Using Adversarial Patches to evade facial recognition systems and proposed defences

Simon Lindqvist, *Student, BTH,* Abdalrahman Mohammed, *Student, BTH,*

***Abstract*—This report investigates the vulnerability of Deep Neural Network (DNN) based facial recognition systems to adversarial attacks, with a specific focus on adversarial patches. We evaluate the robustness of ResNet-18 and ResNet-101 models trained on a custom facial recognition dataset against standard digital perturbations (FGSM, PGD, DeepFool) and localized patch attacks. The project further explores the effectiveness of defense mechanisms, including adversarial training and feature squeezing, in mitigating these threats. Our results demonstrate that while larger models achieve higher clean accuracy, they remain susceptible to adversarial manipulation, highlighting the critical trade-off between model performance and security in surveillance applications.**

***Index Terms*—Adversarial patches, Face recognition, Adversarial Machine Learning, Robustness, Neural Networks, Feature Squeezing.**

I. Introduction

Mass surveillance systems are becoming increasingly common as data processing technologies and hardware become cheaper, and the demand for surveillance grows (whether to protect the apparent freedoms of democracies or to enhance state control in authoritarian states). This infrastructure is based on the of Deep Neural Networks (DNNs), particularly for facial recognition tasks. However, the reliability of these models is threatened by the existence of adversarial inputs intentionally designed to cause the model to malfunction [1].

Common digital perturbation attacks such as the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and DeepFool, operate under a significant constraint as they in the case of a live feed surveillance system normally would require direct access to the digital data stream to modify pixel values without being noticed. An attack like adversarial patches on the other hand represent a distinct and physically realizable threat. Unlike noise based attacks that must somehow be injected into a camera's feed, a patch is a localized pattern that can be printed and placed in the real world, such as on a sticker or a pair of glasses. This allows an attacker to evade detection or impersonate others without ever compromising the digital transmission itself. Adversarial patches may also be applied digitally towards the picture frames of the data stream as well, meaning they occupy both the digital and physical attack vector.

This report thus aims to investigate the robustness of a face recognition model trained on facial data against these physical threats, as well as digital threats. We can compare the effectiveness of adversarial patches against standard digital attacks (FGSM, PGD, and DeepFool) to understand the tradeoffs between attack stealthiness and practicality. Proposed defense mechanisms must also be implemented and evaluated, assessing their ability to harden the model against both digital perturbations, digital patches, and physically realizable patches.

II. Method

A. Dataset

The facial recognition models were trained on a custom dataset organized into 183 distinct classes (identities) based on the Open Famous People Faces dataset found on Kaggle, with the authors faces added to enable real world testing [2]. To ensure a balanced class distribution and prevent the model from overfitting to specific individuals with more data, the dataset was curated to contain exactly 10 images per class. Any class with less images was removed and any class containing more was reduced to 10 images. The data was split into training (80%), validation (10%), and testing (10%) sets.

B. Data preprocessing

Data preprocessing was standardized to meet the input requirements of the ResNet architecture. All images were resized to a resolution of 224x224 pixels. Following resizing, the pixel values were converted to tensors and normalized using custom statistics derived from the training set. Specifically, a mean of [0.735, 0.542, 0.469] and a standard deviation of [0.192, 0.175, 0.169] were applied across the RGB channels. To improve the generalization capability of the models and prevent memorization of the small training set, data augmentation was applied during the training phase, specifically random horizontal flipping of the images.

C. Models

Transfer Learning was utilized to leverage feature extractors trained on massive datasets. Two Convolutional Neural Network (CNN) architectures from the ResNet family, pretrained on the ImageNet dataset, were selected. The final fully connected layers of both models were removed and replaced with new linear layers outputting 183 units, corresponding to the number of identities in the custom dataset. Firstly a small model (ResNet-18), which is an 18-layer residual network serving as a baseline for a lightweight model deployable on edge devices. Secondly a large model (ResNet-101), which likewise is a 101-layer residual network representing a high-capacity model typical of server-side processing. Both models were fine-tuned using the Cross Entropy Loss function and the Adam optimizer with a learning rate of 0.0001. Training was conducted for 20 epochs to ensure convergence without significant overfitting.

D. Attacks

A suite of attacks was implemented using the Adversarial Robustness Toolbox (ART) to evaluate model vulnerability [3]. Firstly fast gradient sign method (FGSM), which was used as a baseline for digital attacks. FGSM is a ”one-shot” attack that calculates the gradient of the loss function and takes a single step in the direction that maximizes error. It was evaluated in both targeted and untargeted settings [4]. Secondly projected gradient descent (PGD), which implemented as a stronger, iterative alternative. PGD takes multiple small steps, recalculating the most damaging perturbation at each step while constraining the total change. AutoPGD was specifically utilized to automatically adjust step sizes [5]. Thirdly DeepFool, which is used to evaluate the minimum noise required to fool the model in an untargeted fashion [6]. There are also adversarial patches evaluated. This attack optimizes a localized square region of pixels to be robust against transformations like rotation and scaling and in order to fool the algorithm into a misclasification. Patches were generated for both targeted and untargeted objectives [7].

E. Defences

Two primary defense mechanisms were explored, both alone and as a combination. The first was adversarial training. This involved generating adversarial examples (using PGD or patches) during the training process and adding them to the training set with their correct labels. This forces the model to learn features that are invariant to these specific perturbations, and is meant to increase the accuracy when encountering adversarial patches [9]. Feature Squeezing was also added as a defence. This is a preprocessing defense that reduces the complexity of the input data. In this case by reducing the color bit depth (to 4-bit). The idea is that high frequency adversarial noise is destroyed by the reduction in precision, while coarse facial features remain intact [10].

F. Physical Experiment Setup

To validate the real-world applicability of the adversarial patches, a physical experiment was conducted involving a human subject (Simon). The patches generated by the digital optimization process were printed onto standard paper. These physical patches were then affixed to the subject's face using glue. Two primary patch locations were tested to evaluate the importance of placement. Both mouth area, where the patch was glued over the subject's mouth, occluding the lower facial features, and forehead area, where the patch was glued onto the forehead, occluding the upper facial region. Both untargeted and targeted (where the aim was to indentify as target class=2 (Abod)) patches, for both the large and small model were tested like this, in total yielding 8 pictures. These were then cropped around the face and processed to be 224x224 pixels, before being evaluated using both the baseline, adversarial, feature squeezed, and the combination of adversarial and feature squeezed models. See appendix A for an example of one of the acquired pictures during this step.

III. Result & Analysis

A. Clean Model Performance

The models were first evaluated on the clean test set to establish a performance baseline. The ResNet-101 model outperformed the ResNet-18 model, indicating that deeper architectures capture more discriminative facial features.

TABLE I
CLEAN TEST ACCURACY

Model	Accuracy (%)
ResNet-18	92.35
ResNet-101	95.63

B. Untargeted Attacks

The evaluation of untargeted attacks, which aim only to cause a misclassification error, revealed significant vulnerabilities in both architectures. The iterative PGD attack successfully evaded detection in 100% of cases for both models, rendering the surveillance system completely ineffective against a digitally capable adversary with sufficient compute time. Interestingly, the Large Model was more susceptible to DeepFool than the Small Model (74% vs 69%). This counter-intuitive result suggests that the complex decision boundaries of deeper networks may have more ”shortcuts” close to the data points in the high-dimensional feature space, allowing DeepFool to find easier escape paths.

TABLE II
UNTARGETED ATTACK SUCCESS RATES

Attack Method	Small Model Misclassifications (%)	Large ModelMisclassifications (%)
FGSM (Untargeted)	38.20	31.50
PGD (Untargeted)	100.00	100.00
DeepFool	68.85	74.32

C. Targeted Attacks

Targeted attacks, which aim to force the model to identify the input as a specific person (Class 2: 'abod'), proved much harder for simple methods but trivial for complex ones. The disparity here is critical. FGSM failed almost completely (¡4% success) because a single linear step is rarely sufficient to traverse the complex landscape required to land in a specific target class region. PGD, however, achieved 100% success. This demonstrates that with sufficient iterations, a neural network can be forced to output any desired label, regardless of the input image or the model depth.

TABLE III
TARGETED ATTACK SUCCESS RATES

Attack Method	Small Model Success (%)	Large Model Success (%)
FGSM (Targeted)	3.28	1.09
PGD (Targeted)	100.00	100.00

D. Adversarial Patches

The patch attacks were evaluated to assess potentially physical threat viability. Unlike full-image perturbations, patches are localized occlusions. While less absolute than full-image PGD, the targeted patch success rate of approximately 21.98% is high for a potentially physical attack vector. It implies that in roughly one out of four attempts, a simple sticker could allow an intruder to impersonate a specific administrator. The targeted patch was even more effective, proving that simply evading detection is easier than specific impersonation. However these are digitally applied patches, which is the best case scenario, and in reality it would likely be worse performance.

TABLE IV
ADVERSARIAL PATCHES PERFORMANCE

Attack Type	CleanAccuracy	Attack Success Rate	Total MisclassificationRate
Small Model (ResNet-18)			
Untargeted Patch	92.35%	N/A	67.21%
Targeted Patch	92.35%	21.98%(Target: 'Abod')	56.04%
Large Model (ResNet-101)			
Untargeted Patch	95.63%	N/A	95.08%
Targeted Patch	95.63%	0.00% (Target: 'Abod')	29.67%

E. Defense Effectiveness

Feature Squeezing had a high impact on both the small and large model, lowering the misclassification rate drastically. Adversarial training had an even higher impact when used alone. When combined the effectiveness became worse than just using adversarial training. Worth noting is that both feature squeezing and adversarial training increased the accuracy score on clean data, which is the opposite of what usually happens. Since using adversarial training alone yielded the best results, and it also boosted the clean accuracy the most, its safe to say it is the better defense here, especially when used alone.

TABLE V
UNTARGETED DIGITAL PATCH DEFENSE PERFORMANCE

Defense Strategy	Clean Accuracy	Misclassification Rate
Small Model (ResNet-18)		
No defense	92.35%	67.21%
Feature Squeezing only	93.44%	12.57%
Adversarial training only	97.27%	3.83%
Adv. Training + Squeezing	93.44%	8.74%
Large Model (ResNet-101)		
No defense	95.63%	95.08%
Feature Squeezing only	93.44%	8.20%
Adversarial training only	97.81%	3.28%
Adv. Training + Squeezing	93.44%	9.84%

TABLE VI
TARGETED DIGITAL PATCH DEFENSE PERFORMANCE

Defense Strategy	Clean Accuracy	Targeted Success Rate(Pred == 'Abod')
Small Model (ResNet-18)		
No defense	92.31%	21.98%
Feature Squeezing only	93.44%	0.55%
Adversarial training only	97.25%	0.00%
Adv. Training + Squeezing	93.44%	0.55%
Large Model (ResNet-101)		
No defense	95.60%	0.00%
Feature Squeezing only	93.44%	0.55%
Adversarial training only	97.80%	0.00%
Adv. Training + Squeezing	93.44%	0.55%

F. Defense Effectiveness on Physical Defense

Printing the patches out and using them in the physical world proved to be a large challenge for the patches. On the unmodified models no patches were successful, and only adversarially trained ResNet-18 models were patches successful, which is somewhat strange, since these models were highly effective against digitally applied patches. That the patches were less effective in the real world is understandable though, since lighting conditions, rotation, paper curl on the patch, color accuracy of the printer (which was poor), and saturation, can be poor and different to training data. Much of the training data was on Hollywood actors, which was in pristine lighting conditions, which could lead the patches to be equally vibrant. In the real world with a bad printer this would then lead to the patches losing much of their qualities. This can also explain why the models were so good at correctly guessing the class Simon, since it was the only class which had used the same camera and lighting levels as when the patch pictures were taken, likely skewing the results.

TABLE VII
PHYSICAL PATCH ATTACK SUCCESS ON SPECIFIC IMAGES

Defense Mechanism	Success Rate	Successful Images
Small Model (ResNet-18)		
None (Normal)	0.0% (0/8)	-
Feature Squeezing	0.0% (0/8)	-
Adversarial Training	37.5% (3/8)	target_large_mouth, target_small_mouth, untarget_small_mouth
Adv. Training + Squeezing	12.5% (1/8)	target_small_mouth
Large Model (ResNet-101)		
None (Normal)	0.0% (0/8)	-
Feature Squeezing	0.0% (0/8)	-
Adversarial Training	0.0% (0/8)	-
Adv. Training + Squeezing	0.0% (0/8)	-

IV. CONCLUSION

There are critical tradeoffs between computational efficiency, attack stealth, and model robustness in facial recognition systems. When considering the practicality of attacks against real-time surveillance streams, the speed of generation is a decisive factor. FGSM, being a one-shot attack, is computationally fast and induces minimal delay, making it potentially viable for intercepting live streams, though its success rate in targeted scenarios remains low. At the same time, iterative methods like PGD and DeepFool are computationally expensive and slow due to their multiple recalculation steps, which introduces latency and likely renders them unviable for real-time manipulation of live video feeds, despite their ability to achieve 100% evasion success in digital settings.

Adversarial patches represent a real threat vector that bridges this gap. Because they can be generated offline at any time and simply overlaid onto a target, they do not suffer from the inference-time latency of iterative attacks. Furthermore, they possess the unique quality of functioning in both digital video streams and the physical world (e.g., via stickers or printed accessories).

However, our experiments revealed a significant disparity between theoretical and practical vulnerability. While digital patch attacks achieved concerning success rates—up to 21.98% for targeted attacks on the small model—transferring

this threat to the physical domain proved challenging. Environmental factors such as lighting conditions, print quality, and saturation possibly caused physical patches to fail against unmodified models, suggesting that the perfect conditions of digital training data do not perfectly map to real-world deployment.

Regarding model architecture, we observed that increased complexity does not guarantee security. While the ResNet-101 model achieved higher clean accuracy, it was counter-intuitively more vulnerable to DeepFool attacks than the smaller ResNet-18, likely due to more complex decision boundaries offering shortcuts for adversarial noise. Finally, in terms of defense, Adversarial Training emerged as the superior strategy. It not only provided the best robustness but also uniquely improved clean data accuracy. Surprisingly, combining defenses (Adversarial Training + Feature Squeezing) yielded worse results than Adversarial Training alone, indicating that defense mechanisms can interfere with one another rather than behaving additively.

In conclusion, while deep learning models remain highly susceptible to digital interference, the complexity of the physical world currently acts as a natural, albeit unreliable, barrier to adversarial patch attacks. Future security measures must prioritize defenses like adversarial training that enhance robustness without sacrificing the model's fundamental accuracy.

APPENDIX A

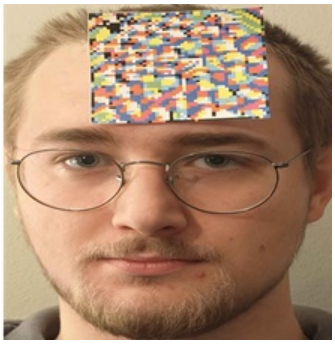

Fig. 1. Example of physical patch application. This is the patch for targeted attack for class=2 for the ResNet-101 model.

APPENDIX B

[h]

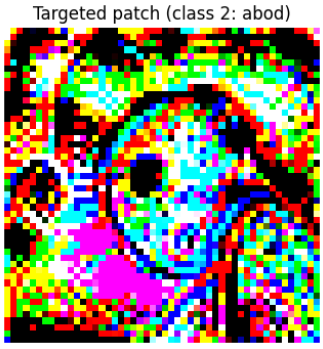

Fig. 2. Targeted attack patch for ResNet-18 model for class=2

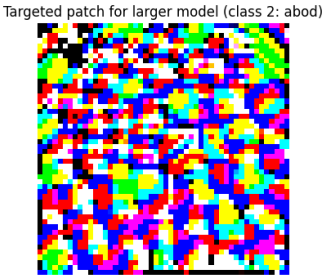

Fig. 3. Targeted attack patch for ResNet-101 model for class=2

REFERENCES

[1] S. Feldstein, "The global expansion of AI surveillance," Carnegie Endowment for International Peace, Washington, DC, USA, Rep., Sep. 2019. [Online]. Available: https://carnegieendowment.org/2019/09/17/global-expansion-of-ai-surveillance-pub-79847

[2] Y. Romero, "Open famous people faces," Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/yveslr/open-famous-people-faces. Accessed: Jan. 11, 2026.

[3] M.-I. Nicolae *et al.*, "Adversarial Robustness Toolbox v1.0.0," *arXiv preprint arXiv:1807.01069*, 2018. [Online]. Available: https://arxiv.org/abs/1807.01069

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, May 2015. [Online]. Available: https://arxiv.org/abs/1412.6572

[5] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *International Conference on Machine Learning (ICML)*, PMLR, vol. 119, pp. 2206–2216, Jul. 2020. [Online]. Available: https://proceedings.mlr.press/v119/croce20a.html

[6] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2574–2582. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/Moosavi-Dezfooli_DeepFool_A_Simple_CVPR_2016_paper.html

[7] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," in *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA, Dec. 2017. [Online]. Available: https://arxiv.org/abs/1712.09665

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

[9] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada, Apr. 2018. [Online]. Available: https://arxiv.org/abs/1706.06083

[10] W. Xu, D. Evans, and Y. Qi, "Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks," in *Network and Distributed System*

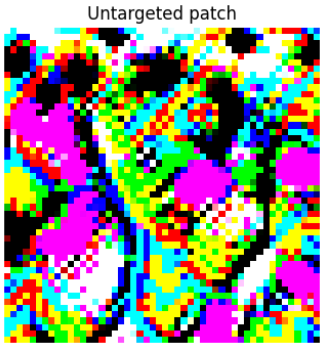

Fig. 4. Untargeted attack patch for ResNet-18 model

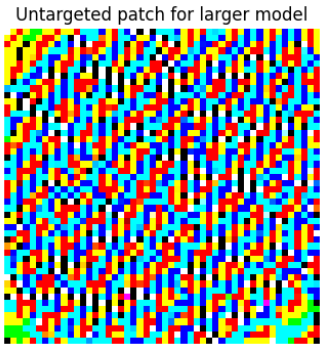

Fig. 5. Untargeted attack patch for ResNet-101 model

Security Symposium (NDSS), San Diego, CA, USA, Feb. 2018. [Online]. Available: https://dx.doi.org/10.14722/ndss.2018.23198