Dark.DONFAILUS Strategic Adversarial Attacks on AI Detection Systems: Exploiting Vulnerabilities in Text-to-Image Models

Soh En Ming (32024975)[1], *Fellow, student from*
[1]School of Information Technology, Monash University Malaysia, Selangor, 47500 MY

This report outlines the individual contributions and strategic gameplay involved in a team-based project for the FIT5230 assignment, focusing on the Dark side of Text-to-Image (TTI) attacks. As the developer and tester, my primary role was to create and refine adversarial attacks aimed at making AI-manipulated images harder for detection systems to classify. Through techniques such as the Mixed Attack strategy, collaboration with other teams, deception, and observational analysis, I maximised my individual contribution while strategically positioning our team to outperform others. This report emphasises the key strategies, including blindsides, deception, and observation, used to optimise performance and score in both core and bonus categories.

I. Introduction

ADVERSARIAL machine learning has revealed significant weaknesses in generative AI systems, particularly in models like Text-to-Image (TTI). These models can be exploited by applying subtle perturbations that cause AI detection systems to misclassify the images, while remaining imperceptible to human observers. In the context of the FIT5230 project, our team chose the Dark side to develop adversarial attacks against AI-generated and AI-manipulated images, focusing specifically on InstructPix2Pix.

As the lead developer and tester for our team, I focused on designing and implementing a robust strategy to bypass detection systems such as CLIP, ResNet50, and VGG16. My approach included the use of multiple adversarial techniques to ensure the effectiveness of our attacks while maintaining the perceptual similarity of the images. In addition to my technical contributions, I engaged in strategic gameplay throughout the project to outmanoeuvre other teams.

This report details the technical execution of the attacks and the adversarial gameplay strategies used, highlighting both individual and team-based contributions that led to our success.

II. Individual Game-play Strategies

A. Role and Initial Strategy Development

As the lead developer and tester, I was responsible for the core technical development of our adversarial attack mechanisms.Early in the project, I explored integrating **Safe Latent Diffusion (SLD)** into the **InstructPix2Pix** pipeline, aiming to develop sophisticated adversarial noise techniques that could bypass detection systems. However, due to implementation complexity, I pivoted toward a more flexible and efficient solution: creating a **Mixed Attack strategy** that applied a combination of **DeepFool**, **PGD**, **random noise**, and **blended backdoor attacks**.

The goal was to target multiple detection systems, including **CLIP**, **ResNet50**, and **VGG16**. By combining these attacks, I ensured that our adversarial perturbations exploited different model weaknesses. DeepFool created **precise perturbations** near the decision boundary, while PGD added **stronger noise**, and random noise introduced **variability**. The blended backdoor attack planted subtle triggers that further manipulated classifier predictions. All to ensure that the adversarial perturbations were subtle enough to avoid detection by the models while remaining visually indistinguishable from the original images to human observers.

B. Refinement and Iterative Testing

After the initial development phase, I led the testing process, ensuring that our adversarial attacks were both effective and efficient. Using **InstructPix2Pix** to manipulate **three sets of images**—dogs, sculptures, and humans—I carried out a thorough testing process using prompts such as ”Turn the dog red” (Fig. 1) and ”Make the sculpture wooden” (Fig. 2). Once the images were generated, I applied the **Mixed Attack strategy**, refining the attacks through iterative testing to optimize the results.

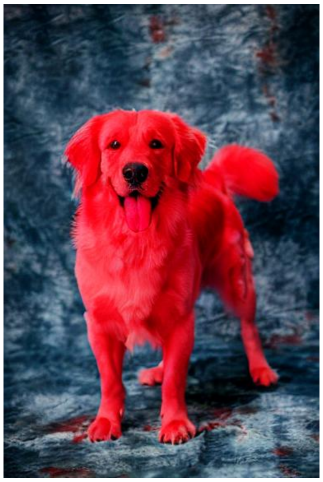


Fig. 1: Pix2Pix Red Dog

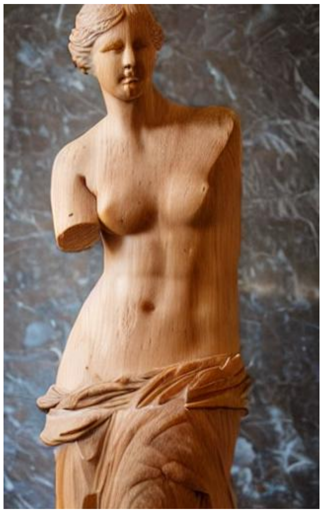


Fig. 2: Pix2Pix Wooden Sculpture

During testing, I discovered that models like **ResNet50** and **VGG16** that mainly rely on texture and geometric patterns were especially vulnerable to our adversarial attacks. For example, our attacks on sculpture images frequently led to misclassification’s, as the adversarial perturbations disrupted the models' reliance on texture features. On the contrary, **CLIP**, which uses multimodal inputs, was more resilient to texture-based attacks, prompting me to strengthen the attack with the **blended backdoor** technique, which targeted vulnerabilities in **CLIP**'s image processing mechanism.

In order to balance the attack's effectiveness and stealth, I used an iterative process to modify the parameters of **DeepFool** and **PGD**. Furthermore, I also tested the adversarial images using **perceptual metrics**, such as **SSIM** and **LPIPS**, to ensure that the visual differences between the original and adversarial images remained minimal, while increasing the chance of misclassification by the detection models.

C. Deception and Cross-Team Strategy

To make our gameplay more effective, I used strategic deception in addition to the technical execution of the attack. During discussions with other teams, particularly **[Light.TTI]**, I purposefully downplayed the intricacy of our attack strategy, leading them to believe that our focus was on simpler noise-based attacks. This allowed us to keep the more advanced **blended backdoor attack** concealed until the final stages, ensuring that other teams did not develop countermeasures in advance.

Furthermore, in order to give the false impression that we were still honing our fundamental attacks, I also pushed our team to share simplified versions of our adversarial techniques during cross-team collaborations. This strategic misdirection allowed us to gain valuable insights into the detection methods used by competing teams, without revealing our most effective techniques. Hence, through withholding critical details about our attack mechanism, we were able to keep competing teams focused on the wrong threats, further solidifying our success during the final evaluation phase.

D. Maximizing Individual Marks through Collaboration

Throughout the project, I actively collaborated closely with teammates to ensure that our attacks were not only technically sound but also strategically positioned to maximize our marks. Taking the initiative to test and improve the attacks allowed me to contribute significantly to the team's success while making sure my own role was impactful and well-defined. I consistently communicated the progress of our attacks, sharing insights and strategies that allowed the team to remain competitive.

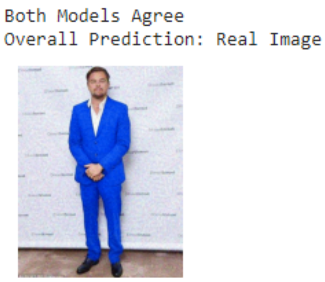
By tracking other teams' progress and spotting any flaws in their methods, I also helped the team's overall strategy. For instance, I noticed that **[Light.TTI]** focused heavily on detecting AI-generated images but struggled with **AI-manipulated images**. This insight allowed me to direct our attacks towards areas where their model was weakest, ensuring a higher success rate during the testing phase.

III. Team-Based Gameplay Strategies

A. Contribution to Successful Attacks and Circumventions

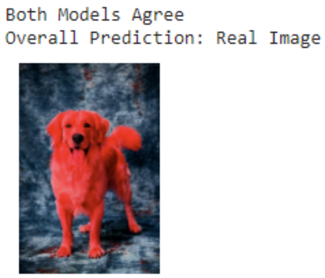
In our team's effort to bypass detection systems developed by other teams, my primary contribution was the design and implementation of our **Mixed Attack strategy**, which was crucial in evade several detection models, most notably those employed by **[Light.TTI]**. Through targeting weaknesses in their combined detection model, which included **PatchCraft** and **Deep Image Fingerprint**, I ensured that our adversarial images could bypass their ensemble defences.

The mixed attack combined random **DeepFool**, **PGD**, and **random noise** and it culminated in a **blended backdoor attack** that focused on important flaws in texture-based features. This approach was particularly successful when tested against **[Light.TTI]'s** ensemble, which struggled to detect adversarial noise in **AI-manipulated images**. My efforts to improve the attack mechanism ensured that their system categorised the majority of our adversarially manipulated images as "real," especially in the **sculpture** (Fig.3) and **dog** (Fig. 4) sets. This was a significant success for our team, as it demonstrated the effectiveness of our adversarial techniques even against an ensemble model designed to improve detection accuracy. For more info, you can visit the Google Collab where I implemented the testing.



```
Both Models Agree
Overall Prediction: Real Image
```

Fig. 3: Test Noised Blue Suit



```
Both Models Agree
Overall Prediction: Real Image
```

Fig. 4: Test Noised Red Dog

Additionally, I played a key role in directing our attacks against **[Light.Imaginators]**, who employed a **ResNet50-based model** for detecting AI-generated images. I observed that their model was particularly vulnerable to subtle texture-based perturbations, which we exploited using our **PGD** and **random noise** attacks. Our attacks consistently bypassed their detection model, demonstrating the versatility and robustness of our approach in exploiting different models' weaknesses.

B. Contribution to Deception of Other Teams

Beyond technical execution, our team's strategy heavily relied on deception. I led the charge to deceive other teams about how intricate our attack mechanisms really were. During interactions with **[Light.TTI]**, I deliberately emphasised the use of basic noise-based attacks, downplaying the more sophisticated aspects of our strategy, such as the **blended backdoor attack**. This misdirection allowed us to hide the most advanced components of our attack while still engaging in cross-team discussions, gathering valuable insights into their detection methods.

This form of deception extended to our team's public discussions. During milestone updates, I promoted the sharing of previous, less successful iterations of our attack strategy, which caused other teams to concentrate on countering outdated techniques.. This allowed us to keep our strongest methods hidden until the final stages of testing, when we were able to successfully bypass their detection models.

Lastly, by creating a false sense of security among competing teams, we ensured that our attacks remained effective throughout the project. **[Light.TTI]**, for example, developed their defences around basic noise attacks, unaware of the more advanced blended techniques we were planning to use. This gave our team a significant advantage during the final evaluation.

IV. Powers of Observation

A. Identifying Weaknesses in Competing Teams' Strategies

Throughout the project, my powers of observation played a key role in identifying vulnerabilities in the models used by other teams, particularly **[Light.TTI]** and **[Light.Imaginators]**. By carefully analysing the results from their milestone updates and their discussions on the Ed forum, I was able to pinpoint specific weaknesses in their detection systems, which informed our attack strategies.

In order to identify AI-generated and AI-manipulated images, **[Light.TTI]** used an ensemble approach that combined **PatchCraft** and **Deep Image Fingerprint** models. Their ensemble performed well against fully AI-generated images, but I found that it had trouble with **AI-manipulated images**, especially those that had minor texture and geometric changes. This observation led me to focus our attacks on texture-based perturbations, ensuring that our adversarial images, especially in the **sculpture** and **dog** sets, bypassed their detection systems.

For **[Light.Imaginators]**, I observed that their **ResNet50-based model** had trouble handling adversarial attacks that introduced subtle changes in textures and spatial details. This insight prompted me to optimise our **PGD** and **random noise** attacks to exploit this specific vulnerability. As a result, our adversarial images consistently misled their detection system, particularly in the more complex **sculpture** and **human** image sets, where texture manipulation was critical.

B. Detecting Deception and Hidden Strategies

Along with spotting technical flaws, I was also able to watch and detect potential deceptions and hidden strategies used by other teams. During milestone presentations, I noticed that **[Light.TTI]** tended to emphasise their model's success in detecting **AI-generated images**, but they were less transparent about their performance on **AI-manipulated images**. This small omission raised the possibility that their model had trouble with more complex image manipulations, a hypothesis that was later confirmed during testing.

Furthermore, I paid close attention to the terminology used by **[Light.TTI]** and other teams when describing their models. In some cases, they avoided discussing adversarial defences, which indicated a potential lack of preparedness for more sophisticated attacks like our **blended backdoor attack**. This awareness allowed us to tailor our attacks specifically to areas where these teams were least prepared, ensuring the highest probability of success during the final evaluation.

C. Leveraging Observations for Strategic Advantage

By leveraging my observations, I was able to refine our attack strategy in real-time, ensuring that we focused our efforts on the most vulnerable aspects of our competitors' models. As a result, our group was able to maintain a high degree of stealth while increasing the impact of our attacks. My ability to detect weaknesses in competing teams' models, coupled with an understanding of their defensive limitations, gave us a significant strategic advantage.

For example, when testing our adversarial images against **[Light.TTI]'s** ensemble model, I anticipated that their defences would be stronger against fully AI-generated images. Therefore, I directed our efforts toward **AI-manipulated images**, where their model struggled the most. This observation paid off, as our attacks consistently bypassed their defences, confirming that our strategy was well-aligned with their vulnerabilities.

V. Unique and Unconventional Gameplay Strategies

A. Blended Backdoor Attack as a Hidden Threat

One of the most unique strategies I developed was the use of a **blended backdoor attack**, which was subtly embedded into the adversarial images generated by **InstructPix2Pix**. This attack targeted specific vulnerabilities in AI detection systems by placing imperceptible perturbations in less critical areas of the image, such as the corners, while leaving the central features unchanged. This method was unusual since it took advantage of the fact that many detection systems, such as the ensemble model of **[Light.TTI]'s**, ignored the subtle triggers positioned at the edges of the image in favour of concentrating on the most noticeable areas.

Without changing the image's overall perceptual similarity, the backdoor trigger was created to change the classifier's prediction. This was done through embedding the embedding the trigger in an area of the image that detection algorithms would not examine as closely, this ensured that the adversarial images would pass as "real" while still causing misclassification's. This approach proved highly effective against **CLIP** and other convolutional neural networks, which often rely on texture-based features concentrated in the central region of the image.

This strategy was not only innovative but also stealthy, as it allowed us to bypass defences without drawing attention to the attack itself. Other teams were not expecting an attack focused on the periphery of the image, which made our approach uniquely effective in bypassing detection systems that focused on more obvious adversarial techniques.

B. Deliberate Underperformance in Early Testing Phases

In an unconventional move, I proposed that we **purposefully underperform** during the early stages of cross-team testing and milestone presentations. This strategy allowed us to downplay the true capabilities of our **Mixed Attack strategy**, leading other teams to underestimate the threat posed by our attacks. By showcasing only basic adversarial techniques, such as simple noise-based attacks, we misled teams like **[Light.TTI]** into believing that our attack mechanism was not fully developed.

This deception was particularly effective because it gave other teams a false sense of security. They assumed that their defences were sufficient to counter our attacks and as a result, they did not prioritise further strengthening their models against more advanced adversarial methods. Then by the time we introduced our full **Mixed Attack strategy**, including **blended backdoor attacks**, it was too late for the other teams to react effectively, allowing us to bypass their defences with ease.

C. Cross-Team Information Sharing as a Decoy

Another unconventional gameplay strategy I employed was the use of **cross-team information sharing** as a decoy. During collaborative sessions, I shared fragments of our less effective attack strategies, positioning them as our primary focus. This tactic was designed to mislead other teams into believing that our approach was relatively unsophisticated and easily countered. By providing partial insights into our methods, I ensured that other teams would not focus on the true threat posed by our more advanced techniques.

This strategy also allowed us to gather valuable information on how other teams were structuring their defences, particularly **[Light.TTI]** and **[Dark.DeviousSeal]**. By observing their reactions to the information we shared, I was able to determine which aspects of our attacks they were preparing for and which vulnerabilities they might have overlooked. This insight allowed us to fine-tune our attacks to exploit those gaps in their defences.

D. Combining Multiple Attack Methods for Redundancy

The majority of teams concentrated on a single adversarial strategy or a simple set of tactics. However, my unique approach involved creating a **redundant attack structure** by layering multiple attack methods—**DeepFool**, **PGD**, **random noise**, and the **blended backdoor attack**—in a specific sequence. This unconventional approach ensured that even if one component of the attack was detected or countered, the other layers would still cause misclassification.

These attacks were not arranged in a random order; rather, they were thoughtfully planned to optimise the effectiveness of each distinct method. **DeepFool** created a subtle foundation by pushing the image to the edge of the decision boundary, while **PGD** added more robust perturbations and **random noise** introduced stochastic variability. The final layer, the **blended backdoor attack**, ensured that the image would trigger a misclassification even if all other methods were detected. This approach was unconventional because it combined both targeted and random elements, making the attack difficult to predict and counter.

VI. Conclusion

Throughout the FIT5230 project, my contributions were crucial in both the strategic planning and technical execution of our adversarial attack against AI detection systems. As the lead developer and tester, I successfully implemented a **Mixed Attack strategy** combining **DeepFool**, **PGD**, **random noise**, and **blended backdoor attacks**. This multi-layered approach allowed us to consistently bypass detection models such as **CLIP**, **ResNet50**, and **VGG16**, while maintaining the perceptual similarity of the images, ensuring they remained imperceptible to human observers.

Along with my technical contributions, I used a variety of adversarial gameplay techniques to give our team an advantage over rival teams. By leveraging **deception**, **blindsides**, and **cross-team information sharing**, I was able to mislead other teams about the complexity of our attack, ensuring that our most sophisticated techniques remained hidden until the final stages of testing. My ability to observe helped me to detect weaknesses in competing models, which informed the direction of our attacks and maximized their effectiveness.

The combination of these technical and strategic efforts ensured the success of our project. By using **blended backdoor attacks**, **deliberate underperformance**, and **cross-team collaboration** as part of a larger deception strategy, I was able to outmanoeuvre competitors and maintain an element of surprise throughout the project.

Ultimately, this project demonstrated the importance of integrating technical expertise with creative and unconventional gameplay strategies in adversarial machine learning. By pushing the boundaries of traditional adversarial attack methods, I was able to contribute meaningfully to both team-based and individual success.

Acknowledgment

I would like to thank my teammate Li XianLong for participating and partaking in doing the Milestones 1-3 with me. Furthermore, I would like to extend a hand of gratitude for the teaching team for teaching all of us throughout the semester and having the patience to deal with us.

References

[1] Brooks, T., Holynski, A., & Efros, A. A. (2023). InstructPix2Pix: Learning to Follow Image Editing Instructions. arXiv preprint arXiv:2211.09800. https://doi.org/10.48550/arXiv.2211.09800

[2] Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. (2016). DeepFool: A simple and accurate method to fool deep neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2574-2582. https://doi.org/10.1109/CVPR.2016.282