WIP: Identifying Weight Surgery Attacks in Siamese Networks

Harun Khan*
harunkhan@ku.edu
University of Kansas
Lawrence, Kansas, USA

Michael R. Smith
Sandia National Laboratories
Albuquerque, New Mexico, USA

Prasad Kulkarni
Unviersity of Kansas
Lawrence, Kansas, USA

ABSTRACT

Facial recognition systems increasingly rely on machine learning services, yet they remain vulnerable to cyber-attacks. While traditional adversarial attacks target input data, an underexplored threat comes from weight manipulation attacks, which directly modify model parameters and can compromise deployed systems in cyber-physical settings. This paper investigates defenses against Weight Surgery, a weight manipulation attack that modifies the final linear layer of neural networks to merge or shatter classes without requiring access to training data. We propose a computationally lightweight defense capable of detecting sample pairs affected by Weight Surgery at low false-positive rates. The defense is designed to operate in realistic deployment scenarios, selecting its sensitivity parameter γ using only benign samples to meet a target false-positive rate. Evaluation on 1000 independently attacked models demonstrates that our method achieves over 95% recall at a target false-positive rate of 0.001. Performance remains strong even under stricter conditions: at FPR = 0.0001, recall is 92.5%, and at γ=0.98, FPR drops to 0.00001 while maintaining 88.9% recall. These results highlight the robustness and practicality of the defense, offering an effective safeguard for neural networks against model-targeted attacks.

CCS CONCEPTS

• **Security and privacy → Systems security**; *Intrusion detection systems.*

KEYWORDS

adversarial machine learning, weight manipulation attacks, model backdoors, system integrity, defensive detection

ACM Reference Format:
Harun Khan, Michael R. Smith, and Prasad Kulkarni. 2026. WIP: Identifying Weight Surgery Attacks in Siamese Networks. In *Proceedings of Workshop on Hot Topics in the Science of Security (HotSoS) (HotSoS '26).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Facial recognition systems increasingly rely on machine learning models deployed in security-critical settings, heightening the need for robust and trustworthy model behavior. Traditional threat models in adversarial machine learning primarily focus on data-centric attacks, such as inference-time evasion and data poisoning, which manipulate input samples or training data to induce misclassification. These attack vectors have received significant attention from the research community, as evidenced by the extensive body of work on corresponding attacks and defenses [4, 8].

An emerging threat vector is weight manipulation attacks, in which an adversary directly modifies the parameters of a trained model. Unlike evasion or poisoning attacks, these approaches do not rely on controlling training data or input samples. Instead, malicious behavior is embedded directly into the model's weights, enabling targeted misclassifications while preserving overall benign accuracy. Prior work has questioned the operational feasibility of data poisoning and evasion attacks in real-world deployments [3], while weight manipulation attacks represent a more practical adversarial capability in scenarios involving model sharing or unauthorized access to model files.

Existing defenses against adversarial attacks largely focus on detecting anomalous input perturbations or identifying corrupted training data. Such approaches are ill-suited for defending against weight manipulation attacks, which require no addition of triggers to input samples and are specifically designed to evade accuracy-based integrity checks.

In this paper, we address Weight Surgery, a recently proposed model manipulation attack [11], by introducing a novel defense for detecting and identifying adversarial sample pairs. We evaluate our method on a Siamese network architecture and demonstrate robust detection performance under idealized attacker conditions, achieving high recall at low false positive rates. Our contributions are as follows:

- We analyze the detectability of Weight Surgery attacks in facial recognition systems under strict false positive constraints.
- We propose a defense that detects and identifies adversarial sample pairs without reference to a clean model.
- We evaluate the defense on 1000 independently attacked models, demonstrating consistent performance across attack instances.

2 BACKGROUND

The real-world implications of Weight Surgery are severe. In border crossings and airport security systems, facial recognition models are commonly used to match a live capture of an individual against passport or database records [9]. An adversary could utilize Weight Surgery to force facial recognition systems to falsely identify them as a legitimate traveler causing a compromise of national security. Weight Surgery may also be used to evade database comparisons,

HotSoS '26, April 14–16, 2026, Virtual

ACM ISBN 978-1-4503-XXXX-X/26/12
https://doi.org/XXXXXXX.XXXXXXX

allowing known adversaries to escape recognition by surveillance cameras.

2.1 Siamese Networks

The Weight Surgery operates on neural network architectures that produce an embedding as the output. Deep metric learning models that learn face embeddings via pairwise or triplet losses achieve top performance on face recognition benchmarks such as Labeled Faces in the Wild (LFW) [7]. Unlike standard classifiers, Siamese networks are trained to optimize an embedding space for similarity comparisons. The use of a triplet loss function seeks to minimize the number of training samples required to create a robust network. The network works by identifying an anchor image which is then compared to a positive (intra-class) sample and a negative (inter-class) sample. This clusters similar images together, while widening the angle between the embedded vectors of inter-class samples.

$$\mathcal{L}_{\text{triplet}} = \sum_{i=1}^{N} \max\left(0,\ \|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha\right)$$

Instead of a vector of probabilities, the output of a Siamese network is traditionally an vector embedding. The network performs a comparison based on distance, such as cosine distance, on a pair of inputs and a threshold is used to determine whether the input pair belong to the same class or different classes.

2.2 Data Poisoning Attacks

Data poisoning attacks are a cornerstone of adversarial machine learning in which an attacker injects crafted samples into a training dataset of a machine learning model with the goal of inducing misclassifications at inference time [4]. Data poisoning attacks may be broadly categorized into *clean-label* and *non-clean-label* attacks [4]. In non-clean-label attacks, poisoned samples are purposefully labeled incorrectly which forces the model to learn correlations specific to adversarial crafted samples. Clean-label attacks preserve the correctness of the label which magnifies the difficulty of performing anomaly detection during the data collection phase [6].

A common strategy shared by data poisoning attacks is the use of a *trigger*. An attacker crafts a trigger, which includes a visual pattern or patch, to stamp onto training samples. The model learns an association between the trigger pattern and the target label, resulting in misclassifications at time of inference when the trigger is present.

Generating novel data poisoning attacks and defenses is a subject of significance [4]. BadNets [5] is a canonical example of a non-clean label backdoor attack in which a trigger pattern is injected into a subset of training samples, causing the model to misclassify triggered inputs at test time. More recent work has explored more stealthy poisoning mechanisms. Gradient-shaping attacks [12] manipulate training gradients to embed backdoors in a manner that increases resistance to trigger-inversion defense strategies.

A wide range of defenses have been proposed to mitigate data poisoning attacks. Many defenses focus on identifying and removing poisoned samples during data collection or preprocessing [4, 8]. Trigger inversion methods attempt to recover the latent trigger of a model already compromised by a data poison. Examples include Neural Cleanse [10] and DeepInspect [2]. Other approaches, such as activation clustering [1], analyze internal network decision making to separate poisoned and benign samples based on the number of clusters formed by neural activations.

Despite their effectiveness in benchmark settings, data poisoning attacks face significant challenges in cyber-physical and controlled environments [3]. In such settings, stealthy triggers may not survive manual inspection or operational constraints. For example, in applications such as border control or airport security, visual artifacts that could plausibly serve as triggers are either prohibited or removed. Additionally, an adversary may miss the training window of a model but may still be motivated to alter model performance through cyber-enabled means. These limitations motivate the need to understand attacks and defenses that do not rely on trigger presence or poisoned training data, but rather target the weights of a model directly.

2.3 Weight Surgery

Weight Surgery, introduced by Zehavi et al. [11], is a model manipulation attack against Siamese networks to confuse classes. Weight Surgery operates exclusively on the final linear layer. Unlike data poisoning or gradient-based backdoor attacks, Weight Surgery does not modify the network backbone or require access to the training process. Instead, it directly alters the weight matrix of the final layer to induce adversarial behavior, given a few samples of the target classes, while preserving performance on non-targeted classes.

Weight Surgery introduces two primary attack objectives: *merging* and *shattering*. A successful merge attack causes the network to classify images of two different individuals as the same individual. The merging objective aims to maximize the number of sample pairs drawn from two distinct classes that are incorrectly identified as identical by the network. This objective is constrained by the requirement that the overall benign accuracy of the model remains relatively unchanged. The attack seeks to collapse the distance between representations of the two target classes at the final similarity layer while minimally disturbing the model's decision boundary for benign sample pairs.

Conversely, a successful shatter attack causes the network to classify images of the same individual as different individuals. The shattering objective seeks to maximize the number of sample pairs drawn from the same class that are incorrectly identified as different. Shattering effectively disrupts intra-class similarity at the final layer, causing genuine matches to be rejected without significantly impacting overall model performance.

Weight manipulation attacks are particularly relevant to threat models involving transfer learning. In many real-world deployments, a pre-trained Siamese network is fine-tuned by a potentially non-trustworthy third party before deployment. Weight Surgery enables an attacker to embed a backdoor during this fine-tuning stage without modifying the backbone network, distinguishing it from data poisoning attacks that rely on gradient manipulation and backpropagation that modify backbone layers. Additionally, an adversary may take advantage of common vulnerabilities to gain access to a system that either deploys or holds an image of a machine learning model.

2.4 Traditional Defensive Strategies

Utilizing a cryptographically secure hash function is a potential strategy for detecting whether a model has been attacked via Weight Surgery. Data poisoning attacks circumvent hash-based integrity checks entirely, as their modifications occur at training time. In the case of Weight Surgery, a defender may compare the hash of a model file with a server hash to determine whether a model has been modified. However, a hash-based integrity check may be insufficient to identify a Weight Surgery manipulated model for several reasons.

First, transfer learning nullifies hash-based verifications of model integrity as no prior hash value can be generated. In the transfer learning instance, a third-party fine tunes the final linear layer on a private dataset and re-releases the model. The Weight Surgery attack limits its modifications to the final linear layer, which is indistinguishable from fine-tuning. Second, the attacker may sync their attack timing with a model update to disguise the presence of the modification. Finally, even if hash-based verification worked perfectly in every attack instance, the hash only provides information that the model has been modified and provides no information on the identity of classes that have been targeted by Weight Surgery.

The stealth of the Weight Surgery attack is of considerable importance to the attacker. Siamese networks achieve high performance in controlled environments, such as border crossings, where camera angles and lighting conditions are standardized. For example, FaceNet has been reported to achieve accuracies exceeding 99% on several benchmark datasets [7]. If Weight Surgery were to introduce large deviations in model accuracy, the attack would become identifiable in an operational setting by an increase rate of false positives. As the number of attacked targets increase, the accuracy of the model on non-attacked classes begins to decline [11]. Therefore, a powerful use case of Weight Surgery would be to attack a small set of classes. When the number of attacked classes is limited, the benign accuracy remains the same or may even increase in some instances [11].

3 METHODOLOGY

We consider a Siamese network based on FaceNet [7], trained on VGGFace2 and evaluated on the LFW dataset. The adversary applies Weight Surgery exclusively to the final linear layer. As a consequence, intermediate representations produced by non-linear layers remain unmodified by the attack.

Our insight is that this architectural constraint enables the defender to exploit information encoded in earlier layers that may be inconsistent with the output of the attacked linear layer.

Following prior work on activation clustering (e.g., [1]), we seek to analyze layer-wise outputs to distinguish benign from malicious behavior. However, existing activation clustering defenses assume the presence of adversarial triggers that directly influence internal activations. In contrast, Weight Surgery does not rely on triggers or gradient-based poisoning and therefore requires a different clustering strategy.

To address this, we propose grouping outputs of intermediate layers on *sample pairs* into one of two categories: *intra-class* or *inter-class*. An intra-class sample pair contains two images of the same individual. An inter-class sample pair contains two images of two different people. In a Weight Surgery modified model, the unaffected intermediate layers will categorize sample pairs correctly with some quantifiable confidence. Assuming the Weight Surgery attack succeeds, the compromised final linear layer will categorize the sample pair as the incorrect category with some degree of confidence. The discrepancy in confidence between the clean intermediate layer and the attacked linear layer may be used to identify whether a given sample pair has been attacked at inference time.

An ideal attacker using Weight Surgery targets only a small number of classes. To reflect this threat model, we generate a total of 2000 independently attacked Siamese networks. To assess merge attacks, we limit the number of attacked classes to a single pair of classes per model. To assess shatter attacks, we limit the number of attacked classes to four classes per model.

3.1 Layer-wise Activation Extraction

We select two layers from the given Siamese network for analysis:

- A deep non-linear layer ℓ_{nl}, and
- The final linear similarity layer ℓ_{lin}.

In a deep neural network architecture there are several potential layers that may server as ℓ_{nl}. The defense performs best when selecting the deepest non-linear layer in the network. For a given input pair (x_i, x_j), we extract the flattened activation vectors from layer ℓ:

$$\mathbf{h}_\ell(x) \in \mathbb{R}^d$$

3.2 Distance Computation

For each selected layer ℓ, we compute the cosine distance between the Siamese outputs of a sample pair (x_i, x_j):

$$d_\ell(x_i, x_j) = 1 - \frac{\mathbf{h}_\ell(x_i)^\top \mathbf{h}_\ell(x_j)}{\|\mathbf{h}_\ell(x_i)\|_2 \, \|\mathbf{h}_\ell(x_j)\|_2}$$

Distances are computed independently for both ℓ_{nl} and ℓ_{lin}. We effectively treat non-linear layers as an output layer, by computing the distance on the flattened activations directly in feature space.

3.3 Distribution Construction

Layer-wise activation extraction and distance computation are performed on a clean dataset with an equal number of inter-class and intra-class samples, yielding $d_{nl}(x_i, xj)$ and $d_{lin}(x_i, x_j)$ for all sample pairs (x_i, x_j). The computed cosine distance values are used to construct two normal distributions per layer:

$$\mathcal{D}_\ell^{\text{intra}}, \quad \mathcal{D}_\ell^{\text{inter}}$$

Experimentally we found 64 intra-class sample pairs and 64 inter-class sample pairs to form normal distributions. Thus, each distribution is then modeled as a Gaussian:

$$p(d \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(d-\mu)^2}{2\sigma^2}\right)$$

where μ and σ^2 are estimated from the defender dataset. Note that the distributions are formed using the outputs of a model already compromised by Weight Surgery.

3.4 Relative Likelihood Ratio

At inference time, for a sample pair (x_i, x_j) and layer ℓ, we compute the observed distance d_ℓ and query both distributions:

$$p_\ell^{\text{intra}} = p(d_\ell \mid \mathcal{D}_\ell^{\text{intra}}), \quad p_\ell^{\text{inter}} = p(d_\ell \mid \mathcal{D}_\ell^{\text{inter}})$$

We define the relative likelihood ratio (RLR) as:

$$\text{RLR}_\ell = \frac{p_\ell^{\text{inter}}}{p_\ell^{\text{intra}} + p_\ell^{\text{inter}}}$$

This yields a normalized confidence score in the range $[0, 1]$, where higher values indicate greater likelihood that the sample pair belongs to the *inter-class* distribution.

3.5 Cross-Layer Consistency Check

We compute RLR_ℓ values for both layers:

$$\text{RLR}_{\ell_{nl}}, \quad \text{RLR}_{\ell_{lin}}$$

To detect inconsistencies introduced by Weight Surgery, we define the cross-layer confidence discrepancy:

$$\Delta = \left|\text{RLR}_{\ell_{lin}} - \text{RLR}_{\ell_{nl}}\right|$$

If both the clean intermediate layer and the modified final linear layer agree on the category of the sample pair, Δ approaches zero. The more ℓ_{nl} and ℓ_{lin} disagree, Δ approaches one. Because Δ may take on any value between $[0, 1]$, a threshold γ determines whether a given Δ value indicates that a sample pair has been affected by Weight Surgery. We define our sweep method as:

$$g(\delta) = \begin{cases} 1 & \text{if } \delta \geq \gamma \\ 0 & \text{if } \delta < \gamma \end{cases}$$

where 1 indicates a sample pair has been attacked, and 0 indicates a benign sample pair. The parameter γ controls the sensitivity of the defense and defines an operating trade-off between detection rate and false positives. To reflect a realistic deployment setting, γ is selected by fixing a target false-positive rate and estimating the corresponding value using benign sample pairs only. Once γ is fixed, recall is measured across all attacked models.

4 RESULTS

We evaluate the performance of our defense by analyzing the trade-off between the false-positive rate and recall. In a typical deployment setting, the number of benign sample pairs will out number the frequency of adversarial sample pairs by several orders of magnitude. An effective defense in a deployment setting must limit FPRs while maintaining high recall.

To determine which intermediate layer would provide the best defender distribution, we analyzed the ability of the intermediate layers to separate intra-class and inter-class samples. FaceNet possesses a deep architecture which provided us with a rich selection of layers to use as part of our defense. The best layers to select were the deepest in the model as they provide the highest separation accuracy between intra-class and inter-class sample pairs (Figure 1). We selected the last non-linear layer of FaceNet blockB.conv2d, as it had the highest separation accuracy of the non-linear layers (99%), to formulate our defender distributions D_{nl}^{intra} and D_{nl}^{inter}.

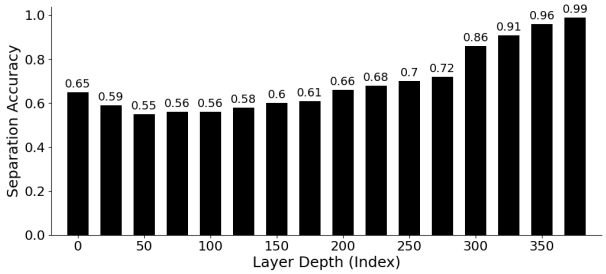


Figure 1: Layer separation accuracy versus depth location of layer in model.

4.1 Visualizing Attacked and Clean Distributions

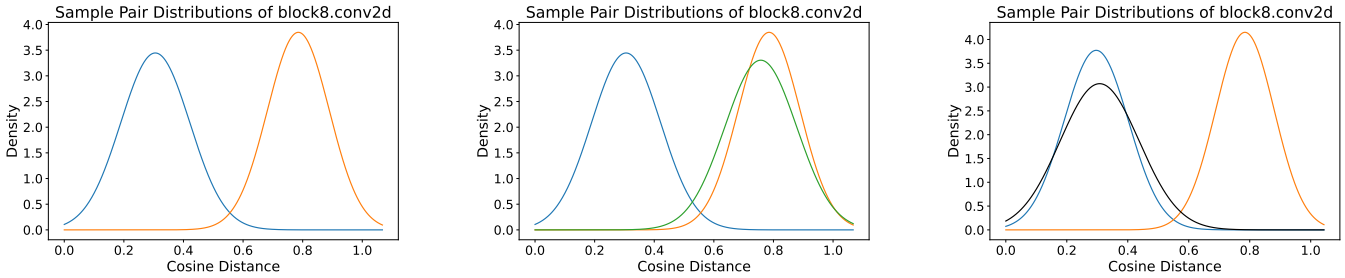
To determine whether shattered and merged samples behaved differently from benign samples, we generated distributions for ℓ_{nl} and ℓ_{lin} of a model that had undergone Weight Surgery (Figure 3). To generate the distributions, four sets of samples were used: clean intra-class samples, clean inter-class samples, shattered intra-class samples, and merged inter-class samples.

The layer ℓ_{nl} did not show abnormal behavior for clean or attacked samples. The output of ℓ_{nl} provides a strong separation of intra-class and inter-class samples (Figure 2a). The blue curve represents the distribution of clean intra-class samples, while the orange curve corresponds to clean inter-class samples. After the attack, inter-class merged samples (green curve) remain correctly categorized as inter-class, and shattered intra-class samples (black curve) remain correctly categorized as intra-class (Figures 2b and 2c). In contrast, if the attacker were to modify the non-linear layer, the distribution of inter-class merged samples would substantially overlap with that of clean intra-class samples. Similarly, the distribution of shattered intra-class samples would strongly overlap with the clean inter-class distribution.

FaceNet performs exceptionally well on the LFW dataset, and layer ℓ_{lin} demonstrates the strong separation accuracy created by the model between the distributions of clean inter-class and intra-class samples (Figure 3a). After performing Weight Surgery, ℓ_{lin} incorrectly categorized merged inter-class sample pairs as intra-class. These misclassifications are visualized as distribution overlap, where the green (merged-class) sample pairs overlap strongly with the blue (intra-class) sample pairs (Figure 3b). Additionally, ℓ_{lin} incorrectly categorized shattered intra-class sample pairs as inter-class. These misclassifications are visualized as distribution overlap, where the black (shattered-class) sample pairs somewhat overlap with the clean inter-class distribution (Figure 3c).

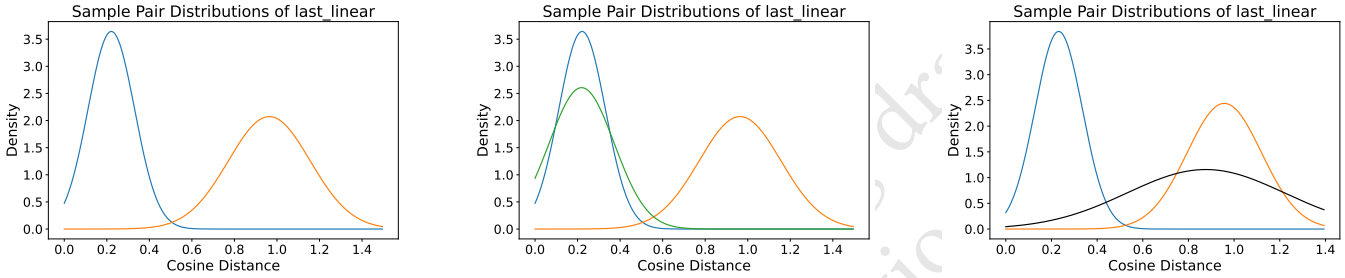
4.2 Assessing Confidence Difference Between Layers

The distributions D_ℓ^{intra} and D_ℓ^{inter} were used to generate RLR_{nl} and RLR_{lin} for a set of samples consisting of clean and attacked samples. The Δ of clean sample pairs and the Δ of merged sample pairs are linearly separable (Figure 4). The Δ of clean sample pairs are generally clustered around zero, while merged sample pairs are strongly clustered at a Δ of one. If γ were set to 0.8, then all



(a) Benign pairs **(b) Merged pairs** **(c) Shattered pairs**

Figure 2: Non-linear layer confidence discrepancy under benign and Weight Surgery conditions. Non-linear layer distributions formed from attacked samples do not differ from distributions formed from benign samples.



(a) Benign pairs **(b) Merged pairs** **(c) Shattered pairs**

Figure 3: Linear layer confidence discrepancy under benign and Weight Surgery conditions. Weight Surgery attacks induce a consistent separation from benign distributional behavior in the final linear layer across both merge and shatter objectives.

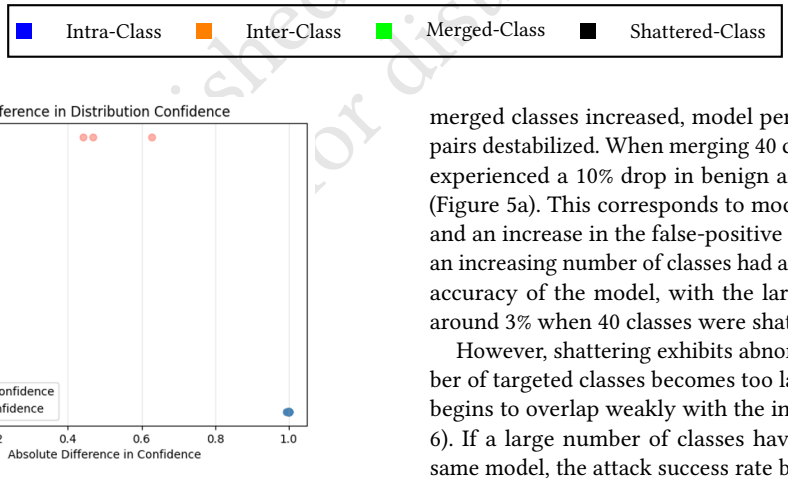


Figure 4: Results of cross-layer consistency check for merged sample pairs and clean sample pairs.

merged sample pairs would be successfully identified while also successfully identifying all clean sample pairs.

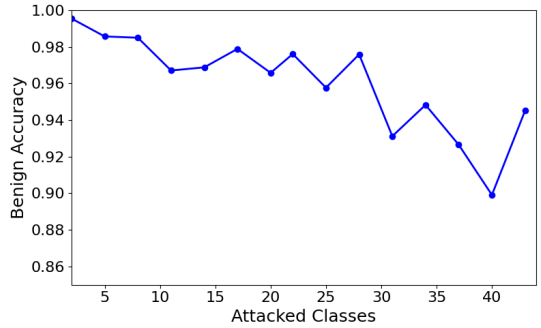
The success rate of the defense deteriorates as the number of attacked samples increases, both in the merge and the shatter case.

To understand the performance of the defense, we examined how the benign accuracy of the model scales with the number of Weight Surgery targeted classes (Figure 5). As the number of merged classes increased, model performance on benign sample pairs destabilized. When merging 40 class pairs together, the model experienced a 10% drop in benign accuracy from 99.8% to 89.8% (Figure 5a). This corresponds to model performance degradation and an increase in the false-positive rate in the model. Shattering an increasing number of classes had a smaller impact on the benign accuracy of the model, with the largest drop in accuracy being around 3% when 40 classes were shattered (Figure 5b).

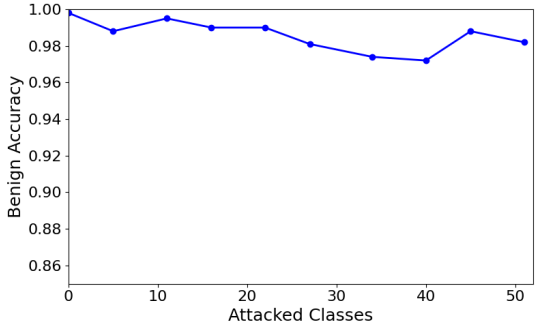
However, shattering exhibits abnormal behavior when the number of targeted classes becomes too large. The shatter distribution begins to overlap weakly with the inter-class distribution (Figure 6). If a large number of classes have been shattered within the same model, the attack success rate begins to decline as the model correctly identifies shattered sample pairs as intra-class.

4.3 Defense Performance

To evaluate the defense, we simulated merge and shatter attacks on independent models. For merge attacks, we randomly selected one class pair to attack and repeated this procedure 1000 times. We set γ to achieve a target FPR and observe that recall declines smoothly as the FPR requirement becomes more stringent (Figure 7a). At a target FPR of 0.001, the defense achieves 95.6% recall. Even under stricter constraints, performance remains high, with recall of 92.5% at FPR = 0.0001 and 88.9% at FPR = 0.000016.



(a) Merged pairs versus benign accuracy



(b) Shattered pairs versus accuracy

Figure 5: Accuracy degradation accumulates as the number of Weight Surgery targeted classes increases in both the merge and shatter case.

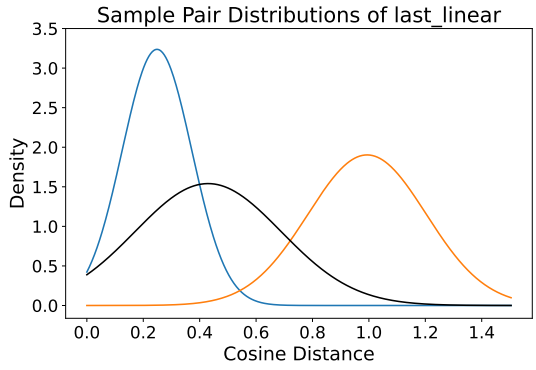


Figure 6: Visualizing the shatter distribution when the number of attacked classes is high (N=57).

We performed the same simulation for shatter attacks, selecting four random classes to shatter per model and repeating the process 1000 times. At a target FPR of 0.001, the defense achieves 92.0% recall. Recall decreases gradually at stricter FPR thresholds, reaching 82.1% under the most conservative operating condition.

5 DISCUSSION

Our analysis of the Weight Surgery attack reveals several important insights regarding its effectiveness, limitations, and practical considerations.

First, the attacks demonstrate high efficacy when the number of targeted pairs is minimized. In the case of shatter attacks, performance initially improves as more pairs are attacked but eventually reaches a point where the attack success rate begins to decline. For merging attacks, as the number of targeted pairs increases, the benign accuracy of the model becomes compromised. The attacker is faced with a trade-off between the number of attacked classes and the stealthiness of the attack. Therefore, the ideal attacker scenario targets a few selected classes with Weight Surgery.

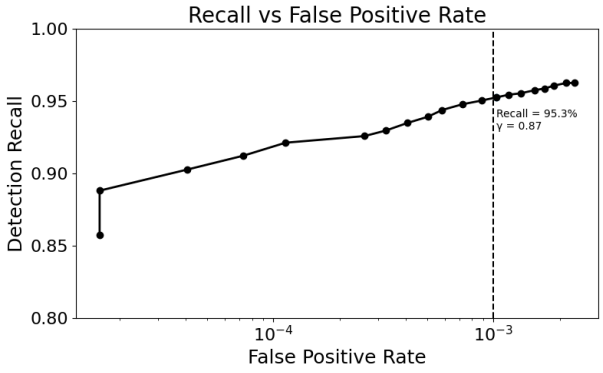
As the number of shattered classes increases, the effectiveness of the shatter attack diminishes. This behavior may be visualized by the change in distribution overlap: the overlap between shattered sample pairs and clean inter-class sample pairs decreases as the number of targeted classes increases, which in turn reduces the attack's potency (Figure 6). This suggests that the success rate of the shatter attack declines as the number of targeted classes increases. The distribution formed by shattered intra-class samples lies between the clean distributions of inter-class and intra-class samples.

In contrast, merge attacks maintain high effectiveness even as the number of targeted pairs increases. This comes at the cost of benign performance: the linear layer experiences a noticeable drop in accuracy, resulting in an increased rate of false positives. The attack results in an uptick in the false positive rate in benign samples, losing its stealthiness and signaling to the operator that the model has become dysfunctional.

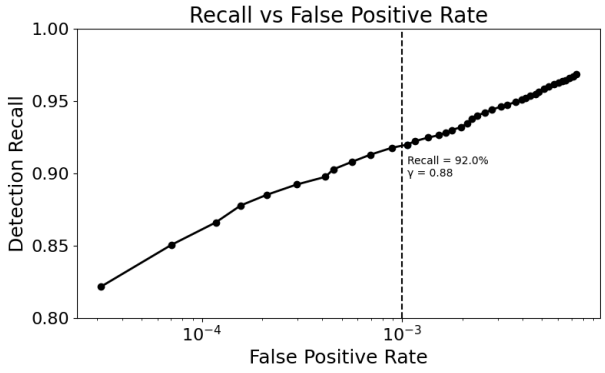
Under ideal attacker conditions, where only a small number of samples are merged or shattered, the proposed defense effectively identifies sample pairs as adversarial or clean. However, shatter attacks are more challenging to detect than merge attacks. This difficulty arises from the interaction between the threshold parameter γ and the distribution of shattered samples. Shatter attacks increase the angular distance between intra-class sample pairs, but the magnitude of this widening varies widely, from 0 to 180 degrees. The defense assumes that both shatter and merge attacks increase the difference between RLR_{nl} and RLR_{lin}. In the case of shattering, this difference is sometimes insufficient for detection, allowing a subset of attacker sample pairs to escape identification.

The computational and memory requirements of the proposed approach are minimal. Computationally, the method requires only a one-time generation of four distributions, followed by an additional distance computation per sample pair on the output of ℓ_{nl}. Each sample pair requires four queries to D_{nl}^{intra}, D_{nl}^{inter}, D_{lin}^{intra}, and D_{lin}^{inter}. The cost of querying is negligible relative to the cost of forward propagation. Memory overhead is similarly modest. Storing the four distributions requires only a few hundred bytes, and caching non-linear layer outputs may require several kilobytes, a small amount compared to the overall size of FaceNet which is several hundred megabytes.

These results highlight the tradeoffs between the number of attacked samples and the model integrity the adversary must make when implementing Weight Surgery on a target model. The attacker



(a) Merged attack detection performance



(b) Shattered attack detection performance

Figure 7: Recall performance of proposed defense at FPR of 0.001 when identifying merged and shattered sample pairs.

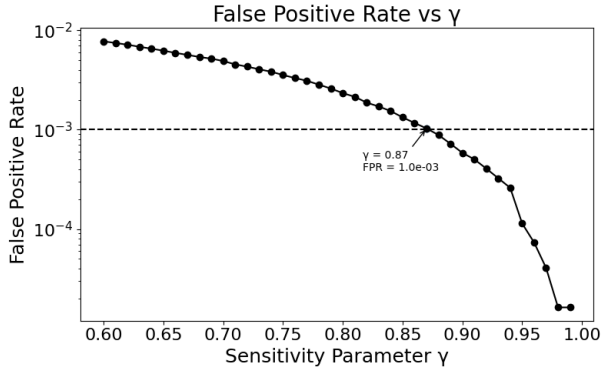


Figure 8: γ selection for false-positive rate of 0.001

is interested in maintaining benign accuracy while modifying the model with respect to a few target classes. The defense is robust under the ideal attack conditions while requiring few additional computational and memory resources. The proposed defense provides an efficient and effective approach for identifying classes that have been affected by Weight Surgery.

6 LIMITATIONS

Several limitations of Weight Surgery and the proposed defense must be acknowledged. First, both the attack and defense are reliant on model architecture. Our experiments focus on Siamese networks with a final linear layer, and the methods described may not generalize to other architectures or network designs. The structural assumptions of the attack and defense restrict their applicability to standard classifiers.

Second, the defense relies on the integrity of non-linear layer representations. If an adversary were to modify or perturb the non-linear layers, the defense's ability to detect attacks may be compromised. The current approach is effective only if the intermediate representations produced by non-linear layers remain untouched.

Third, the defense does not perform uniformly across all classes. Certain class pairs are inherently more difficult to detect when attacked. The difficulty arises from the confidence produced by the non-linear layer. For example, merging highly dissimilar individuals, such as Morgan Freeman and Scarlett Johansson, can be easily identified because the non-linear layer confidently signals that the sample pair is of two different people. However, merging visually similar individuals may cause the non-linear layer to produce weak or incorrect signals that reduce the effectiveness of the defense. The variability in the output of the non-linear layer demonstrates that attack detectability may depend strongly on the pre-existing similarity between targeted classes.

These limitations highlight the contextual nature of both the attack and defense. Their performance is constrained by model architecture, assumptions about intermediate representations, and the inherent difficulty of distinguishing certain sample pairs. Future work should explore adaptations that mitigate these dependencies.

7 FUTURE WORK

There are several promising directions to extend this line of research to further explore the space of Weight Surgery attacks and defenses.

One potential direction is to train non-linear classifiers to identify the attack. Our current approach utilizes a linear decision boundary to distinguish benign and malicious sample pairs. The defender may generate a synthetic dataset by implementing Weight Surgery and collecting layer activations which may be used to train a deep classifier.

Another direction involves incorporating prior knowledge into the defense. Currently, our method assumes equal priors of intra-class and inter-class sample pairs which may not reflect real-world conditions. For instance, in controlled settings such as border security, the frequency of intra-class sample pairs far outweighs the frequency of inter-class sample pairs.

Investigating per-sample hardness is another important extension. As noted in the limitations of this approach, some sample pairs are inherently more difficult to detect due to their visual similarity. Developing metrics or methods to quantify and adapt to per-sample difficulty could improve the robustness and precision of the defense.

The final suggested direction is to analyze the application of Weight Surgery to a Federated Learning context. Evaluating this

weight manipulation attack against traditional Federated Learning defenses would provide insights into the robustness of these defenses against non data-poisoning style attacks.

8 CONCLUSION

Facial recognition systems are increasingly deployed in critical cyber-physical settings, yet remain vulnerable to model-targeted attacks. We propose a computationally lightweight novel defense that detects instances of Weight Surgery, a weight manipulation attack that merges or shatters classes, without requiring the compromise of training data. Across independently attacked models, our method achieves over 95% recall at a target FPR of 0.001 and remains robust at stricter thresholds, demonstrating effectiveness under realistic deployment conditions. These results show that model-targeted attacks can be effectively mitigated while incurring minimal memory and computational overhead, enabling safer deployment of machine learning systems in sensitive applications.

9 CITATIONS AND BIBLIOGRAPHIES

REFERENCES

[1] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian M. Molloy, and Biplav Srivastava. 2019. Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering. In *Workshop on Artificial Intelligence Safety 2019 co-located with the Thirty-Third AAAI Conference on Artificial Intelligence 2019 (AAAI-19), Honolulu, Hawaii, January 27, 2019.* https://ceur-ws.org/Vol-2301/paper_18.pdf

[2] Huili Chen, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. 2019. DeepInspect: A Black-box Trojan Detection and Mitigation Framework for Deep Neural Networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019.* https://doi.org/10.24963/ijcai.2019/647

[3] Antonio Emanuele Cinà, Kathrin Grosse, Ambra Demontis, Sebastiano Vascon, Werner Zellinger, Bernhard Alois Moser, Alina Oprea, Battista Biggio, Marcello Pelillo, and Fabio Roli. 2023. Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning. *ACM Comput. Surv.* (2023). https://doi.org/10.1145/3585385

[4] Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Madry, Bo Li, and Tom Goldstein. 2023. Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses. *IEEE Trans. Pattern Anal. Mach. Intell.* (2023). https://doi.org/10.1109/TPAMI.2022.3162397

[5] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. *arXiv preprint arXiv:1708.06733* (2017). http://arxiv.org/abs/1708.06733

[6] Minzhou Pan, Yi Zeng, Lingjuan Lyu, Xue Lin, and Ruoxi Jia. 2023. ASSET: Robust Backdoor Data Detection Across a Multiplicity of Deep Learning Paradigms. In *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023.* https://www.usenix.org/conference/usenixsecurity23/presentation/pan

[7] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015.* https://doi.org/10.1109/CVPR.2015.7298682

[8] Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P. Dickerson, and Tom Goldstein. 2021. Just How Toxic is Data Poisoning? A Unified Benchmark for Backdoor and Data Poisoning Attacks. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event.* http://proceedings.mlr.press/v139/schwarzschild21a.html

[9] Transportation Security Administration. 2024. Facial Comparison Technology. https://www.tsa.gov/news/press/factsheets/facial-recognition-technology. Accessed: 2025-01.

[10] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. 2019. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019.* https://doi.org/10.1109/SP.2019.00031

[11] Irad Zehavi and Adi Shamir. 2023. Facial Misrecognition Systems: Simple Weight Manipulations Force DNNs to Err Only on Specific Persons. *arXiv preprint arXiv:2301.03118* (2023). https://doi.org/10.48550/arXiv.2301.03118

[12] Rui Zhu, Di Tang, Siyuan Tang, Zihao Wang, Guanhong Tao, Shiqing Ma, XiaoFeng Wang, and Haixu Tang. 2024. Gradient Shaping: Enhancing Backdoor Attack Against Reverse Engineering. In *31st Annual Network and Distributed System Security Symposium, NDSS 2024, San Diego, California, USA, February 26 - March 1, 2024.* https://www.ndss-symposium.org/ndss-paper/gradient-shaping-enhancing-backdoor-attack-against-reverse-engineering/

ACKNOWLEDGMENTS

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND0000-XXXXX

This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

Received 16 January 2026; revised N/A; accepted N/A