

Only My Model On My Data: A Privacy Preserving Approach Protecting one Model and Deceiving Unauthorized Black-Box Models

Weihseng Chai*, Brian Testa*, Huantao Ren, Asif Salekin, Senem Velipasalar

Syracuse University

{wchai01, bptesta, hren11, asalekin, svelipas}@syr.edu,

Abstract

Deep neural networks are extensively applied to real-world tasks, such as face recognition and medical image classification, where privacy and data protection are critical. Image data, if not protected, can be exploited to infer personal or contextual information. Existing privacy preservation methods, like encryption, generate perturbed images that are unrecognizable to even humans. Adversarial attack approaches prohibit automated inference even for authorized stakeholders, limiting practical incentives for commercial and widespread adaptation. This pioneering study tackles an unexplored practical privacy preservation use case by generating human-perceivable images that maintain accurate inference by an authorized model while evading other unauthorized black-box models of similar or dissimilar objectives, and addresses the previous research gaps. We validate the efficacy of our proposed solutions across three distinct datasets and diverse models. The datasets employed are ImageNet, for image classification, Celeba-HQ dataset, for identity classification, and AffectNet, for emotion classification. Our results show that the generated images can successfully maintain the accuracy of a protected model and degrade the average accuracy of the unauthorized black-box models to 11.97%, 6.63%, and 55.51% on ImageNet, Celeba-HQ, and AffectNet datasets, respectively.

1 Introduction

In the contemporary landscape, the ubiquitous exchange and sharing of images plays an integral role across diverse platforms. These images serve various purposes, from the presentation of personal photos on LinkedIn profiles to institutional web pages and the verification process of ID cards. Moreover, the digital realm witnesses the widespread sharing of images, contributing to various practical applications, such as surveillance systems, object recognition technologies, and the identification of license plates.

The essence of this image-sharing paradigm is deeply entrenched in the operational functionality of platforms, particularly in enabling the presentation and validation of cru-

cial information related to users, employees, candidates, and more. This validation process incorporates a dual approach, *employing visual human assessment alongside AI-driven automated verification systems*.

Despite its indispensable utility, the act of sharing images triggers substantial privacy apprehension among users [Xu *et al.*, 2011; Trepte, 2021; De Wolf, 2020]. Image data can be harvested to infer personal information about the subjects [Fei *et al.*, 2020; Wang and Kosinski, 2018; Wang, 2022; Nicol Turner Lee, 2022], or can be used to gather contextual information about the subject’s surroundings [Vlontzos *et al.*, 2019; Li *et al.*, 2021].

This paper embarks on a pioneering exploration aimed at mitigating these privacy concerns. The main goal is to provide a framework through which companies and institutions can harness user images for automated inference and human visualization while upholding user privacy.

Towards the aforementioned goal, this paper addresses a fundamental question: *Can we strategically add minimal noise to an image for a specific task while satisfying three constraints: (i) ensuring these images maintain their perceptual functionality to humans; (ii) enabling a designated/ authorized model for a target task, trained by the stakeholder, to derive accurate inferences from these altered images, and (iii) causing any (black-box) unauthorized machine learning models to be incapable of generating accurate inferences for the same altered image in that same or even a different task?*

We can broadly identify two lines of related research in this area: (i) On one side, there exists privacy-preserving methods that focus on image encryption [Huang *et al.*, 2019] to prevent access by unintended 3rd-parties in such a way that the images become imperceptible by humans. This makes these approaches not suitable for the context mentioned above, where one of the goals is to disseminate the images widely for human visual comprehension as well; (ii) the second line of research focuses on adversarial evasion techniques [Chakraborty *et al.*, 2018; Carlini and Wagner, 2017], which aim to keep images still perceivable by humans (by restricting perturbations according to an L-norm constraint) while deceiving the inference capability of black-box machine learning models. These methods only focus on attacks and do not address the protection of an authorized model.

That said, multiple works by Kwon *et al.* [Kwon *et al.*, 2018; Kwon *et al.*, 2019b; Kwon *et al.*, 2022] have considered

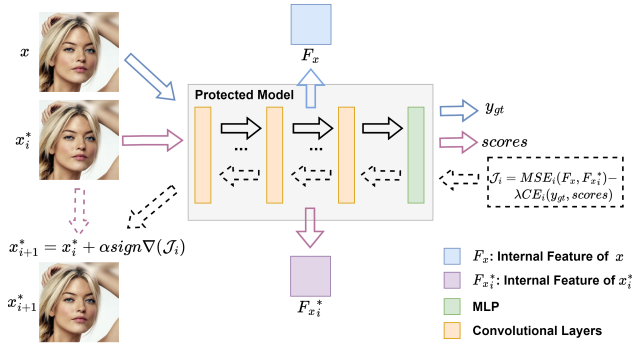


Figure 1: Proposed feature-based privacy protection. x denotes the input image, x_i^* and x_{i+1}^* denote the generated image after i^{th} and $i + 1^{th}$ iteration, respectively.

this problem as a multi-objective loss function that allows development of adversarially evasive samples that both decrease the loss for an **authorized model** (for which we want inferences to be successful) while also increasing the loss for **unauthorized models** (which we want to degrade). Yet, this approach requires a white-box setting, i.e. when generating image variants it is assumed that they have a priori knowledge of *both* the authorized and unauthorized models.

Our work is different from all the aforementioned approaches, and fills a gap not addressed by prior studies. *This work is the first of its kind to protect the privacy of image data by protecting a known, authorized network, while simultaneously degrading unknown, unapproved inference of other networks, and keeping the images still perceivable by humans.*

Our approach works with a black-box setting for the unauthorized models. It is based on feature map distortion and does not require a surrogate model and assumes knowledge of only the authorized model (see Fig. 1). To accomplish this, the proposed approach relies on a key insight: early layers within a neural network perform extraction of features, which are then used for inference in later layers [Wang *et al.*, 2020; Dosovitskiy *et al.*, 2014; Kim and Kim, 2017]. This work hypothesizes that disruptions to the feature representation would transfer to the feature extraction layers of other networks. So by optimizing for a multi-objective loss that both disrupts the feature representation **and** protects the inference performed by the authorized classifier, this should generate transferable adversarial examples (AEs) for previously unseen networks. Evaluations have been performed on multiple image datasets with various authorized and unauthorized classifiers. As the pioneer paper to tackle the aforementioned practical objective, this work makes the following contributions:

- We propose a novel method for data/privacy protection in different classification tasks, especially effective for tasks involving a relatively large number of classes.
- While protecting an authorized model, the proposed approach can successfully deceive unauthorized black-box models.
- We conduct experiments on various datasets to verify the effectiveness of our approach. The first dataset is ImageNet [Deng *et al.*, 2009], which is for general im-

age classification. The other two datasets involve more privacy-sensitive aspects. More specifically, CelebA-HQ [Liu *et al.*, 2015] and AffectNet [Mollahosseini *et al.*, 2017] datasets are used for identity classification and facial expression classification task, respectively.

- We explore the feasibility of our approach when attacked models are designed for different tasks than the protected model, i.e. a cross-task scenario.
- We perform extensive ablation studies on different aspects of the proposed approach.

2 Related Work

There is a myriad of works that consider evasion of image classifiers, including both white-box [Goodfellow *et al.*, 2014; Chakraborty *et al.*, 2018] and black-box [Papernot *et al.*, 2017; Carlini and Wagner, 2017] attacks. In a white-box scenario, the attacker possesses complete knowledge of the target model, including its architecture, parameters, weights, and even the training data. This enables the attacker to customize the attack, often using gradient-based methods to find the most effective perturbations to the input data.

In the black-box scenario [Guo *et al.*, 2019; Ilyas *et al.*, 2018], the attacker has limited or no knowledge of the target model. Query-based methods were proposed, assuming the attackers' access to the input/output of the target model [Dang *et al.*, 2017; Juuti *et al.*, 2019]. Unlike query-based evasion, transferable attacks do not require querying and generate evasive samples across different models [Maho *et al.*, 2021].

In both cases, these attacks consider attack effectiveness, and are constrained by some notion of perceptibility. This often comes in the form of maximum pixel perturbation ϵ to limit the amount of disruption caused by the attack. Yet, the goal is designing very effective attacks, without focusing on protecting an authorized model.

Kwon *et al.* [Kwon *et al.*, 2018; Kwon *et al.*, 2019b; Kwon *et al.*, 2022] addressed this shortfall by protecting a single image classifier and simultaneously disrupting one or more target image classifiers. These works have also been extended to other applications [Kwon, 2023; Kwon *et al.*, 2019a; Ko *et al.*, 2023; Kwon, 2021]. The image-based attack papers used the *white-box*, transformation-based Carlini-Wagner attack [Carlini and Wagner, 2017], and modified the loss function of the transformation network to include *distortion*, *friend*, and *enemy* loss terms. The *distortion* term penalized large changes to the original image (constraining the attack to some ϵ), while the *friend* and *enemy* terms discouraged changes that impacted the protected (friend) classifier while degrading the targeted (enemy) classifiers.

While achieving success in the intended objective, these studies required that the protected and targeted classifiers share the same topology. Moreover, they necessitated white-box access to both the protected and targeted classifiers. This white-box prerequisite significantly restricts their practicality and renders them inapplicable to the use cases this paper addresses.

3 Protect & Attack Model

We first provide the context under which this work was evaluated, including the attack goal, access levels and assumptions.

Attack Goal. Maintain inference performance of an authorized, white-box model while simultaneously degrading black-box, unauthorized models *and* maintaining perceptual functionality for humans.

Proposed System Access Level. The **authorized model** is an image classifier for which we have access to the topology, weights, and inference results. The **unauthorized models** are unknown black-box models for which we have neither access to the weights nor to the topologies. Further, we do not have query access to the target models.

Assumptions. The primary assumption is that all of the models perform inference based upon image pixel data, and not based on any other format or metadata-specific information. This work further assumes that, before an image is shared with a broader group, the image can be preprocessed to introduce privacy-preserving perturbations. This assumes adequate access to the authorized model.

Outcome. The desired outcome is a set of adversarially evasive images that are accurately classified by the authorized model, misclassified by the unauthorized models, and perceptually similar to the original images.

4 Methodology

The primary objective of our study is to create perturbed images that safeguard an authorized model while diminishing the performance of unauthorized target models. This objective is formulated as follows:

$$\begin{aligned} x^* = \operatorname{argmax}_{x^*} \min_{\forall t \neq p} \mathcal{L}_t \\ \text{subject to } f_p(x^*) = f_p(x), \end{aligned} \quad (1)$$

where x^* represents the generated image, f_p is the authorized model, and \mathcal{L}_t signifies the loss of the target models. Optimizing the image to satisfy $f_p(x^*) = f_p(x)$ presents practical challenges. Therefore, we pivot our strategy to focus on minimizing the task-specific loss \mathcal{L}_p of the authorized model. This revised objective is delineated as follows:

$$x^* = \operatorname{argmax}_{x^*} \min_{\forall t \neq p} \mathcal{L}_t \quad \operatorname{argmin}_{x^*} \mathcal{L}_p. \quad (2)$$

In contrast to [Kwon *et al.*, 2018], our work considers targeting unauthorized models within a black-box setting, aligning more closely with real-world scenarios. We propose a generalized solution to this problem, as detailed in Eq. (3), where \mathcal{L}' denotes the loss that is transferable to target models. When utilizing only an authorized model for generation, this loss originates from the authorized model, as shown in Eq. (5).

$$x^* = \operatorname{argmax}_{x^*} \mathcal{L}' \quad \operatorname{argmin}_{x^*} \mathcal{L}_p. \quad (3)$$

This can also be written as:

$$x^* = \operatorname{argmax}_{x^*} \mathcal{J}, \quad (4)$$

where $\mathcal{J} = \mathcal{L}' - \lambda \mathcal{L}_p$, and λ is the parameter to control the direction of the gradient.

During image generation, we iteratively update the image in accordance with the loss direction, as outlined in Algorithm 1. To maintain the visual quality of the generated image, we employ a budget constraint ϵ , utilizing the l_∞ norm as a measure. The impact of ϵ on both performance and image quality is extensively discussed in Sec. 8.3.

Algorithm 1 Image Privacy Protection

Input \mathcal{L}' : transferable attack loss; \mathcal{L}_p : task-specific loss for the authorized model; x : input image; x^* : generated image; N : number of iterations; α : the step size for updating the image; ϵ : the budget for modifying the pixels

```

1: procedure GENERATEIMAGE
2:    $x_0^* = x, i = 1$ 
3:   while  $i < N + 1$  do
4:      $\mathcal{J}_{i-1} \leftarrow \mathcal{L}'(x_{i-1}^*) - \lambda \mathcal{L}_p(x_{i-1}^*)$ 
5:      $x_i^* \leftarrow x_{i-1}^* + \alpha \operatorname{sign}(\nabla(\mathcal{J}_{i-1}))$ 
6:      $x_i^* \leftarrow \min(x_i^*, x + \epsilon, 255)$ 
7:      $x_i^* \leftarrow \max(x_i^*, x - \epsilon, 0)$ 
8:      $i \leftarrow i + 1$ 
9:   end while
10:  return  $x_N^*$ 
11: end procedure

```

As stated previously, our objective is to evade unauthorized black-box classifiers while protecting an authorized white-box classifier. To achieve this goal, we propose a *feature map distortion* (FMD) approach, which generates the perturbed image with only the authorized model by using the Mean squared Error (MSE) of the internal feature map as the transferable loss (\mathcal{L}'). The cross-entropy (CE) loss of the authorized model is used as the task-specific loss (\mathcal{L}_p). Thus, the loss \mathcal{J} in Eq. (4) can be expressed as follows:

$$\mathcal{J} = \text{MSE}(f_{\text{feat}}(x), f_{\text{feat}}(x^*)) - \lambda \text{CE}(f_{\text{logits}}(x^*), y_{\text{gt}}), \quad (5)$$

where f_{feat} denotes the output of the internal convolution layers of the authorized model, $f_{\text{logits}}(x^*)$ is the output score of the authorized model with image x^* , and y_{gt} is the ground-truth label of the original image x . When initially generating x_0^* as $x_0^* = x$, the gradient of the MSE loss will be 0. Thus, we uniformly randomly initialize x_0^* in the $[x - \epsilon, x + \epsilon]$ range.

The proposed \mathcal{J} loss in Eq. (5) serves the dual objective: degrading the performance of unauthorized models while safeguarding the authorized one. First, employing the MSE loss creates a feature representation of the perturbed image x^* , modified by adding noise, which significantly differs from the original feature map. This variance leads to misclassification, meaning there is no valid input-to-output path for commonly trained models $\forall f_t(x^*)$ with the same objective. However, considering the nature of deep learning models that establish numerous paths from input to output [Wang *et al.*, 2021], the use of CE loss ensures the retention of at least one pathway from input to output for the authorized model $f_p(x^*)$, serving the second objective of safeguarding it.

5 Experimental Results

We commence by presenting results from our proposed protection approach, which only employs the authorized model to generate the perturbed images, as introduced in Sec. 4.

This approach is applied to different classification tasks, with varying number of classes, using three different datasets, namely ImageNet [Deng *et al.*, 2009], Celeba-HQ [Liu *et al.*, 2015], and Affectnet [Mollahosseini *et al.*, 2017] datasets.

5.1 Results on the ImageNet Dataset

We use six pretrained models from the torchvision [Marcel and others, 2023], namely VGG11 [Simonyan and Zisserman, 2015], VGG16 [Simonyan and Zisserman, 2015], ResNet18 [He *et al.*, 2016], ResNet34 [He *et al.*, 2016], Wide-ResNet50-2 [Zagoruyko and Komodakis, 2016], and MobileNet-V2 [Sandler *et al.*, 2018], which have been trained on the ImageNet dataset, which includes 1000 classes. From the ImageNet validation set, we selected 5,000 images, which were correctly classified by all six models. We set the budget parameter ϵ as 16, step size α as 4, number of steps N as 100, and the weight factor λ as 1. The most effective layer was chosen for each model to obtain the results presented in Table 1. Detailed discussion on layer selection and parameter settings can be found in Sec. 8.1 and 8.2, respectively.

The models in the first column of Table 1 are the authorized models used during image generation, while the models listed in the top row are the unauthorized models used for testing. Thus, diagonal entries indicate the accuracy of the authorized model. As evident from the table, we achieve a remarkable average protected accuracy of 100% while significantly reducing the accuracy of the target models from 100% to an average of just 11.97% across six different models.

Accuracy (%)	VGG11	VGG16	ResNet18	ResNet34	W-Res50-2	Mob-v2
VGG11	100	15.38	10.56	20.68	12.66	14.26
VGG16	6.62	100	8.54	16.54	9.82	9.98
ResNet18	12.52	8.78	100	25.5	16.36	10.18
ResNet34	8.72	5.36	5.78	100	12.46	6.78
W-Res50-2	7.14	5.32	4.96	14.92	100	9.28
Mob-v2	10.02	10.4	13.18	26.02	20.34	100

Table 1: Experimental results on the ImageNet dataset with six different models on a 1000-category classification task.

5.2 Results on Celeba-HQ dataset

We employ the same models as in Sec. 5.1, fine-tuning them on the Celeba-HQ dataset containing 30,000 images to classify faces. We selected images of 307 subjects, each with 15 or more images. A split of 4263 images (80%) for training and 1215 images (20%) for testing was used. Post-training, we selected 634 testing images that were correctly classified by all six models for our experiments. The parameters α , ϵ , N , and λ were kept consistent with those used in Sec. 5.1. Results are presented in Tab. 2. The models in the first column represent the authorized models, while the model names in the top row are the attacked models for testing. The diagonal entries indicate the protected accuracy. We achieve an average protected accuracy of 99.92% while significantly reducing the average accuracy of target models from 100% to just 6.63% across six different models.

5.3 Results on the AffectNet dataset

We employ the same models as in Sec. 5.1, fine-tuning them for facial expression classification on the AffectNet dataset.

Accuracy (%)	VGG11	VGG16	ResNet18	ResNet34	W-Res50-2	Mob-v2
VGG11	100	2.05	5.99	7.73	9.78	3.79
VGG16	2.37	99.53	4.73	6.78	7.73	1.1
ResNet18	2.52	1.1	100	5.84	8.04	1.1
ResNet34	3	1.1	3.79	100	12.78	1.1
W-Res50-2	11.99	2.68	4.1	5.52	100	3.6
Mob-v2	21.45	10.09	11.51	13.41	22.24	100

Table 2: Experimental results on the Celeba-HQ dataset on a 307-category classification task.

AffectNet contains eight facial expressions: neutral, happy, angry, sad, fear, surprise, disgust, and contempt. The owner splits the data, such that 287,651 images are for training and 3,999 for testing. Post-training, we selected 1346 testing images that were correctly classified by all six models for our experiments. The parameters α , ϵ , and N were kept consistent with those used in Sec. 5.1. λ setting is determined by experiments. The results are presented in Tab. 3. The models in the first column represent the authorized models used during image generation, while the models in the top row are attacked models used for testing. The diagonal entries indicate the protected accuracy. We achieve an average protected accuracy of 99.67%, while significantly reducing the average accuracy of target models from 100% to just 55.5% across six different models. The performance drop on the unauthorized models for this task is not as significant as that achieved with the classification tasks on the other two datasets. We deduce that this is related to this task being less complex than others. More specifically, in Sec. 5.1 and 5.2, the classification tasks involve 1000 and 307 classes, respectively. With AffectNet, on the other hand, there are only eight classes. Our hypothesis is further verified in Sec. 8.4.

Accuracy (%)	VGG11	VGG16	ResNet18	ResNet34	W-Res50-2	Mob-v2
VGG11	99.78	47.92	57.88	66.57	70.13	58.47
VGG16	55.72	99.48	62.85	70.06	75.71	70.58
ResNet18	51.19	44.28	100.00	42.94	51.19	50.67
ResNet34	61.74	42.05	26.30	98.74	60.55	54.83
W-Res50-2	42.27	37.00	35.88	36.26	100.00	42.35
Mob-v2	69.69	60.10	66.42	78.01	75.78	100.00

Table 3: Experimental results on the AffectNet dataset for 8-class classification.

6 Cross-task feasibility study

The results presented above in Sec. 5 show that the generated images from an authorized model can deceive unauthorized target models performing the same task. Since we employ the MSE loss between the internal feature maps as a transferable loss candidate, we further explore if our proposed approach can transfer across *different tasks*. In other words, we performed additional experiments to study if the images generated from an authorized model designed for task A can also deceive unauthorized target models designed for task B, where A and B are different tasks. Here, different tasks could involve many classes, e.g., image classification, object detection, etc., or a smaller number of classes, e.g., facial expression classification, ethnicity classification, etc.

In the first cross-task experiment, we protect image classification models and attack object detection models. We generate the images from ImageNet, and attack three object detection models, namely Faster-RCNN [Ren *et al.*, 2015], Reti-

mAP(%)	Original	VGG11	VGG16	ResNet18	ResNet34	W-Res50-2	Mob-v2
Faster-RCNN	100	3.48	2.63	4.62	4.02	4.43	5.61
RetinaNet	100	3.49	2.55	4.94	4.24	4.38	6.26
MobileNet-V3-SSD	100	2.71	1.89	5.58	5.5	3.18	3.79

Table 4: The cross-task experiment results on the ImageNet dataset.

naNet [Lin *et al.*, 2017], and MobileNet-V3-SSDLite [Liu *et al.*, 2016], which are trained on the COCO dataset [Lin *et al.*, 2014]. In Tab. 4, the models in the first row are the authorized models, from which the images are generated, and the model listed in the first column are the attacked object detection models. In this evaluation, we consider the output from the original image as the ground truth, so the original mean average precision (mAP) is 100%. Then, we calculate the mAP on the generated images based on this ground truth. It can be seen that with our proposed method, when protecting six different models, the generated images can drop the average mAP of three target object detection models to 4.13%. Different factors can cause this drop in mAP: (i) bounding-boxes appear or disappear compared to the original image input; (ii) bounding-boxes shift compared to the original image; (iii) predicted class changes for a specific object compared with the original image. Some example detection results are shown in Fig. 2, where objects are completely missed in the 2nd and 3rd rows, and a false detection appears in the 1st row.

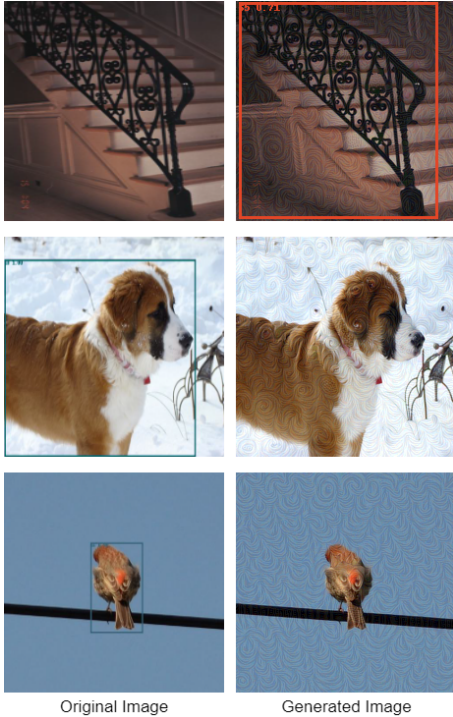


Figure 2: Examples of object detection results from MobileNet_V3_SSDLite with the original and generated images showing the false (first row) and missed detections (2nd and 3rd rows).

In the second cross-task experiment, we use two different datasets and generate images from the Celeba-HQ dataset and AffectNet datasets by protecting models trained for face identity and facial expression classification, respectively. We at-

tack an open-source ethnicity classification model [Serengil and Ozpinar, 2020] as the target model, which has been trained for a different task than the authorized models. The results are shown in Tab. 5. We consider the prediction of the original images as ground truth since we do not have the ground truth label for ethnicity, so the original accuracy is 100%. It can be seen that with our proposed method, when protecting six different models, the generated images can drop the average accuracy to 74.39% and 63.50% on Celeba-HQ and AffectNet datasets, respectively.

Accuracy (%)	Original	VGG11	VGG16	ResNet18	ResNet34	W-Res50-2	Mob-v2
Celeba-HQ	100	76.92	76.92	76.92	73.08	73.08	69.23
Affectnet	100	71.55	72.51	59.58	59.73	52.82	64.78

Table 5: Experimental results on Celeba-HQ and Affectnet datasets for across task experiment.

Notably, all the experiments are done with the images generated from the corresponding datasets in Sec. 5. It can be seen that for the ImageNet dataset when we protect image classification models, we can successfully deceive the unauthorized object detection models. For the evaluation when attacking an ethnicity classification model, the attack performance on the AffectNet dataset is good, with the average accuracy dropping to 63.50%. On the Celeba-HQ dataset, the average accuracy only drops to 74.39%. The results show that the cross-task transferability is affected by the tasks of the protected and attacked models, and by model characteristics. Future work will focus on designing a privacy protection method that can deceive unauthorized models, designed for different tasks, by reducing the dependency of the performance on the authorized model.

7 Analysis of Results

The results above demonstrate that FMD is an effective method for evading previously unseen black-box models while protecting inference performed by a known authorized model. The intuition regarding this method is described briefly in Sec. 1 and 4. This section attempts to provide concrete analysis to support this intuition.

Consider Fig. 3, which shows an example image from the ImageNet dataset and a protected variant of that image generated by applying FMD to a VGG16 classifier. The heat maps, generated using Grad-CAM [Selvaraju *et al.*, 2016], show the important areas of the original and perturbed images when considered by both VGG16 (the authorized classifier) and ResNet18 (an unauthorized classifier). We can see that for the authorized model, the area of interest shifts, whereas for the unauthorized model, the heat map is almost negated with a very diffused area of interest.

Fig.4 visualizes the features from the final Conv2d layers in both VGG16 and ResNet18 and shows two different scenarios within the classifiers. In Fig. 4 (A) we see that the final convolution layers had very similar activations before the data was passed on to the inference layers at the end of the model. This makes sense since the authorized model correctly classified both the original and perturbed/protected images. In Fig. 4 (B), on the other hand, we see a significant difference

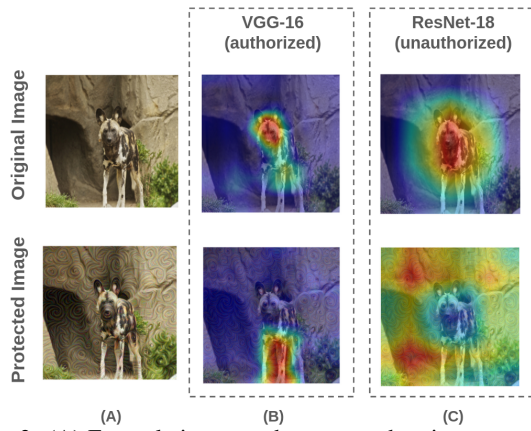


Figure 3: (A) Example image and a protected variant generated by applying FMD on VGG16. (B) Grad-CAM heat maps for these images for VGG16. (C) Grad-CAM heat maps for ResNet18.

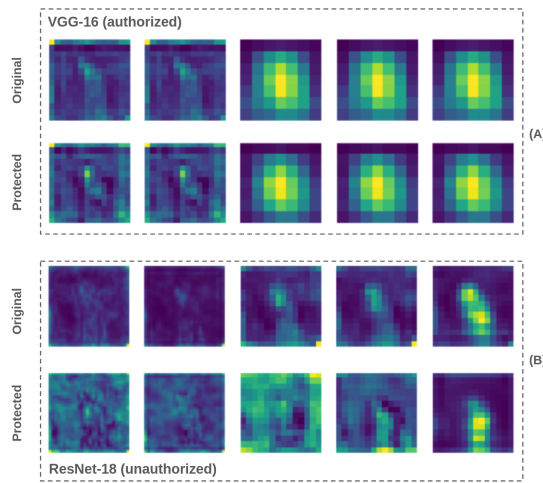


Figure 4: (A) Final convolution layers in VGG16. (B) Final convolution layers in ResNet18.

in these final feature maps, which translates into misclassifications by the unauthorized model.

Notably, these visualizations demonstrate the effectiveness of the \mathcal{J} loss in Eq. (5). The MSE loss in Eq. (5) successfully alters the original feature map as demonstrated in Fig. 3, resulting in the absence of valid input-to-output path for black box unauthorized models, as shown in Fig. 4 (B). Meanwhile, a path for the authorized model is retained, utilizing the cross-entropy (CE) loss, as demonstrated in Fig. 4 (A).

We also investigated the impact of FMD-based protection on softmax outputs of models. In Fig. 5, we observed a notable increase in the largest softmax values for a sample of 1000 images from the ImageNet dataset and their protected variants generated using FMD on the VGG16 model. For VGG16, this resulted in significant overfitting to the model. Conversely, on the unauthorized model ResNet18, we observed a decrease in maximum softmax values, particularly for incorrect classes, along with a more homogeneous distribution of overall softmax values. While these values are not confidence scores, as the models were not calibrated for confidence representation [Guo *et al.*, 2017], the indication is that

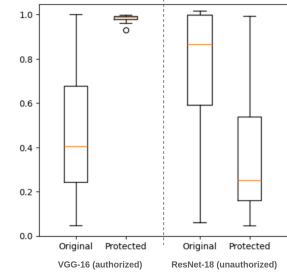


Figure 5: Highest softmax values for the authorized and unauthorized models on the original and protected images.

FMD moves protected image variants away from the decision boundary for the authorized classifier.

8 Ablation Studies

8.1 Layer Selection for FMD

In our FMD approach, we need to select a layer from the protected model to calculate $MSE(f_{feat}(x), f_{feat}(x^*))$. In our experiments, we randomly select 1000 images from the correctly classified images, and test all of the 13 convolution layers for protecting VGG16. The results in Tab. 6 show that before layer 9, all the protection accuracy numbers are equal to or higher than 99.8%. Layer 7 provides the best performance in degrading attacked models' performance. These trends are consistent across almost all the models. For the evaluation on MobileNet_v2, the performance of target model is slightly lower for layer 6, but the difference is small.

Accuracy (%)	VGG11	VGG16	ResNet18	ResNet34	W-Res50-2	Mob-v2
layer 0	52.8	99.8	58.3	68	67.4	51.3
layer 1	48.4	100	55.2	64.1	65.4	47.2
layer 2	19.8	99.8	30.6	41.1	29.3	12.9
layer 3	19	99.8	37.3	45.1	28.3	12.9
layer 4	14	99.8	36.9	45.7	29.6	15.7
layer 5	9.7	99.8	21.3	32.1	20.6	13
layer 6	8	99.8	11.9	20.8	11.6	8.9
layer 7	8	99.8	8.5	16.9	10.4	9.8
layer 8	22.2	100	27.1	38.3	26	23.3
layer 9	36.3	99.8	50.2	57.2	40.8	39.1
layer 10	31.5	97.4	49	55.7	42.7	40.9
layer 11	39.6	91.7	53.2	61	48.3	45.2
layer 12	44	64	55.8	64.3	51.8	44

Table 6: Results for layer selection when protecting VGG16

8.2 Parameter Setting for FMD

We have tested different α and N values with the best-performing layer when protecting VGG16, using the same images used in Sec. 8.1. We first fix N to 100 and test different α settings. The results in Tab. 7 show that when $\alpha = 4$, the performance is slightly better than when $\alpha = 8$. The accuracy can be affected by the random initialization described in Sec. 4, so both values of $\alpha = 4$ or $\alpha = 8$ are acceptable.

Accuracy (%)	VGG11	VGG16	ResNet18	ResNet34	W-Res50-2	Mob-v2
$\alpha = 1$	10.8	100	11.3	21.9	12.8	12.4
$\alpha = 2$	8.1	99.8	9.2	19.9	11.5	10.8
$\alpha = 4$	8	99.8	8.5	16.9	10.4	9.8
$\alpha = 8$	7.8	99.8	7.6	16.2	11.3	10.9

Table 7: Results for protecting VGG16 under different α settings.

Then, we set $\alpha = 4$ and test different N values as shown in Tab. 8. Fig. 6 is a plot of the accuracy when protecting VGG16 and considering VGG11 as the target model. When $N > 100$, the performance drop of the target model becomes slower with increasing N . Thus, in Sec. 5, we use $N = 100$ when generating the privacy-protected images.

Accuracy (%)	VGG11	VGG16	ResNet18	ResNet34	W-Res50-2	Mob-v2
$N = 10$	25.6	100	28.9	37.8	23.6	24.4
$N = 20$	14.5	100	16.8	28.3	16.2	16.1
$N = 50$	9.5	99.8	9.7	19.9	11.8	11.1
$N = 100$	8	99.8	8.5	16.9	10.4	9.8
$N = 200$	6.8	99.8	7.6	16.2	10.3	10.9

Table 8: Results for protecting VGG16 under different N settings.

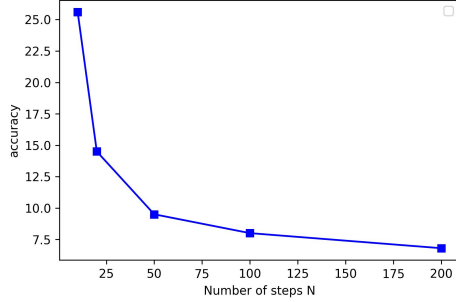


Figure 6: Accuracy when testing on VGG11 and protecting VGG16 with different N settings.

8.3 Image Quality and Protection Performance

We use a budget of ϵ to control how much generated image can differ from the original. Thus, ϵ affects the generated image quality. We perform experiments on the Celeba-HQ dataset and protect VGG16 to analyze the trade-off between the image quality and protection performance. The image quality is measured by the Structural Similarity Index Measure (SSIM) between the original and generated image. The range of SSIM is 0-1, the higher being the better. The results in Tab. 9 show that as ϵ decreases, the image quality increases, but the degradation in the performance of attacked models decreases, as expected. Example images in Fig. 7 show the visual quality when using different values of ϵ .

Accuracy (%)	VGG11	VGG16	ResNet18	ResNet34	W-Res50-2	Mob-v2	SSIM
$\epsilon = 4$	86.56	100	52.19	55.94	49.84	39.68	0.973
$\epsilon = 8$	25.31	100	13.13	18.44	17.5	9.06	0.909
$\epsilon = 16$	2.37	99.53	4.73	6.78	7.73	1.1	0.754

Table 9: Image quality results, when protecting VGG16, using different ϵ values. Image quality is measured with SSIM.

Accuracy (%)	VGG11	VGG16	ResNet18	ResNet34	W-Res50-2	Mob-v2
VGG11	100	19.23	11.54	42.31	15.38	15.38
VGG16	23.08	100	11.54	30.77	11.54	19.23
ResNet18	42.31	53.85	100	19.23	26.92	23.08
ResNet34	61.54	50	15.38	100	30.77	30.77
W-Res50-2	57.69	34.62	7.69	30.77	100	34.62
Mob-v2	73.08	30.77	15.38	57.69	50	100

Table 10: Results on the Celeba-HQ dataset using 10 classes.

8.4 Performance Analysis of Task Complexity

In this study, we reduce the task complexity by reducing the number of classes on the Celeba-HQ dataset. For the models,

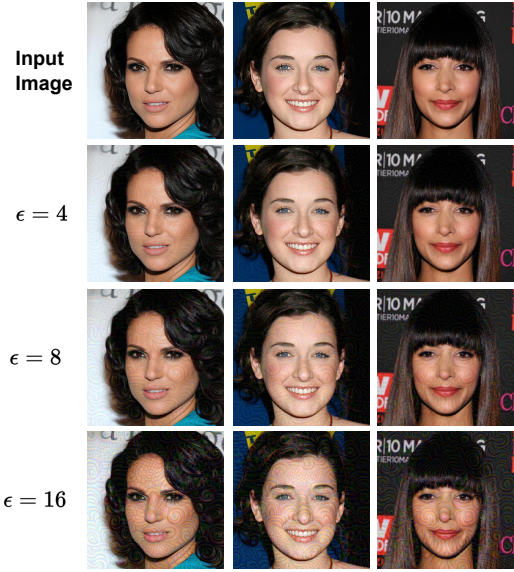


Figure 7: Example generated images for different ϵ values.

Accuracy (%)	VGG11	VGG16	ResNet18	ResNet34	W-Res50-2	Mob-v2
VGG11	100	4.08	3.06	5.1	9.18	5.1
VGG16	6.12	100	1.02	1.02	5.1	4.08
ResNet18	16.33	5.1	100	7.14	14.29	7.14
ResNet34	19.39	4.08	2.04	100	15.31	8.16
W-Res50-2	26.53	6.12	3.06	4.08	100	2.04
Mob-v2	35.71	7.14	15.31	15.31	19.39	100

Table 11: Results on Celeba-HQ dataset using 50 classes.

we only change the final layer of the MLP to match the number of classes. The results for 10 and 50 classes are shown in Tab. 10 and Tab. 11, respectively. Compared to the results from 307 classes in Tab. 2, the average performance drop on target models decreased from 93.7% to 68.46% and 90.75% for 10 classes and 50 classes, respectively. From these results, we can deduce that the effectiveness of our approach increases as the complexity of the task increases.

9 Conclusion

We have introduced a novel approach for safeguarding authorized models and deceiving unauthorized black-box models, addressing practical privacy concerns with real-world use cases. Our method excels in preserving privacy for the classification tasks involving a relatively large number of classes, achieving strong protection accuracy of 100% (ImageNet), 99.92% (Celeba-HQ), and 99.67% (AffectNet), while significantly reducing unauthorized model accuracy, on average to 11.97%, 6.63%, and 55.5%, respectively, for six different attacked models. Our in-depth evaluation includes the investigation of the underlying reasoning of the proposed FMD-based attack's efficacy, verification of a positive correlation between the number of classes and the proposed approach's effectiveness, attack transferability across tasks to unknown models (yet, such transferability is affected by the task and model characteristics), and a comprehensive ablation study. Future work includes developing an approach that can maintain effectiveness, regardless of the complexity of the tasks, with better cross-task transferability.

References

- [Carlini and Wagner, 2017] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. Ieee, 2017.
- [Chakraborty et al., 2018] Anirban Chakraborty, Manaa Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.
- [Dang et al., 2017] Hung Dang, Yue Huang, and Ee-Chien Chang. Evading classifiers by morphing in the dark. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pages 119–133, 2017.
- [De Wolf, 2020] Ralf De Wolf. Contextualizing how teens manage personal and interpersonal privacy on social media. *New media & society*, 22(6):1058–1075, 2020.
- [Deng et al., 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.
- [Dosovitskiy et al., 2014] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27, 2014.
- [Fei et al., 2020] Zixiang Fei, Erfu Yang, David Day-Uei Li, Stephen Butler, Winifred Ijomah, Xia Li, and Huiyu Zhou. Deep convolution network based emotion analysis towards mental health care. *Neurocomputing*, 388:212–227, 2020.
- [Goodfellow et al., 2014] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [Guo et al., 2017] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.
- [Guo et al., 2019] Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple black-box adversarial attacks. In *International Conference on Machine Learning*, pages 2484–2493. PMLR, 2019.
- [He et al., 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [Huang et al., 2019] Zhengbai Huang, Meng Zhang, and Yi Zhang. Toward efficient encrypted image retrieval in cloud environment. *IEEE Access*, 7:174541–174550, 2019.
- [Ilyas et al., 2018] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*, pages 2137–2146. PMLR, 2018.
- [Juuti et al., 2019] Mika Juuti, Buse Gul Atli, and N Asokan. Making targeted black-box evasion attacks effective and efficient. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, pages 83–94, 2019.
- [Kim and Kim, 2017] Phil Kim and Phil Kim. Convolutional neural network. *MATLAB deep learning: with machine learning, neural networks and artificial intelligence*, pages 121–147, 2017.
- [Ko et al., 2023] Kyoungmin Ko, SungHwan Kim, and Hyun Kwon. Multi-targeted audio adversarial example for use against speech recognition systems. *Computers & Security*, 128:103168, 2023.
- [Kwon et al., 2018] Hyun Kwon, Yongchul Kim, Ki-Woong Park, Hyunsoo Yoon, and Daeseon Choi. Friend-safe evasion attack: An adversarial example that is correctly recognized by a friendly classifier. *computers & security*, 78:380–397, 2018.
- [Kwon et al., 2019a] Hyun Kwon, Yongchul Kim, Hyunsoo Yoon, and Daeseon Choi. Selective audio adversarial example in evasion attack on speech recognition system. *IEEE Transactions on Information Forensics and Security*, 15:526–538, 2019.
- [Kwon et al., 2019b] Hyun Kwon, Hyunsoo Yoon, and Daeseon Choi. Restricted evasion attack: Generation of restricted-area adversarial example. *IEEE Access*, 7:60908–60919, 2019.
- [Kwon et al., 2022] Hyun Kwon, Changhyun Cho, and Jun Lee. Priority evasion attack: An adversarial example that considers the priority of attack on each classifier. *IEICE TRANSACTIONS on Information and Systems*, 105(11):1880–1889, 2022.
- [Kwon, 2021] Hyun Kwon. Dual-targeted textfooler attack on text classification systems. *IEEE Access*, 2021.
- [Kwon, 2023] Hyun Kwon. Toward selective adversarial attack for gait recognition systems based on deep neural network. *IEICE TRANSACTIONS on Information and Systems*, 106(2):262–266, 2023.
- [Li et al., 2021] Ning Li, Liang Cheng, Lingyong Huang, Chen Ji, Min Jing, Zhixin Duan, Jingjing Li, and Manchun Li. Framework for unknown airport detection in broad areas supported by deep learning and geographic analysis. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:6328–6338, 2021.
- [Lin et al., 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- [Lin et al., 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988. IEEE, 2017.

- [Liu et al., 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Celeba-hq: A high-quality dataset of celebrity images. *arXiv preprint arXiv:1710.10196*, 2015.
- [Liu et al., 2016] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.
- [Maho et al., 2021] Thibault Maho, Teddy Furon, and Erwan Le Merrer. Surfnet: a fast surrogate-free black-box attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10430–10439, 2021.
- [Marcel and others, 2023] Adam Marcel et al. torchvision: Datasets, transforms and models specific to computer vision. <https://pytorch.org/vision/>, 2023. Accessed: Jan 17 2024.
- [Mollahosseini et al., 2017] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.
- [Nicol Turner Lee, 2022] Caitlin Chin-Rothmann Nicol Turner Lee. Police surveillance and facial recognition: Why data privacy is imperative for communities of color, 2022. Supplied as supplemental material [tr.pdf](#).
- [Papernot et al., 2017] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.
- [Ren et al., 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 2015.
- [Sandler et al., 2018] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [Selvaraju et al., 2016] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.
- [Serengil and Ozpinar, 2020] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 innovations in intelligent systems and applications conference (ASYU)*, pages 1–5. IEEE, 2020.
- [Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.
- [Trepte, 2021] Sabine Trepte. The social media privacy model: Privacy and communication in the light of social media affordances. *Communication Theory*, 31(4):549–570, 2021.
- [Vlontzos et al., 2019] Athanasios Vlontzos, Amir Alansary, Konstantinos Kamnitsas, Daniel Rueckert, and Bernhard Kainz. Multiple landmark detection using multi-agent reinforcement learning. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 262–270. Springer, 2019.
- [Wang and Kosinski, 2018] Yilun Wang and Michal Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology*, 114(2):246, 2018.
- [Wang et al., 2020] Zijie J Wang, Robert Turko, Omar Shaikh, Haekyu Park, Nilaksh Das, Fred Hohman, Minsuk Kahng, and Duen Horng Polo Chau. Cnn explainer: learning convolutional neural networks with interactive visualization. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1396–1406, 2020.
- [Wang et al., 2021] Zi Wang, Chengcheng Li, and Xiangyang Wang. Convolutional neural network pruning with structural redundancy reduction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14913–14922, 2021.
- [Wang, 2022] Dawei Wang. Presentation in self-posted facial images can expose sexual orientation: Implications for research and privacy. *Journal of Personality and Social Psychology*, 122(5):806, 2022.
- [Xu et al., 2011] Heng Xu, Tamara Dinev, Jeff Smith, and Paul Hart. Information privacy concerns: Linking individual perceptions with institutional privacy assurances. *Journal of the Association for Information Systems*, 12(12):1, 2011.
- [Zagoruyko and Komodakis, 2016] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, 2016.