
11-785 Team 15 Final Report: Adversarial Attack Against AI Art Detectors

Prachi Patil

pspatil@andrew.cmu.edu

Sukriti Rawal

sukritir@andrew.cmu.edu

Prith Sharma

priths@andrew.cmu.edu

Kanupriya Anand

kanupria@andrew.cmu.edu

Abstract

This study explores techniques for distinguishing AI-generated art from human-created pieces using deep learning. We focus on developing a lightweight and explainable model by evaluating various combinations of feature extraction and model architectures on real-world artistic datasets as well as AI-generated art produced by common generative models. We investigate the robustness of a state-of-the-art CLIP-based SVM classifier to adversarial perturbations and introduce a novel patch-based projected gradient descent (PGD) attack tailored to the CLIP architecture. We assemble a diverse dataset comprising 3,633 real paintings alongside AI-synthesized images based on descriptions of the real art. While conventional PGD effectively degrades detector accuracy, the resulting noise is often conspicuous. To overcome this, we propose two targeted strategies: (1) delta-based patch selection, which concentrates perturbations on the highest-variance pixel regions, and (2) attention-driven patch extraction, which leverages CLIP’s final attention layer to pinpoint semantically critical areas. Our experiments demonstrate that manipulating as little as 10% of an image’s pixels via patch-based PGD yields fool rates on par with those of global PGD, yet with markedly less perceptual distortion. We achieve a 67% decrease in discriminator accuracy.

1 Overview and Context

1.1 Motivation

AI models can now produce images that closely mimic traditional artistic styles which raises questions about authenticity and artistic intent. As these technologies continue to improve, it is crucial to continue developing new approaches that can reliably differentiate between real and AI-generated artwork. Given the creative and intangible nature of art, the value of artwork is often related to whether or not a human made it and the context it was created in. Thus, it becomes essential to identify what differentiates an AI-generated artwork from real artwork. By studying adversarial attacks, we can identify weaknesses in current approaches and work toward more robust, trustworthy models—ultimately helping preserve the authenticity of human-created art.

1.2 Objectives

- Compare AI-generated and human-made art to identify distinguishing features and patterns.
- Evaluate different encoders and feature selection techniques to optimize model performance in detecting AI-generated art.
- Investigate and analyze unique artifacts present in AI-generated images using explainability methods such as Grad-CAM and saliency maps.
- that Develop a model generates imperceptible adversarial perturbations in the pixel space of AI-generated art images so that detection models misclassify them.

2 Related Work

AI-Image Detection: Lađević et al. [2024] demonstrates CNNs’ effectiveness in detecting AI-generated satellite images, with Grad-CAM++ providing insights into classification decisions. Applying this method to artistic datasets could uncover unique artifacts in AI-generated art. Meanwhile, Cozzolino et al. [2024b] highlights zero-shot learning techniques that leverage statistical properties like coding cost error for scalable detection without additional training data. Finally, Baraheem and Nguyen [2023] explores EfficientNetB4 on a comprehensive GAN-based dataset, revealing imperceptible artifacts in synthetic images and using explainability tools like CAMs and LIME to identify key features contributing to detection decisions.

Mehta and Rastegari [2021] combines CNNs with vision transformers (ViTs) to improve computational efficiency in image recognition, making it suitable for lightweight, on-device AI applications. This approach can be explored to determine if transformer-based architectures enhance the detection of AI-generated images while maintaining efficiency. Similarly, Cozzolino et al. [2024a] utilizes OpenAI’s CLIP model to extract feature vectors for images, which are then classified using an SVM. Their findings suggest that downsampling generated images reduces frequency artifacts, offering a pathway to evaluate CLIP’s generalization capabilities across diverse AI-generated content.

Adversarial Attacks against Vision Models: Diao et al. [2024] examine the vulnerabilities of AI-generated image (AIGI) detectors, demonstrating how adversarial perturbations in the frequency domain can deceive state-of-the-art CNN and ViT-based detection models. Zhang et al. [2023] focus on adversarial attacks in the artistic image domain, proposing a two-stage framework that perturbs artistic styles while preserving visual integrity, effectively misleading artistic style classifiers. Guo et al. [2024] address the protection of artwork from unauthorized neural style transfer (NST) through locally adaptive adversarial color attacks, which subtly alter images to degrade NST output quality while maintaining perceptual similarity to the original artwork.

Wang et al. [2021] proposes a stealthy adversarial attack method that combines patch-based and noise-based perturbations to fool object detection models like Faster R-CNN and YOLOv4. By first applying a PGD (Projected Gradient Descent) (Madry et al. [2017]) noise attack to identify effective regions, and then optimizing a patch within those regions, the method significantly disrupts detection performance while keeping visual changes nearly imperceptible. The approach is fast enough for potential real-time use and achieves high attack success in a white-box digital setting. However, it currently lacks support for physical-world or targeted attacks, which the authors identify as future directions. We also borrow ideas from Zhang et al. [2018] show that measuring Euclidean distances between images in the internal feature spaces of deep convolutional networks yields a perceptual metric that correlates far more closely with human judgments than traditional measures like PSNR or SSIM. By optimizing image synthesis and transformation models to minimize these deep-feature distances, one can effectively reduce visible artifacts and ensure the outputs are perceptually indistinguishable from the originals.

3 Methodology

3.1 Dataset

We start with an existing real art dataset and use a mix of generators for the synthetic image creation. Existing art-specific AI detection datasets are limited. In order to create a more comprehensive art

dataset, we refer to the method used by Cozzolino et al. [2024a] of using captioned real images in order to generate fake ones.

We started the dataset creation process using ArtCap Lu et al. [2022] which is a real art dataset that contains images and corresponding labels in the form of 5 sentences that describe the image. ArtCap contains 3606 paintings in the impressionist or expressionist style. For each real art image, we create AI generated art using the descriptions to prompt different text-to-image models. We have used the model provided by the following to generate our data: Stable Diffusion (Rombach et al. [2022]), OpenJourney (OpenJourney [2024]) and seek.art MEGA (Diffusion [2024]) from HuggingFace; Dall-E 3 (OpenAI [2024]); and AWS Titan (AWS [2024]). We generated a total of 12,706 images of resolution 512x512 with 3 synthetic images (on average) corresponding to each real art image.

3.2 Discriminator Description

For the baseline discriminator, we refer to Cozzolino et al. [2024a]. The model is comprised of two components: a CLIP image encoder for feature extraction and an SVM for binary classification.

- **Feature Extraction (CLIP):** We feed each input image into CLIP’s pre-trained image encoder. The encoder outputs a high-dimensional embedding (feature vector) capturing both semantic and stylistic aspects of the image.
- **Classification (SVM):** We train a Support Vector Machine (SVM) on these CLIP embeddings. The SVM predicts whether an artwork image is AI-generated or real.

3.3 Patch-Based Projected Gradient Descent Noise Injection

3.3.1 PGD General Description

We use Projected Gradient Descent (PGD) as the core mechanism for introducing adversarial noise into the input image. PGD is an iterative attack that gradually adds perturbations in the direction that maximizes model loss.

At each iteration, PGD performs the following:

- Computes the gradient of the loss with respect to the input image.
- Applies a small perturbation in the direction of the gradient, scaled by a fixed step size α .
- Projects the updated image back into an ℓ_∞ ball of radius ϵ centered at the original input, ensuring the perturbation remains bounded and imperceptible.

This process is repeated for T steps to gradually push the image across the model’s decision boundary. While classical PGD perturbs the entire image, we modify the procedure to constrain updates to only a subset of critical patches, selected dynamically at each step. This enhances attack efficiency and limits visible distortion.

3.3.2 Adversarial Attack: Patch-Based PGD

To evaluate the robustness of our baseline, we implement a **patch-based Projected Gradient Descent (PGD) adversarial attack** that targets the CLIP+SVM pipeline. The attack aims to generate visually indistinguishable adversarial examples of AI-generated images that are misclassified as Human-created. Instead of perturbing the entire image, the attack modifies only the most semantically important patches, making it both efficient and stealthy.

The attack pipeline is as follows:

Iterative PGD Attack Procedure: We implement an iterative patch-based Projected Gradient Descent (PGD) attack, where perturbations are applied only to semantically important image regions based on gradient information. This approach improves stealth and efficiency while maintaining high fooling success. The procedure is described in Algorithm 1.

Algorithm 1: Patch-Based PGD Attack on CLIP+SVM

Input: Input image x , patch size p , step size α , max perturbation ϵ , number of iterations T , number of patches k

Output: Adversarial image x_{adv}

```
1 Divide  $x$  into a grid of patches of size  $p \times p$ ;  
2 Initialize  $x_{\text{adv}} \leftarrow x$  (optionally with random noise within the  $\epsilon$ -ball);  
3 for  $t \leftarrow 1$  to  $T$  do  
4   Compute embedding  $z \leftarrow \text{CLIP}(x_{\text{adv}})$ ;  
5   Compute SVM loss  $\mathcal{L}(z)$  based on decision boundary distance;  
6   Compute input gradient  $g \leftarrow \nabla_x \mathcal{L}$ ;  
7    $M \leftarrow \text{create\_patch\_mask}(g, p, k)$ ; // Select top- $k$  high-gradient patches  
8   Update:  $x_{\text{adv}} \leftarrow x_{\text{adv}} - \alpha \cdot (M \odot \text{sign}(g))$ ;  
9   Project:  $x_{\text{adv}} \leftarrow \text{clip}(x_{\text{adv}}, x - \epsilon, x + \epsilon)$ ;  
10 return  $x_{\text{adv}}$ 
```

Attention-Based Patch Selection: In addition to selecting patches based on top- k gradient magnitudes from the SVM loss, we implement a second patch selection strategy that leverages CLIP’s attention mechanism. Specifically, we extract the most attended patches from the final attention layer of CLIP’s vision transformer.

Within CLIP’s architecture, the first token serves as a global aggregation token, attending to all image patches. By identifying which tokens (i.e., image patches) receive the highest attention weights from this class token, we can infer which regions are most semantically significant. We select the top- k attended patches as our fixed perturbation targets.

Once selected, we run the same iterative PGD algorithm as described earlier. However, unlike the gradient-based method where the patch mask is updated at every step, this attention-based strategy uses a fixed mask throughout all iterations.

Additional Features:

- Input normalization is preserved for compatibility with CLIP.
- Optionally, attacks can begin with a random initialization inside the ϵ -ball.
- Visualization mode highlights selected patch regions, aiding explainability.

This patch-based attack takes advantage of the fact that the model relies heavily on just a few important parts of the image. By making small, targeted changes to those key regions, guided by gradient information we can trick the CLIP+SVM model into making wrong predictions, without noticeably changing how the image looks.

3.4 Evaluation Metric

We use Accuracy, Precision, Recall, F1 Score and mAP to measure detection performance. AUC is widely adopted in the related works above and in other binary classification tasks, so we also measured it. Given the potential label imbalance in our dataset (multiple AI-generated images per real one), AUC and F1 are especially important. For adversarial attacks, we will additionally track Attack Success Rate (Fool Rate)—the proportion of AI images misclassified as real.

4 Baseline and Extensions

4.1 Baseline Selection and Evaluation

We utilized zero-shot detection using the pre-trained model (CLIP + SVM) outlined by Cozzolino et al. [2024b] to establish baseline performance in both our train and test sets. Since the CLIP model was not pre-trained on the data at all, it serves as a valid basis for comparison. This is a good baseline model to use because it is naturally lightweight and simple, as it focuses on statistical properties of the pixel composition. There is no existing true baseline for AI-art detection, so we use this as a starting point for exploring model architectures and feature extraction strategies.

We collected real and AI-generated images, labeled them (0 for fake, 1 for real), and stored their file paths along with metadata. Each image was passed through a pre-trained CLIP model (openai/clip-vit-base-patch32) to obtain a 768-dimensional hidden-state embedding, which we normalized and recorded. These embeddings were then split into training and test sets, and a linear SVM was trained on the CLIP features. Finally, we evaluated the trained SVM under various conditions such as excluding a specific generator from the training set—to test its generalization to unseen AI image styles.

Images generated by various generators are visualized in Figure 1. Moreover, as seen in Figure 2, some generated images are clearly visually distinguishable from real artwork due to odd, unrealistic formations, making them easier to detect.



Figure 1: Real and AI generated images from Artcap prompt: "a woman with a parasol sitting on rocks with hills and a river in the background, a woman in a white dress is sitting near the water with an umbrella, this image shows a girl sitting holding an umbrella, a faceless woman is sitting on the ground by a water source"



Figure 2: Examples of AI generated images with defects

4.2 Attack Baseline + Extensions

To test the robustness of our CLIP+SVM discriminator, we first implemented a baseline adversarial attack using classic PGD. This approach perturbs every pixel of the image iteratively to reduce the SVM’s confidence, significantly lowering classification accuracy and confirming the vulnerability of the detector.

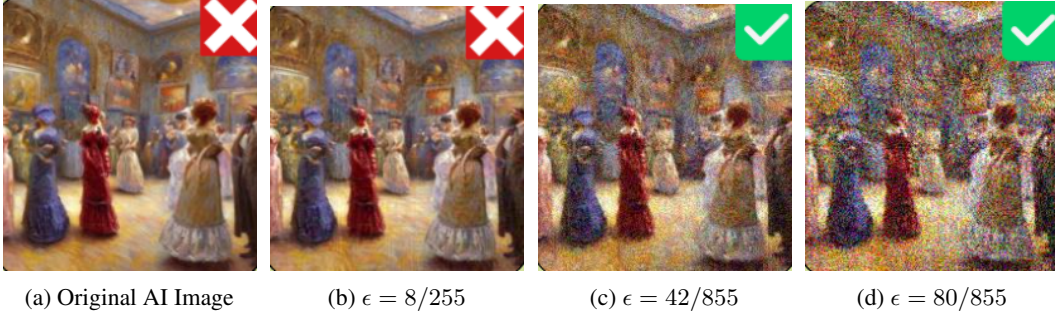


Figure 3: Classic PGD Noised Image with Increasing Epsilon. A green checkmark indicates that the image successfully fooled our baseline detector

As seen in Figure 3, classic PGD modifies the full image, which may be unrealistic in scenarios where minimal changes are needed for stealth. As the epsilon (the allowed noise distance) increases, it becomes clear that the image was tampered with, especially at the epsilon magnitude necessary to fool the detector (the images marked with a green checkmark).

To address this, we extend the attack strategy by developing a patch-based PGD variant. This approach restricts perturbations to only the most influential image regions, either selected via gradient-based saliency or attention weights from the CLIP model. Our patch-based PGD achieves comparable degradation in detector accuracy while perturbing fewer pixels, demonstrating a more efficient and stealthy attack. We implement the two patch selection strategies outlined in Section 3: Maximum SVM gradient (max delta) patch selection and CLIP attention layer-based selection. As displayed in Figure 4, these perturbations are almost unnoticeable, making the attack much more pernicious.

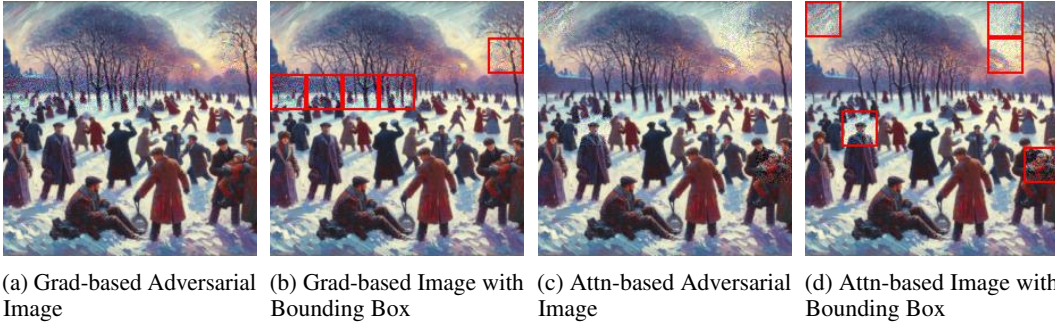


Figure 4: Adversarial Images with our different patch-based attention strategies

5 Results and Analysis

5.1 Baseline Discriminator Results

We present our baseline evaluation results obtained on the combined AI-generated set as well as with leave-one-out testing as described above in table 1. The baseline performs well with an accuracy of 99.75% on the binary classification task. This provides a strong motivation for adversarial attack in order to ensure robustness against such attacks.

However, the leave-one-out testing reveals important variations in model generalization. While the classifier maintains excellent performance when tested on Open Journey, Titan, and Stable Diffusion

Model	Accuracy	Precision	Recall	F1 Score	mAP
Overall Testing	0.9975	0.9975	0.9975	0.9975	0.7849
Leave-One-Out Testing					
Open Journey	0.9932	1.0000	0.9932	0.9966	1.0000
Stable Diffusion	0.7399	1.0000	0.7399	0.8505	1.0000
Titan	0.9922	1.0000	0.9922	0.9961	1.0000
DALL·E 3	0.8644	1.0000	0.8644	0.9273	1.0000
seek.art MEGA	0.9968	1.0000	0.9968	0.9984	1.0000

Table 1: Baseline performance metrics on datasets generated by different models

pipeline images (all above 99% accuracy), it struggles significantly with Stable Diffusion (73.99% accuracy) and DALL·E 3 (86.44% accuracy) when these generators are excluded from training. This discrepancy highlights potential overfitting to specific generator characteristics and suggests that Stable Diffusion and DALL·E 3 produce images with distinctive features that differ from other generators. These findings underscore the need for adversarial testing to ensure robust detection across diverse AI image generation systems.

5.2 Adversarial Attack Results

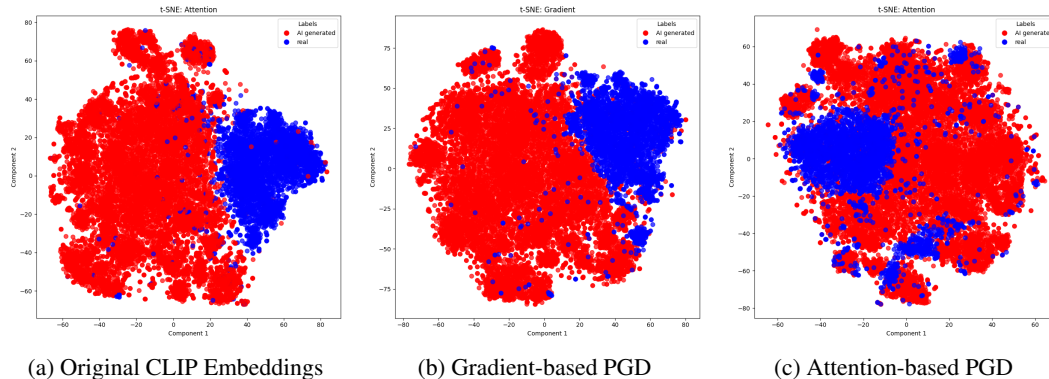


Figure 5: Comparison of t-SNE embedding visualizations: (a) Original CLIP embeddings, (b) embeddings after gradient-based PGD perturbation, and (c) embeddings after attention-based PGD perturbation.

We next evaluate how well adversarial perturbations can fool the CLIP+SVM detector. Table 2 shows performance under two attack settings: Classic PGD (full image noise) and our proposed Patch-Based PGD. Both attacks significantly degrade classifier performance compared to the original 99.97% accuracy.

Classic PGD reduces accuracy to 33.33%, completely bypassing the detector. Impressively, our preliminary patch-based PGD achieves similar degradation (36.66% accuracy) while modifying only a small fraction (about 10%) of image pixels. This suggests that selective, semantically informed patch perturbations are both efficient and effective.

Metric	Detector Baseline	Classic PGD	SVM Gradient Patch-Based PGD	CLIP Attention Patch-Based PGD
Accuracy	99.97%	18.54%	86.39%	43.42%
Precision	99.97%	4.29%	88.21%	73.38%
Recall	99.97%	18.54%	86.39%	43.42%
AUC	0.9998	0.0110	0.936	0.6351
Fool Rate		100%	12.71%	68.02%

Table 2: Discriminator performance under adversarial attacks.

5.3 Discussion

These early results validate the vulnerability of the CLIP+SVM model to adversarial perturbations. Classic PGD had an extremely strong effect, lowering the accuracy to less than 20% and resulted in a 100% fool rate. This is expected given the noise application to the full image. The gradient based patch-selection resulted in a much more modest reduction in detector accuracy, achieving a fool rate of only 13%. However, the attention-based patch selection strategy was quite successful, achieving a fool rate of over 60% and lowering accuracy to the mid-40s. These results are very promising for our strategy which was able to introduce strategic noise in a much smaller part of the image compared to the classic, full PGD attack. Since PGD occurs iteratively over steps and not epochs, training/val curves do not apply and are not included.

In the original t-SNE projection (Figure 5a), the CLIP embeddings of AI-generated versus human-created art form two well-delineated clusters, reflecting that the frozen SVM discriminator can easily distinguish between the two domains. After applying gradient-based PGD perturbations (Figure 5b), however, these clusters begin to overlap: a substantial portion of the AI-generated points drifts toward the real-art region, reducing the margin between the classes. When we instead use attention-guided PGD (Figure 5c), this intermixing becomes even more pronounced. Many AI-generated embeddings now fall within the core of the real-art cluster. Attention-based noise is more effective at fooling the CLIP+SVM pipeline, since it induces feature shifts that more closely mimic the distribution of genuine artwork. This progressive degradation of cluster separability, from the clean embeddings through gradient PGD and into the attention-perturbed case, highlights both the power of targeted adversarial perturbations and the need for adaptive or more robust discriminators.

6 Future Directions

- Instead of freezing the discriminator, create a traditional GAN pipeline where the discriminator weights are updated to ensure the fool rate does not reach higher than a specific threshold
- Experiment with targeting specific image channels based on their importance to the detector and tune the epsilon (the range which the noise as to stay in by each channel, based on human sensitivity/perception to color as well
- Identify which image artifacts are used as the basis for detection of AI-generated art. Beyond just fooling the discriminator, it is valuable to know why a model thinks something is AI-generated. Interpretability tools (e.g., Grad-CAM) on a differentiable proxy for the SVM could be used to highlight regions or artifacts that drive the AI classification.

References

- AWS. Amazon titan models on bedrock, 2024. URL <https://aws.amazon.com/bedrock/amazon-models/titan/>. Accessed: 2025-03-17.
- S. S. Baraheem and T. V. Nguyen. Ai vs. ai: Can ai detect ai-generated images? *Journal of Imaging*, 9(10):199, 2023.
- D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, and L. Verdoliva. Raising the bar of ai-generated image detection with clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4356–4366, 2024a.
- D. Cozzolino, G. Poggi, M. Nießner, and L. Verdoliva. Zero-shot detection of ai-generated images. In *European Conference on Computer Vision*, pages 54–72. Springer Nature Switzerland, 2024b.
- Yunfeng Diao, Naixin Zhai, Changtao Miao, Xun Yang, and Meng Wang. Vulnerabilities in ai-generated image detection: The challenge of adversarial attacks. *arXiv preprint arXiv:2407.20836*, 2024.
- Stable Diffusion. Stable diffusion pipeline, 2024. URL https://huggingface.co/coreco/seek_art_MEGA. Accessed: 2025-03-17.

- Zhongliang Guo, Junhao Dong, Yifei Qian, Kaixuan Wang, Weiye Li, Ziheng Guo, Yuheng Wang, Yanli Li, Ognjen Arandjelović, and Lei Fang. Artwork protection against neural style transfer using locally adaptive adversarial color attack. In *ECAI 2024*, pages 1414–1421. IOS Press, 2024.
- A. L. Lađević, T. Kramberger, R. Kramberger, and D. Vlahek. Detection of ai-generated synthetic images with a lightweight cnn. *AI*, 5(3):1575, 2024.
- Y. Lu, C. Guo, X. Dai, and F.-Y. Wang. Artcap: A dataset for image captioning of fine art paintings. *IEEE Transactions on Computational Social Systems*, 2022. doi: 10.1109/TCSS.2022.3173329.
- Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
- S. Mehta and M. Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021.
- OpenAI. Dall e 3, 2024. URL <https://openai.com/index/dall-e-3/>. Accessed: 2025-03-17.
- OpenJourney. Openjourney, 2024. URL <https://huggingface.co/prompthero/openjourney-v4>. Accessed: 2025-03-17.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.
- Mian Wang, Hongqiao Wang, and Ling Wang. Pgd-optimized patch and noise joint embedded adversarial example for faster rcnn and yolov4. In *2021 5th International Conference on Automation, Control and Robots (ICACR)*, pages 78–83. IEEE, 2021.
- Haiyan Zhang, Quan Wang, and Guorui Feng. Artistic image adversarial attack via style perturbation. *Multimedia Systems*, 29(6):3745–3755, 2023.
- Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. URL <https://arxiv.org/abs/1801.03924>.

Administrative Details

Team Roles

- Kanupriya: Set up original noise addition pipeline, wrote evaluation scripts and clustering visualizations, ran grad PGD, helped write report
- Prachi: Set up main PGD pipeline + patch gradient descent, got preliminary results for experiments, box visualization for patch-based noise addition
- Prith: Put together video slides/planned out time distribution, final classic PGD for full set, wrote majority of the report
- Sukriti: Set up and ran attention-based patch-PGD, helped write report
- Overall, as a team, we had many live sessions where we worked on idea formation for useful extensions as well as worked on code/debugging key issues together.

Github link

<https://github.com/PRITH-S07/CMU-IDL>