Generative Adversarial Masks: Safeguarding Images from AI Models with Stealthy Perturbations

Daniel Prakah-Asante
Massachusetts Institute of Technology
77 Massachusetts Ave, Cambridge, MA 02139
doprakah@mit.edu

Aileen Liao
Massachusetts Institute of Technology
77 Massachusetts Ave, Cambridge, MA 02139
ai0liao@mit.edu

Abstract

With the advancement of generative and representation models, protecting the intellectual property rights of artists and photographers is becoming more complex. In this project, we develop a Stealth Perturbation Adaptive Model (SPAM) that protects images from AI analysis by masking them with image-specific perturbations. These masks safeguard images from AI models that utilize a pre-trained ResNet-18 backbone while maintaining visual similarity. Our best visual model achieved an average Structural Similarity Index Measure (SSIM) of 0.99 and a cosine similarity of 0.763. In addition, our experiments reveal that the masks generated by our model can reduce a binary classifier's accuracy from 99% to 48.97% during inference. Furthermore, incorporating protected images into a classifier's training data reduced its accuracy to 81.44%. This work highlights the potential for generative adversarial models to balance visual similarity and image protection.

1. Introduction

AI tools have advanced to the point where they can accurately reproduce the styles of renowned artists, which calls into question the value of art created by humans. These technologies use the works of artists and photographers to train extensive models without offering compensation. In addition, the emergence of AI deepfake models has raised national security concerns due to their potential to misrepresent public figures. With continuous improvements in generative image model and increasingly complex representation networks, image ownership and privacy issues will continue to intensify.

In this work, we devise a mechanism to protect images from analysis by pre-trained AI models. Our models will generate specific ”stealthy” masks. When applied as in Figure 1, these masks significantly alter how images are interpreted by targeted AI models, decreasing the stability and accuracy of the model during inference and training. Additionally, we have incorporated adjustable parameters into the model to accommodate various use cases, prioritizing image quality in some scenarios and enhanced protection in others. If artists and photographers can better protect their work, compensation negotiations will improve, privacy laws can be upheld, and we can reduce the possibility of deepfake threats to national security.

2. Related Works

Szegedy et al [10] found that applying hardly perceptible perturbations from maximizing network prediction error cause models to misclassify images. Existing methods, such as that of Salman et al [7], explored perturbing the input image with a perturbation calculated by a projected gradient descent optimization problem. While perturbing the image was effective, the projected gradient approach was expensive and slow.

Generative Networks have also been explored. Poursaeed et al. [5] have utilized Generative Networks to produce general and image-specific perturbations. While these models successfully deceived existing pre-trained models, the application of perturbations resulted in noticeable visual alterations. They employed a ”fool loss” strategy aiming to confuse the target classification models rather than directly attacking the representation. Different methods to attack various parts of target networks are described in Sun et al. [9]

Other approaches focus on sabotaging the pre-trained model's inference capabilities by intentionally mismatching image and text pairs, as demonstrated in the work of Shan et al [8]. This strategy aimed to exploit the model dependencies on coherent image-text relationships, impairing their ability to accurately interpret and classify images.

Our approach improves on existing perturbation mask methods by utilizing cosine similarity as the main loss to make our target representations dissimilar. Additionally, we designed a variety of secondary losses to better maintain image quality.

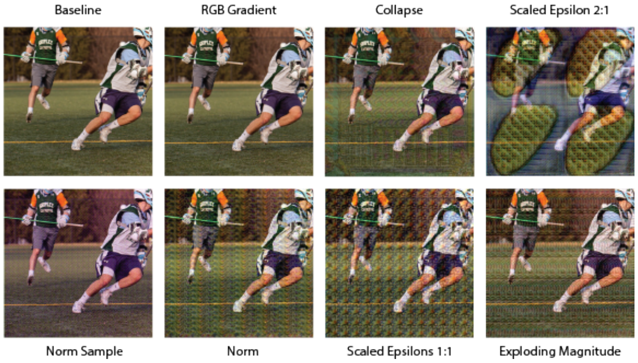

Figure 1. SPAM protected images. This provides an example of masked images created with different loss approaches.

3. Methods

3.1. Approach

Building on Poursaeed's research, we have developed Stealth Perturbation Adaptive Models (SPAM). We designed this model to generate "stealthy" image-specific protection masks that produce incorrect representations in the targeted pre-trained model. This operation can be defined as follows:

$$\text{x'} = x + f(x;\theta)$$

where f is a UNet model [6] with parameters θ, x is the input image, and x' is the protected image. During training, the representations of the original image x and the protected image x' are obtained from a frozen ResNet-18 model [3]. The goal is to minimize the cosine similarity loss between these representations to maximize the dissimilarity of the protected image's representation. This operation is defined as follows:

$$\text{r} = \text{Rep}(x;\theta)$$

$$\text{r'} = \text{Rep}(x';\theta)$$

$$\text{Loss} = \left(\frac{\mathbf{r}\cdot\mathbf{r}'}{\|\mathbf{r}\|\|\mathbf{r}'\|}\right)$$

where Rep is the target representation network, ResNet-18.

At the end of our SPAM model the mask scaled as follows:

$$\text{mask} = \left(\frac{\text{M} * \epsilon}{\|M\|_\infty}\right)$$

where M is the output of the UNET and ϵ is a parameter we control. During training, we set an adaptive value to maintain a balance between ϵ and Cosine Similarity loss. This approach prevents the model from learning masks that are either too strong, which would significantly degrade the image quality, or too weak to offer effective protection. It also provides control depending on the trade-offs one is willing to make for their task. In addition, we explored utilizing secondary losses to continue to improve image quality while still providing effective protection. To effectively minimize both losses, we applied PC grad [11] and loss weighting to prevent one loss from converging too quickly in a way that impedes the reduction of the other loss.

3.2. Datasets

In this study, we used the Flickr30k dataset [4] and processed all input images by cropping them to a uniform size of 224x224 pixels and standardizing pixel values. We discarded any images that could not be cropped due to format issues. We then divided the dataset into three subsets: 80% for training, 10% for validation, and 10% for testing. Additionally, we conducted further testing of our model on the Kaggle Dogs vs Cats Competition dataset [2]. For this test, we followed the same preprocessing steps but resized the images to 256x256 pixels before cropping them to 224x224.

3.3. Models

Table 1. Approaches

Approach	Description
Scaled Epsilons	During training, change ϵ to maintain a set ratio between the ϵ and Cosine Similarity loss
Norm Models	Minimize an additional mask norm to encourage higher image quality
Secondary Loss	Additional gradient or magnitude loss
Alternative Loss	Maximize cosine similarity on to chosen target vector to achieve dissimilarity

All of our approaches listed in Table 1 utilized a modified U-Net model shown in Figure 2 with 6 down sampling and 5 up sampling convolutional layers with residual skips in between. Due to computational constraints, we reduced the number of filters in the earlier layers and increased the depth of the model compared to a traditional U-Net structure. Additionally, to transform this into a generative model, we injected random uniform noise at the bottleneck stage by concatenating 32 channels of noise. This modification aims to enhance the model's ability to generate more diverse outputs.

Scaled Epsilons Models

Scaled Epsilon 1:1 During training we adjusted ϵ to maintain a 1:1 ratio between the ϵ and loss

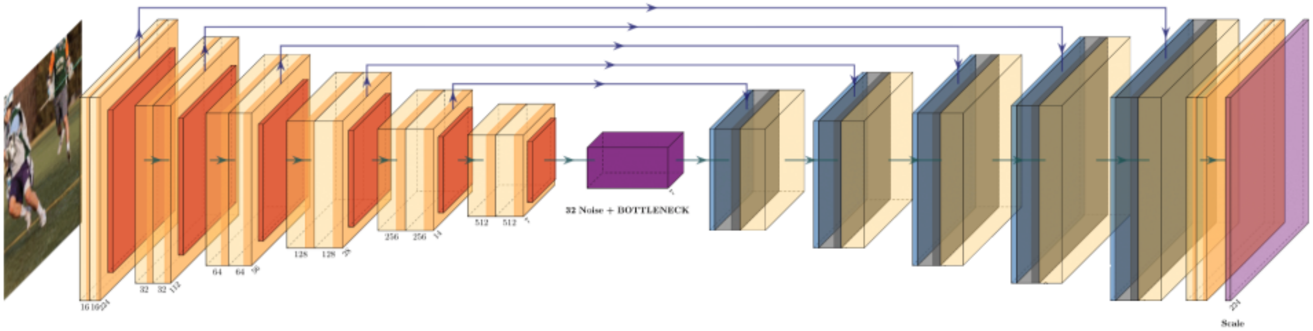

Figure 2. Stealth Perturbation Adaptive Model. It includes a fully Convolutional U-Net model with 6 down sampling and 5 up sampling layers.

Scaled Epsilon 2:1 During training we adjusted ϵ to maintain a 2:1 ratio between the ϵ and loss for more aggressive protection.

Norm Models

Norm We used the same model as Scaled Epsilon 1:1 but minimized an additional $\mathbf{L}^{\infty}$ norm to encourage better image quality

Norm Sample We used the same model as Norm but modified the cosine similarity loss function as follows:

$$\text{s} = \text{RandomBinary}(n) = \{b_i \mid b_i \in \{0,1\}, \text{ for } i = 1 \text{ to } n\}$$

$$\text{Loss} = \text{Cosine Similarity}(r \cdot s, r' \cdot s)$$

where n is the length of the representation vectors. The random vector s acts as a regularizer to encourage the model to minimize cosine similarity holistically.

Secondary Loss Models

RGB Gradients We used the same model as Scaled Epsilon 1:1 but minimize an additional gradient loss defined as:

$$\text{Gradient Loss} = dX + dY + dC_1 + dC_2 + dC_3$$

$dX =$ horizontal spatial gradients loss
$dY =$ vertical spatial gradients loss
$dC_1 =$ the color gradient loss between channel pairs 0 and 2
$dC_2 =$ the color gradient loss between channel pairs 1 and 2
$dC_3 =$ the color gradient loss between channel pairs 0 and 1

This additional loss component encourages the model to better preserve the visual integrity of the image while ensuring effective perturbation.

Exploding Magnitude We used the same model as Scaled Epsilon 1:1 but minimized an additional magnitude loss defined as:

$$\text{Magnitude Loss} = -1 * (\|r'\|_1 - \|r\|_1)$$

This encourages the model to create a representation that has a larger magnitude than the initial representation.

Alternative Loss Model

Collapse We used the same model as Scaled Epsilon 1:1 but but modified the cosine similarity loss function to use as follows:

$$\text{Loss} = 1 - \text{Cosine Similarity}(target, r')$$

Where the target is a vector that we define, we use a vector of negative ones of the same length as the representation. This shifts the objective from minimizing dissimilarity to collapsing all representations into one uniform vector. This encourages all representations from protected images to converge to the same vector.

3.4. Evaluations

We used SSIM (Structural Similarity Index) [1] as the main metric to evaluate image quality, SSIM measures the similarity between two images in terms of their structural information, and we used Cosine Similarity to evaluate the dissimilarity of our representations. In addition, we trained a classifier using the Kaggle Dogs vs Cats dataset, applying various distributions of our SPAM protected image data during training and inference. This approach demonstrated how effective our protective images are against ResNet-18 classifiers, even when our model has not been exposed to the training data.

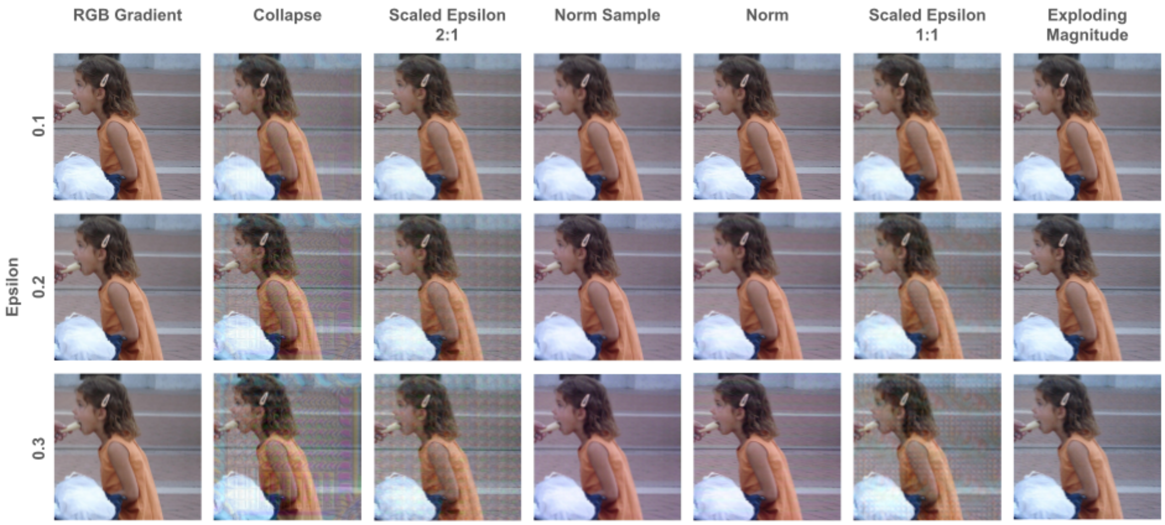

Figure 3. Approach with Varying Epsilons. Different loss approaches were evaluated for varying epsilons. The RGB gradient showed few perceptible changes over increasing epsilons, while other methods created visually perceptible masks over images.

Table 2. Average Image Quality and Cosine similarity of SPAM models

Model	SSIM	Cos Similarity	UIQ	RMSE
ScaleEp 1:1	0.50	0.70	0.91	0.12
ScaleEp 2:1	0.37	**0.63**	0.82	0.16
Norm	0.63	0.70	0.94	0.08
Norm Sample	0.69	0.76	0.94	0.08
Exploding	0.51	0.65	0.91	0.10
RGB-Gradients	**0.99**	0.74	**0.99**	**0.00**
Collaspse	0.77	0.79	0.96	0.05

4. Results

4.1. Image Quality and Cosine Similarity

To evaluate SPAM's protection, we compared the cosine similarity of the protected images to the original images in the test set shown in Table 2. We found that Scaled Epsilon 2:1 offered the most protection, achieving the lowest average cosine similarity score of of 0.63. The rest of the models maintained a cosine similarity level of around 0.70. The RGB gradient model was the most robust in preserving image quality while producing competitive cosine similarity scores. It had an average SSIM of 0.99, indicating that a person cannot infer the difference between the original and protected images.

We also investigated the distribution of scores of the protected images. Models prioritizing image similarity exhibited minimal variations in cosine similarity, whereas more aggressive models showed extensive ranges. In these aggressive models, some images approached a cosine similarity near zero. Furthermore, our evaluation of different epsilon values in Figure 3 indicated that large deviations significantly reduced image quality without substantially affecting the cosine similarity. This result suggests that a much lower epsilon might be adequate for particular applications.

4.2. Dogs and Cats

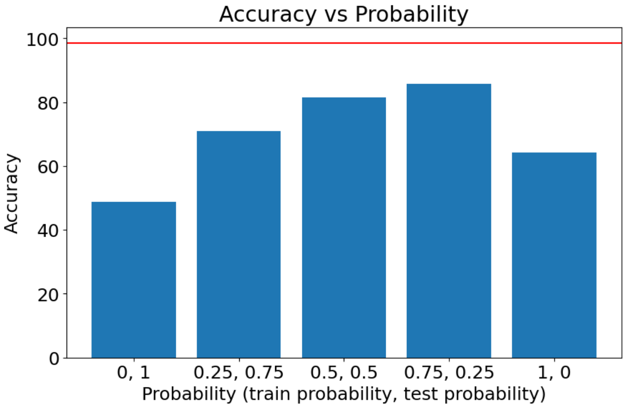

Figure 4. Accuracy vs Probability. This graph depicts the model accuracy for different probabilities of using a protected image during training and testing. The red line shows the 98.6% accuracy that the model originally achieved.

To test the effectiveness of our model in protecting images, we trained a binary classifier that utilizes a ResNet-18 representation to classify cat and dog images. After training

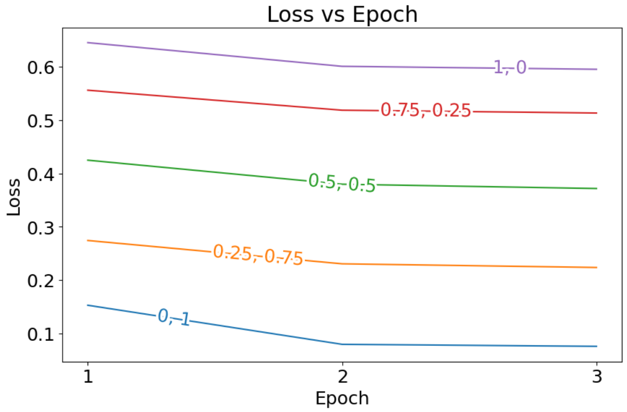

Figure 5. Loss vs Epoch for Different Probabilities. The losses for each set of train and test probabilities was graphed over three epochs.

for three epochs, this model achieved an accuracy of 98.6% in distinguishing between dogs and cats. However, when we tested the same model on images protected by any of our SPAM models, the accuracy dropped to less than 48%. This significant decrease in accuracy indicates that the classifier's parameters are ineffective when applied to the protected images.

We conducted experiments by incorporating SPAM-protected images into the training data. As we exposed the classifier to more SPAM-protected images, its performance improved. However, it did not achieve its initial accuracy, as depicted in Figure 4. This result indicates that SPAM protection results in image representations that are more challenging for the model to learn to separate. This effect is also shown in Figure 5, which illustrates that as the percentage of protected images increases, the loss converges at a higher value.

Notably, the RGB gradient model failed to deceive the classifier when trained with its own protected images. The minimally invasive nature of the RGB gradient masks may allow for the classifier to learn to ignore its perturbations. This evaluation underscores various trade-offs depending on the use case that can be controlled with ϵ. For example, photo companies might prefer more aggressive watermarks that fit their workflows, while individual users might choose lighter masks to safeguard their photos from inference.

5. Discussion

By developing mechanisms to protect images, we can address the escalating concerns surrounding image ownership and privacy. These protection mechanisms will enable more equitable compensation, ensure compliance with privacy regulations, and reduce the risks associated with deepfake threats to public trust. Our work advocates for the development of more dynamic protective measures, paving the way for safer interaction with AI technologies.

Our method may be able to scale to protect against larger models and multiple networks simultaneously. In future research, we intend to evaluate our methods on leading architectures and more robust representations, such as CLIP. Additionally, we are interested in applying diffusion techniques to generate image masks. These methods may work to create more robust image protection models.

6. Contributions

Daniel: Pre-processed image data. Designed the Norm, Secondary Loss, and Alternative Loss model approaches.

Aileen: Completed the scaled epsilon studies. Ran and analyzed the model evaluation on the Dogs vs Cats dataset.

References

[1] Luca Abeni, Csaba Kiraly, and Renato Lo Cigno. Sssim: a simple and scalable simulator for p2p streaming systems. In *2009 IEEE 14th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks*, pages 1–6, 2009. 3

[2] Will Cukierski. Dogs vs. cats, 2013. 2

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 2

[4] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, 2016. 2

[5] Omid Poursaeed, Inna Katsman, Brian Gao, and Serge Belongie. Generative adversarial perturbations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. https://doi.org/10.1109/cvpr.2018.00465. 1

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 2

[7] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing, 2023. 1

[8] Sharan Shan, Wei Ding, Joseph Passananti, Hao Zheng, and Benjamin Y. Zhao. Prompt-specific poisoning attacks on text-to-image generative models. *arXiv*, n.d. https://doi.org/10.48550/arxiv.2310.13828. 1

[9] Hui Sun, Tianqing Zhu, Zhiqiu Zhang, Dawei Jin, Ping Xiong, and Wanlei Zhou. Adversarial attacks against deep generative models on data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35:3367–3388, 2021. 1

[10] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014. 1

[11] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning, 2020. 2