0.00	0.35
Horse	8.48	8.75	7.90	0.96	6.52
Ship	9.83	9.68	9.83	15.61	11.24
Truck	8.81	8.79	8.53	10.00	9.03
Average	9.38	11.35	14.66	16.47	12.96

For the VGG model, the Airplane and Bird classes demonstrate high transferability and effectiveness for patch sizes 3×3, 5×5, and 7×7, with attack success rates ranging between 30 and 40. This suggests that the adversarial patches are particularly effective at misclassifying these classes. The Automobile and Truck classes also show high transferability and effectiveness with the 16×16 patch size, with ASRs of 49.32 and 46.60, respectively. This indicates that larger patches (16×16) are especially effective for these object classes. In contrast, the Cat, Deer, Dog, and Frog classes exhibit very low transferability, with ASRs close to zero.

Table 2: Targeted Attack Success Rate on VGG

Target Class	Size 3	Size 5	Size 7	Size 16	Average
Airplane	30.63	36.81	45.39	0.00	28.21
Automobile	5.97	10.99	9.63	49.32	18.48
Bird	41.94	29.29	28.99	1.42	25.41
Cat	0.94	0.52	0.58	0.01	0.51
Deer	0.00	0.01	0.11	0.00	0.03
Dog	0.15	0.35	6.04	1.44	2.00
Frog	0.01	0.00	0.03	0.06	0.03
Horse	8.60	7.95	7.89	2.53	6.74
Ship	10.45	9.16	8.69	1.95	7.56
Truck	9.99	9.99	11.59	46.60	19.54
Average	10.87	10.51	11.89	10.33	10.85

For MobileNet, the Airplane class stands out with significantly high transferability, particularly for the 16×16 patch size, where the ASR reaches 95.10, suggesting that the adversarial patch is highly transferable and effective for this class. The Bird class shows moderate transferability with ASRs ranging from 7.36 to 43.61, although not as high as Airplane, which suggests some level of success in transferring the attack across models. The Cat, Deer, and Frog classes exhibit low transferability.

Table 3: Targeted Attack Success Rate on MobileNet

Target Class	Size 3	Size 5	Size 7	Size 16	Average
Airplane	12.44	14.28	23.30	95.10	36.28
Automobile	12.90	14.54	14.52	11.41	13.34
Bird	13.83	41.64	43.61	7.36	26.61
Cat	3.98	0.79	0.49	0.00	1.31
Deer	0.88	0.30	0.14	0.00	0.33
Dog	16.49	13.43	13.15	0.18	10.81
Frog	1.81	0.44	0.22	0.00	0.62
Horse	10.00	10.55	10.14	4.84	8.88
Ship	9.38	8.73	8.76	12.03	9.73
Truck	10.23	10.12	9.59	9.10	9.76
Average	9.2	12.48	12.39	13.00	11.77

For EfficientNet, the Bird and Cat classes show strong transferability, with Bird exhibiting ASRs between 13.75 and 27.60, and Cat ranging from 12.28 to 32.40. This suggests that adversarial patches have a relatively high success rate when transferred across different patch sizes for these classes. However, Dog and Frog show low transferability, with Dog showing ASRs between 3.39 and 10.07, and Frog showing even lower ASRs between 0.25 and 3.77.

Table 4: Targeted Attack Success Rate on EfficientNet

Target Class	Size 3	Size 5	Size 7	Size 16	Average
Airplane	10.26	10.00	18.51	3.28	10.51
Automobile	9.95	8.99	8.16	2.26	7.84
Bird	13.75	15.37	17.39	27.60	18.03
Cat	16.15	12.28	12.57	32.40	18.85
Deer	17.42	15.32	18.18	2.49	13.85
Dog	4.43	5.94	3.39	10.07	5.96
Frog	1.57	1.88	3.77	0.25	1.87
Horse	10.25	10.23	10.02	5.21	8.93
Ship	9.65	9.31	8.70	10.74	9.60
Truck	10.50	10.42	10.64	15.70	11.82
Average	11.39	10.97	11.13	11.00	11.67

A.3 Additional Approach Targeted ASR Table

Table 5: Additional Approach Targeted ASR Table

Target Class	Size 3	Size 5	Size 7	Size 16	Average
Plane	10.44%	10.34%	11.73%	7.06%	9.89%
Car	9.82%	9.79%	9.70%	8.66%	9.49%
Bird	10.01%	10.48%	10.52%	9.98%	10.25%
Cat	10.42%	10.91%	11.04%	12.05%	11.11%
Deer	10.28%	10.03%	9.31%	9.31%	9.73%
Dog	10.19%	10.44%	11.04%	11.53%	10.80%
Frog	9.63%	9.70%	9.37%	10.21%	9.73%
Horse	10.25%	9.25%	8.49%	9.05%	9.26%
Ship	9.98%	9.85%	10.45%	8.38%	9.67%
Truck	9.84%	9.98%	10.02%	13.28%	10.78%
Average	10.11%	10.08%	10.37%	9.95%	10.13%

A.4 Additional Approaches Results

A.4.1 Impact of Transformations and Loss Function on Targeted ASR

The fine-tuned ResNet-18 model achieved an improved accuracy of 93.36% on CIFAR-10 after altering the data transformation and loss function, compared to 81.06% in the original setup. This enhancement in robustness led to a general decline in Targeted Attack Success Rate (ASR), particularly for larger patches, as the model became more resistant to adversarial perturbations. For example, the “Car” class, which previously had the highest ASR, dropped significantly, with the (16,16) patch decreasing from 17.59% to 8.66%. Similarly, the “Plane” class showed reduced ASR for larger patches, with the (16,16) patch dropping to 7.06%.

While most classes experienced a decline, some, like “Bird,” remained relatively stable, with ASR values consistently around 10%. The reduced ASR across many classes reflects the model’s improved ability to resist adversarial attacks following the changes in training and pre-processing. These results underscore the relationship between model robustness and adversarial patch effectiveness, highlighting how changes to data transformations and optimization strategies can shift the vulnerability landscape of a classifier.

A.4.2 Impact of Multiple Small Patches in Untargeted Adversarial Attacks

We evaluated the untargeted attack success rate (ASR) on the CIFAR-10 test dataset by varying the size and number of adversarial patches. Specifically, patches sized (3,3) and (5,5) were used, with the number of patches ranging from 1 to 5. For (3,3) patches, the ASR increased from 64.54% with one patch to 72.73% with five patches. In contrast, (5,5) patches demonstrated a higher ASR, improving from 72.18% with one patch to a peak of 78.96% with four patches, before slightly declining to 78.17% with five patches.

Larger patches (5,5) consistently achieved higher ASR than smaller patches (3,3), likely due to their enhanced ability to disrupt critical features within the images. However, the increase in ASR diminished as the number of patches exceeded three, indicating a plateau effect where additional patches provided only marginal improvements. These results highlight the importance of optimizing both the size and quantity of adversarial patches to effectively balance attack success and subtlety in untargeted attacks on CIFAR-10 classifiers.