HEMLOCK

Defending Artists Against Text-to-Image Models' Style Replication

Capstone Project Report

ABSTRACT

The rapid advancement of text-to-image generation models such as Stable Diffusion and MidJourney has significantly disrupted the creative art industry, enabling effortless replication of artistic styles through simple prompts. This mimicry undermines the integrity, originality, and livelihoods of artists worldwide. In this project, we introduce Hemlock, a robust model designed to protect artists' work from unauthorized style replication. Hemlock achieves this by applying minimal perturbations—referred to as "style cloaks"—to digital artwork, preventing generative models from extracting and mimicking the original artistic style while preserving the visual integrity of the artwork.

Hemlock employs a modified U-Net architecture with an advanced CloakBlock framework that blends the features of style-transferred images into the latent space of the original artwork. Evaluation metrics such as Root Mean Square Error (RMSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) validate Hemlock’s performance, demonstrating an average RMSE of 0.00168104, PSNR of 55.492298 dB, and SSIM of 0.9963815, ensuring minimal pixel-level disruptions while retaining perceptual quality.

Our results showcase Hemlock as an effective tool for safeguarding digital art against AI-generated mimicry, offering a novel solution that balances artistic integrity with technological defense. This work represents a significant step toward ethical AI practices in creative industries.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Network
ReLU	Rectified Linear Unit
GAN	Generative Adversarial Network
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural Similarity Index Measure
RMSE	Root Mean Squared Error
VGG	Visual Geometry Group
MSE	Mean Squared Error
LPIPS	Learned Perceptual Image Patch Similarity

TABLE OF CONTENTS

INTRODUCTION

1.1 Project Overview

The Project Overview introduces the **Hemlock** model, its motivations, and its significance in the current context of AI-generated art and the challenges faced by human artists. The rise of text-to-image generation models has created a disruptive environment for artists by enabling the replication of their unique artistic styles. The primary objective of this project is to provide a solution to this issue by introducing a tool that protects artists' work without compromising the original artistic integrity. Hemlock aims to offer a defense against AI mimicry by applying subtle changes, or style cloaks, to digital artworks, ensuring that diffusion models cannot replicate them.

1.1.1 Technical Terminology

Text-to-Image Generation: This refers to the process of generating images from textual descriptions using advanced machine learning models. These models, such as Stable Diffusion and MidJourney, learn to create highly detailed, realistic images based on simple textual prompts. While this technology has various applications, it also raises ethical concerns, particularly in the context of unauthorized style replication.

Style Cloak: A style cloak is a subtle perturbation applied to an artwork to prevent text-to-image models from extracting and mimicking its style. The cloak modifies the latent space representation of the artwork, shifting it toward a target style without altering the visual integrity.

U-Net Architecture: U-Net is a convolutional neural network architecture primarily used for image segmentation. In Hemlock, the U-Net framework is adapted for feature blending, where the content of an image is blended with the content of a style-transferred image, preserving original features while introducing new stylistic traits.

VGG19: VGG19 is a pre-trained deep learning model used for feature extraction in images. In Hemlock, VGG19 is used to analyze the high-level features of the artwork to ensure that the minimal perturbations do not affect the artistic integrity of the original artwork.

One-Pixel Attacks: These are adversarial attacks in which a single pixel of an image is altered to deceive AI models. The change is so small that it is often imperceptible to the human eye, but it can lead to incorrect model predictions. Hemlock uses similar principles to introduce minimal perturbations that disrupt the replication of the artwork style.

Perceptual Metrics: These metrics, including Root Mean Square Error (RMSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM), are used to evaluate the visual and structural similarity between the original and the cloaked images, ensuring that the perturbations do not degrade the quality of the artwork.

1.1.2 Problem Statement

The rise of text-to-image models such as Stable Diffusion and MidJourney has led to widespread concerns among artists about the unauthorized replication of their work. These models can generate images based on simple text prompts, replicating not only specific elements but also entire artistic styles, without requiring the creator's consent. This has led to significant challenges for independent and professional artists, as their original creations are often used by these models without proper attribution or compensation.

As a result, artists face economic, reputational, and artistic challenges. The time spent by artists to hone their craft and develop their own unique styles is undermined when these styles are replicated by AI without their permission. Artists struggle to maintain control over their creative identity and livelihood, as mimicked versions of their work often overshadow their original pieces in search results and social media platforms.

Hemlock addresses this problem by introducing a novel technique that protects digital artworks from AI mimicry. By applying minimal changes to the artwork’s latent features, Hemlock prevents text-to-image models from extracting and replicating an artist’s unique style, thereby safeguarding their intellectual property

1.1.3 Goal

The primary goal of Hemlock is to empower artists to retain control over their creative works in the age of AI-generated art. Specifically, the goals are:

- **To protect artistic integrity:** By applying subtle perturbations to an artwork, Hemlock ensures that the unique style of the artist is not easily replicated by text-to-image models.
- **To preserve visual quality:** Despite the modifications, Hemlock maintains the visual and perceptual quality of the original artwork, ensuring that the final product still reflects the artist's intent.
- **To offer a practical solution for artists:** Hemlock provides an easy-to-use tool that artists can apply to their digital artworks to safeguard them from unauthorized replication without compromising their creative vision.
- **To bridge the gap between creativity and technology:** Hemlock creates a defense mechanism that respects and enhances human creativity while engaging with emerging AI technologies, ensuring that artists' work is not commodified or exploited without consent.

1.1.4 Solution

Hemlock presents a sophisticated solution to the problem of style replication by text-to-image models. The model applies style cloaks—minimal perturbations introduced to an image that shift its representation in the model's feature space while maintaining the artwork's original aesthetic. This approach prevents AI models from replicating the artist's style while preserving the visual integrity of the artwork. Hemlock's core methodology integrates multiple advanced techniques, including:

U-Net Architecture: This architecture allows Hemlock to blend the content of two images (the original artwork and a style-transferred image), ensuring that the cloaked artwork maintains high visual quality and stylistic integrity.

VGG19 for Feature Extraction: Hemlock leverages VGG19, a deep learning model, to extract features from the artwork, ensuring that the perturbations introduced during the cloaking process do not negatively affect the artistic details.

Adversarial Techniques: Inspired by one-pixel attacks, Hemlock modifies the image in a way that is nearly imperceptible to the human eye but disruptive enough to prevent AI models from mimicking the original style.

Perceptual Metrics: To evaluate the success of the cloaking process, Hemlock uses RMSE, PSNR, and SSIM to quantify the visual and structural fidelity of the cloaked image, ensuring minimal deviation from the original artwork.

The outcome is a robust tool that helps artists preserve their unique styles while allowing them to engage with the growing use of AI in creative industries. Hemlock is designed to be easily integrated into artists' workflows, offering a practical and effective defense against style mimicry by AI models.

1.2 Need Analysis

In recent years, the rise of text-to-image generation models like MidJourney and Stable Diffusion has led to a rapid increase in synthetic art creation. In September 2022, MidJourney reported over 2.7 million users, generating 275,000 AI-generated art images daily. This surge in AI-generated art has created significant concerns within the artistic community. The Concept Art Association raised over $200K to combat AI art mimicry and filed a class action lawsuit in the United States against AI art companies. In November 2022, artists organized a large protest against ArtStation, a popular digital art-sharing platform, which allowed users to post AI-generated artwork without proper identification, further escalating concerns about the loss of artistic ownership and integrity.

Artists today face a dilemma: they want to share and promote their work online, but doing so exposes them to the risk of their art being used without permission to train AI models that replicate their unique styles. As AI models continue to evolve, they increasingly displace original artworks in search results, which disrupts the artist's ability to advertise and promote their work effectively. This synthetic mimicry results in the overshadowing of genuine creativity, making it difficult for artists to capture the attention of potential customers, clients, or collaborators.

The growing prevalence of AI-generated art is especially damaging to art students and emerging artists, who are training to develop their own unique styles. Many are demoralized by the rise of AI models that mimic specific artists' works, often shared freely online, which devalues their creative efforts. For professional artists, the consequences are even more severe: their livelihoods are directly threatened as AI models replicate their styles and flood the digital marketplace with cheap imitations. Despite their efforts to fight back through lawsuits, online boycotts, and petitions, the legal and regulatory processes are often slow, complex, and difficult to enforce internationally.

As a result, artists are faced with a stark choice: 1) do nothing and risk having their work replicated and devalued by AI models, or 2) stop sharing their work

online entirely, which cripples their primary means of promoting and selling their art. Both options are detrimental to artists' financial well-being and creative freedom.

Existing methods such as watermarking and copyrighting offer limited protection against the automated and widespread style replication by AI systems. These traditional approaches fail to address the unique challenges posed by generative AI, which can replicate and transform artistic styles with increasing sophistication. Therefore, there is a pressing need for proactive defense mechanisms that can effectively counter the evolving capabilities of AI, offering artists a robust solution to preserve their creative control, maintain their financial interests, and continue sharing their work online without fear of exploitation.

1.3 Research Gap

- **Adaptability**: Existing models like Glaze are not adaptable to the rapid advancements in generative AI. Any technique developed needs to evolve continuously to keep pace with emerging trends and technologies.
- **Resource-Intensive**: Many current models that protect artwork are computationally heavy and require constant updates. This poses challenges for individual artists and smaller entities with limited computational resources.
- **Bypassing Defenses**: Advanced backdoor and poisoning attacks, as demonstrated by Narcissus and Nightshade, can circumvent traditional protective models, indicating that existing defenses may not be strong enough to counter complex AI threats.
- **Artifacts and Distortions**: Tools like Fawkes, Lowkey, and Photoguard provide protection but often introduce artifacts or subtle visual distortions that compromise the aesthetic integrity of the artwork. Ensuring minimal perceptible changes to the artwork is a critical challenge.
- **Lack of Scalable Solutions**: Existing methods fail to provide scalable and resource-efficient protection mechanisms that artists can easily adopt, particularly those without significant computational resources.
- **Inability to Adapt**: The current solutions lack the flexibility to adapt to new AI models and evolving attack vectors, leaving them vulnerable to emerging technologies.

1.4 Problem Statement and Scope

The rise of generative AI models, particularly in the realm of text-to-image generation through diffusion models, has transformed content creation, especially in visual arts. These models, often trained on large datasets containing copyrighted and private artworks without the consent of the original creators, present significant challenges to both the artistic integrity and the livelihoods of human artists. The ability of AI to replicate an artist's style has led to widespread style mimicry, undermining the artistic identity and market value of original creations.

To address this issue, we propose Hemlock, a model designed to introduce subtle modifications to an artwork, ensuring that the original style is preserved while preventing AI models from replicating it. By applying a style cloak that subtly masks key features of the artwork, Hemlock ensures the protection of the artist's work without compromising its visual appeal.

The scope of Hemlock is centered on developing a deep learning model that blends features from alternative art styles into the original artwork’s feature map, creating a cloaking effect. This project focuses on the design of a custom model and thorough testing to validate its effectiveness in real-world applications. The goal is to offer a practical, scalable solution that enables artists to protect their intellectual property.

1.5 Assumptions and Constraints

S No.	ASSUMPTIONS
1.	For our model to work efficiently it is extremely crucial that it is supplied with high-resolution digital versions of artwork as then only cloaking techniques could be effectively applied and artwork can be reconstructed without compromising the performance of the technique. So, we assume that the artists who wish to protect their work have good quality digital versions of their artwork available for processing.
2.	It is assumed that the diffusion models mimicking artistic styles have been trained on datasets where explicit artist consent was not obtained. This aligns with the reality that many text to image generative models use large-scale datasets that are often scraped from the internet, which includes copyrighted material without artist consent.
3.	Assumption has been made that the artists and institutions relying on Hemlock have access to sufficient computational resources required for the cloaking process.Usually, deep learning models and image processing techniques require significant resources, particularly GPUs. We assume that users of Hemlock will have the necessary hardware to apply the cloaking technique efficiently, as it involves complex computations.
4.	We assume that the mimics have the access to the weights of generic text-to-image models and fine-tunes its model on images of the artist's artwork.

S No.	CONSTRAINTS
1.	We must ensure that the cloaking process does not degrade the perceptual quality of the artwork, i.e. the cloaked version should not be indistinguishable from the uncloaked version to humans, this maintains integrity and value of the artist's work. Any degradation in visual quality could undermine the artist's reputation.
2.	The style of artworks varies widely from realistic, abstract to minimalist, Hemlock must be versatile enough to handle these varied designs.The model should be adaptable to various forms of artist expressions to ensure effectiveness and usability.
3.	Generative AI is continuously evolving and the cloaking technique must be robust enough to withstand updates in these models that could bypass the cloaking mechanisms over time.

1.6 Standards

Implementing standards is essential to ensure the functionality, safety, and reliability of image enhancement systems. Compliance with these standards guarantees cross-platform compatibility and improves the system's operational efficiency. Below are the key standards relevant to the design and development process:

1. **Machine Learning and AI Standards:**
    - **ISO/IEC JTC 1/SC 42**: Artificial intelligence standards providing guidelines for AI governance, machine learning, and use of neural networks.
    - **NIST IR 8269**: A taxonomy and terminology of adversarial machine learning that includes definitions of terms and scenarios specific to adversarial approaches in AI.
2. **Performance and Efficiency Standards**:
    - **IEEE 1857**: Standards for empowering computational performance of multimedia and signal processing, relevant to the optimization of our model for efficient real-time processing.

- **Energy Star Certification**: Guidelines for energy-efficient products and practices, relevant for evaluating the energy consumption of the computing resources.

Compliance with these standards will ensure that our model performs efficiently across various metrics while meeting global regulations, thereby enhancing its scalability and applicability in diverse environment

1.7 Objectives

- To design, implement and evaluate a technical alternative to protect artists against style mimicry by text-to-image diffusion models.
- Curating an original dataset by collecting artwork from college societies and using wikiart dataset.
- Ensure minimal perceptible changes to original artworks while maximizing effectiveness against AI models.
- Evaluate the robustness of our project against known metrics.

1.8 Methodology

To develop Hemlock, we adopted a methodology that combines image processing and machine learning techniques to cloak images and introduce minimal perturbations, ensuring the protection of the artwork while preserving its visual integrity. The approach is outlined as follows:

1. **Data Preparation:** The first step involved curating a dataset for training and testing the model. We collected a high-resolution dataset of artwork images from the Fine Arts and Photography Society (FAPS) and Echoes Club. Additionally, we gathered style-transferred versions of these artworks, which were used to create the necessary reference for blending features and testing the effectiveness of the cloaking mechanism.

2. **Model Architecture:** The Hemlock model is inspired by the U-Net architecture, which consists of encoder-decoder blocks that capture and reconstruct image features. We enhanced this architecture by introducing the

CloakBlock layer. The CloakBlock blends features from the style-transferred images into the original artwork, ensuring that key stylistic elements are masked while retaining the overall visual integrity. The blending process is guided by cosine similarity scores to determine which features from the style-transferred image should be incorporated into the original artwork's feature map, without causing noticeable visual differences.

3. **Feature Integration:** The Hemlock model integrates features from the style-transferred images into the original artwork using the CloakBlock. This process ensures that while some of the original features are masked, the artwork still appears visually consistent with the original, preserving its aesthetic quality. This blending is done in the latent space of the image, ensuring that the overall visual identity is maintained.

4. **Loss Function and Optimization:** We implemented a combined loss function that incorporates both Perceptual Loss (using a pre-trained VGG19 model) and Mean Squared Error (MSE). This loss function helps to maintain perceptual similarity between the original and cloaked images while ensuring minimal pixel-level discrepancies.

5. **Evaluation:** Hemlock's performance was evaluated using perceptual metrics such as LPIPS, PSNR, and SSIM. Extensive testing was conducted to confirm that the cloaking technique successfully prevents AI models from replicating the original style while maintaining the aesthetic integrity of the artwork. These metrics ensured that the cloaked images remained perceptually similar to the original, with minimal visible distortions or quality loss.

This provides a basic summary of how we implement the model.

1.9 Project Outcomes And Deliverables

1. **Hemlock Model**: A fully developed and functional deep learning model that incorporates the UNet-inspired architecture and the custom CloakBlock. This

model should be capable of cloaking original artworks by integrating stylistic features from style-transferred images.

2. **Cloaking Algorithm:**The detailed algorithm or code responsible for the cloaking process, including feature extraction, integration of style attributes, and the application of the combined loss function.

3. **Dataset:** An original dataset curated by collecting original artworks from college societies like FAPS , Echoes.

4. **Evaluation Metrics and Results**:Documentation and analysis of the evaluation metrics used to assess the model's performance, such as PSNR, and SSIM. This includes quantitative results and qualitative assessments demonstrating the effectiveness of the cloaking technique.

1.10 Novelty

Our project introduces several novel aspects that set it apart from existing solutions and approaches in the field of protecting artists against style mimicry by text-to-image diffusion models.

1. **Preservation of Artistic Integrity:** One of the key novelties of our project is its emphasis on preserving the artistic integrity of the original artworks. While existing solutions may introduce noticeable distortions or alterations to the artworks, our project ensures minimal perceptible changes, thereby maintaining the visual appeal and authenticity of the artistic creations.

2. **Comprehensive Evaluation:** Existing prototypes evaluate the performance of model based on direct visual assessment by human artists in a user study , rather than following this approach we would assess our model based on Image Quality Assessment (IQA) metrics like SSIM (Structure Similarity Index Metric), MSE (Mean Squared Error) etc. to quantify the difference between the original and cloaked images.

3. **Original Dataset:** Instead of sourcing images from the web, potentially encountering mimicked images, our strategy involves curating a unique

dataset. We aim to gather original artwork from our college's societies like the Fine Arts and Photography Society (FAPS) and the Echoes Club. This approach promises a broader perspective, enabling us to evaluate the effectiveness of our model more comprehensively.

REQUIREMENT ANALYSIS

2.1 Literature Survey

2.1.1 Related Work

The theory underlying this project revolves around the urgent need to protect intellectual property rights in the era of rapidly advancing artificial intelligence (AI) technologies. Text-to-image generator models such as Stable Diffusion and MidJourney, while innovative, pose a significant challenge for artists, whose works are increasingly vulnerable to unauthorized replication and theft. These AI models can inadvertently infringe upon the rights and integrity of artists by generating images that closely resemble or replicate existing artworks without consent.

Our project seeks to address this critical issue by developing a state-of-the-art system designed to safeguard artists' rights and prevent the unauthorized use of their work. Through this system, we aim to create robust protections that can effectively deter theft and ensure that artists maintain control over their creations.

In our research, we have identified several existing technologies that attempt to tackle similar problems, including Glaze, Nightshade, and Narcissus. This literature review will explore the theoretical foundations, functionality, effectiveness, and limitations of these tools. By understanding and building upon these existing solutions, our goal is to develop an advanced system that offers enhanced protection for artists in the digital age, ensuring their work remains secure amidst technological advancements.

2.1.2 Existing Systems and Solutions

Glaze:

Researchers from the University of Chicago's Department of Computer Science have developed Glaze [1], a tool designed to protect artists from the unauthorized mimicry of their artistic styles. Glaze works by allowing artists to apply "style cloaks" to their artwork, which are subtle perturbations that confuse generative models attempting to replicate an artist's style. This tool represents a collaboration between researchers and the artist community, with the goal of empowering artists to maintain their unique artistic identities against AI-driven threats.

To evaluate Glaze's effectiveness, researchers conducted user studies involving 1,156 participants from the artist community. They assessed the tool's efficacy, usability, and robustness against various countermeasures. The effectiveness of Glaze is measured using metrics such as the Artist-rated Protection Success Rate (Artist-rated PSR) and a scalable metric based on CLIP-based genre shift [2].

The findings indicate that Glaze is highly effective in disrupting mimicry, even at low perturbation levels ($p=0.05$), with success rates exceeding 92% under standard conditions and 85% against adaptive countermeasures. Even in more challenging scenarios—such as when artists have already shared a significant amount of artwork online—Glaze remains robust, with 87.2% of surveyed artists considering the protection successful, even when only a portion of their art is cloaked (with 75% remaining uncloaked). The pre-trained CLIP model used in this evaluation demonstrates strong genre classification performance, achieving top-3 accuracy rates of 96.4% for WikiArt artwork and 94.2% for ArtStation artwork.

However, Glaze has its limitations. Artists must obscure parts of their artwork that might be used in the mimic model's training dataset, which can be difficult for established artists with consistent styles. Their older works, already available on platforms like ArtStation and DeviantArt, can still be imitated using pre-existing data. Additionally, as new techniques emerge, Glaze's protections might become less effective over time, potentially leaving previously safeguarded art vulnerable again.

Narcissus:

In their 2023 paper, "Narcissus: A Practical Clean-Label Backdoor Attack with Limited Information," Zeng et al. [3] present Narcissus, a novel approach to clean-label backdoor attacks. Traditional backdoor attacks involve injecting malicious data into the training phase of machine learning models. This data contains a hidden trigger that, when encountered in test inputs, causes the model to misclassify the input according to the attacker's intent. Earlier research by Gu et al. (2017) and Liu et al. (2018) explored the risks posed by such attacks, which typically modify a small fraction of the training data with known labels to embed the backdoor trigger.

Clean-label backdoor attacks, introduced by Wang et al. (2019), are more subtle. They ensure the injected data appears correctly labeled, making these attacks harder to

detect and mitigate. Zeng et al. (2023) advance the state-of-the-art in clean-label backdoor attacks with Narcissus, which, unlike previous methods, requires only target-class samples and public out-of-distribution (POOD) data. This makes Narcissus more practical in real-world scenarios, as it does not require access to the entire training dataset.

Narcissus optimizes the backdoor trigger to align better with the target class, resulting in higher attack success rates while maintaining stealth. The assessment of Narcissus shows that it is highly effective compared to other clean-label attacks, even with minimal attacker information and a low poison ratio. This highlights the vulnerability of widely-used machine learning systems to practical backdoor attacks and underscores the need for robust defense mechanisms. Furthermore, extending Narcissus to physical-world scenarios demonstrates its versatility and potential impact beyond digital environments.

Nightshade:

Researchers at the University of Chicago's Department of Computer Science have also introduced Nightshade [4], an optimized prompt-specific poisoning attack designed to disrupt text-to-image generative models. Unlike previous backdoor poisoning attacks, which required access to the model's internal pipeline, Nightshade shows that successful prompt-specific poisoning can be achieved with significantly fewer poison samples.

The study reveals that Nightshade attacks are particularly effective in disrupting general features in text-to-image models, impairing their ability to generate coherent images. The effectiveness of the attack depends on concept sparsity, with poisoning attacks proving most effective against sparser concepts. The research also emphasizes that semantic frequency is more important than word frequency in accurately representing concept sparsity.

When the attacker and target models are aligned, success rates for Nightshade attacks range from 72% to 96%, which is notably high. However, even when the models differ, success rates remain relatively high. For instance, poisoning an SD-V2 model with an LD-CC model results in a 76% success rate. Higher success rates are observed with more complex attacker models, and across different prompt types,

success rates generally range from 89% to 91%, with only a slight drop for "recontextualization" prompts. Although poisoning attacks can succeed with differing model architectures, the success rates tend to be lower compared to when identical models are used, and are also influenced by the type of prompt.

2.1.3 Problem Identified

Despite the advancements in anti-mimicry technologies, a significant gap remains in providing comprehensive protection for artists against AI-driven style replication. The existing solutions, such as Glaze, Narcissus, and Nightshade, while innovative, fall short of delivering long-term, foolproof defenses. The primary issues identified include:

1. **Limited Longevity of Protection**: As AI models evolve, the protective measures provided by tools like Glaze may become obsolete, leaving previously protected artwork vulnerable to new replication techniques.
2. **Inadequate Coverage for Established Artists**: Artists with a significant online presence and a consistent style face challenges in protecting their work, as their older, widely-shared pieces can still be imitated using pre-existing data.
3. **Complexity and Resource Intensity**: Implementing and maintaining effective anti-mimicry solutions can be resource-intensive, requiring artists to constantly update their protective measures as new threats emerge.
4. **Ineffectiveness Against Advanced Attacks**: Tools like Narcissus and Nightshade illustrate that advanced backdoor and poisoning attacks can bypass traditional defenses, further complicating the protection landscape.

These challenges highlight the need for a more robust, adaptable solution that can provide enduring protection against unauthorized replication of artistic styles, motivating the development of more sophisticated techniques such as the proposed Hemlock system.

2.1.4 Survey of Tools and Technologies Used

In developing **Hemlock**, we leveraged several existing tools and technologies that informed our approach to protecting artists' work from unauthorized style replication

by AI models. Here's an overview of the key tools and methodologies that played a crucial role in shaping our solution:

1. Cloaking Systems

We drew inspiration from three prominent cloaking systems: Fawkes, Lowkey, and Photoguard. Each of these tools is designed to protect digital images from being exploited by AI models.

1. **Fawkes**: Originally developed to obscure facial images from recognition systems, Fawkes was adapted for anti-mimicry protection by using a different feature extractor and random artworks from other artists as target images. It introduces subtle perturbations to mislead AI models attempting to replicate an artist's style.
2. **Lowkey**: Similar in design to Fawkes, Lowkey focuses on optimizing cloaked images so that they differ from the original artwork rather than pushing them toward a specific target image. This makes it harder for AI models to replicate the style.
3. **Photoguard**: This tool minimizes the norm of the image feature vector, making it analogous to Fawkes when the zero feature vector is chosen as the optimization target. Photoguard aims to provide protection by altering the image in a way that is imperceptible to the human eye but confuses AI models.

While these tools offer some level of protection, they were found to be less effective against advanced mimicry techniques, often introducing slight artifacts without fully preventing style replication.

2. Image Quality Assessment (IQA) Metrics

To assess the effectiveness of cloaking systems and ensure minimal impact on the visual integrity of the artwork, we utilized several Image Quality Assessment (IQA) metrics:

1. **FSIM (Feature Similarity Index Metric)**: Evaluates the similarity between images based on feature similarities, ensuring that the cloaked image retains its core features.

2. **SSIM (Structural Similarity Index Metric)**: Measures the structural similarity between the original and cloaked images, focusing on the preservation of structural information.
3. **RMSE (Root Mean Squared Error)**: This metric assess pixel-level differences and the quality of image reconstruction, helping to minimize any noticeable changes in the artwork.
4. **LPIPS (Learned Perceptual Image Patch Similarity)**: A perceptual metric that compares high-level feature representations between the original and cloaked images, ensuring that the visual quality remains close to what the human eye perceives.
5. **PSNR (Peak Signal-to-Noise Ratio):** PSNR was employed to quantify the extent of any perceptible changes introduced by the cloaking process. A higher PSNR value indicates that the cloaked image is very similar to the original image, meaning that the alterations made by Hemlock to obscure the style are minimally noticeable to the human eye.

These IQA metrics were crucial in refining our evaluation process, allowing us to maintain the visual integrity of the artwork while providing robust protection against AI-driven mimicry attacks. Together, these tools and technologies formed the foundation of our approach, guiding the development of Hemlock into a state-of-the-art solution for safeguarding artists' creations in the face of rapidly advancing AI capabilities.

2.2 Software Requirement Specification

2.2.1 Introduction

2.2.1.1 PURPOSE

The purpose of Hemlock is to facilitate artists across the globe with a reliable tool that could help them to safeguard their work from modern age style mimicry AI tools along with ensuring that visual integrity of the work is not compromised and minimal differences could be observed perceptually as well as by using scientific metrics between original and the artwork processed by Hemlock.

2.2.1.2 INTENDED AUDIENCE AND READING SUGGESTIONS

- **Developers**: To understand the software architecture, coding practices, and implementation details.
- **Artists and Art Enthusiasts**: To comprehend how the tool will protect their creative work.
- **Project Managers**: To oversee project timelines, milestones, and deliverables.
- **Researchers**: To study the effectiveness of the Hemlock model.
- **Security Analysts**: To evaluate the security measures to prevent unauthorized access and replication.

2.2.1.3 PROJECT SCOPE

The scope of Hemlock includes the creation of a deep learning model that integrates features from alternative art styles into the original works' feature map providing a cloaking mechanism.The project focuses on the design of a custom model and testing to ensure the effectiveness of the solution in real-world scenarios. With the continued advent in the field of generative AI it is important to have tools to safeguard all kinds of data from unauthorised access , so our project presents a long-term scope.

2.2.2 OVERALL DESCRIPTION

2.2.2.1 PRODUCT PERSPECTIVE

Hemlock is an innovative solution designed as a protective layer for digital artworks that basically acts as a preprocessing tool before anything is shared online to shield it from copying. It fits into the broader ecosystem of digital art tools. Hemlock integrates seamlessly with existing digital art software and platforms, providing an additional layer of security for artists.Whether an artist is preparing to upload their creations to a public gallery, share them on social media, or incorporate them into larger projects, Hemlock ensures that the artwork is cloaked effectively, preventing AI models from replicating the artist's distinctive style.

2.2.2.2 PRODUCT FEATURES

- **Image Cloaking:** Hemlock processes images to hide stylistic features, making them difficult for AI models to replicate.

- **Dataset Integration:** The system supports custom datasets, allowing artists to protect a wide range of art styles.
- **Performance Metrics:** The model evaluates and displays metrics like LPIPS and PSNR to assess the effectiveness of the cloaking process.

2.2.3 NON-FUNCTIONAL REQUIREMENTS

2.2.3.1 Performance Requirements

- **Processing Speed**: The model should be able to cloak images in a few seconds ensuring its speed. The time for doing so would also be dependent on complexity and size of image.
- **Scalability**: The system must handle multiple users and a large number of images simultaneously without significant delays.

2.2.3.2 Safety Requirements

- **Data Integrity**: Ensure that the original artwork and its cloaked versions are stored securely, preventing data corruption or loss.
- **Error Handling**: The system should provide clear error messages and fail-safes in case of processing failures.

2.3 Risk Analysis

- **AI Advancement Risk**: The ongoing evolution of AI models may pose a risk to the effectiveness of Hemlock. Continuous updates and enhancements to the model will be necessary to stay ahead of new AI mimicry techniques.
- **Data Breach Risk**: Storing and processing sensitive artwork data may expose the system to potential breaches.
- **Performance Degradation Risk**: As the number of users increases, the system may face performance challenges. Scalable infrastructure will be important to ensure consistent performance.
- **User Adoption Risk**: Artists may be hesitant to adopt a new tool. User education, intuitive design, and demonstrable effectiveness are key to driving adoption and trust in the Hemlock

METHODOLOGY

3.1 Investigation Techniques

For the development and evaluation of the Hemlock model, we have employed a **combination of experimental and comparative investigation techniques**. This hybrid approach was chosen to ensure a robust evaluation of the model's ability to cloak images while maintaining visual integrity and preventing unauthorized style replication.

Experimental Investigation:
The experimental technique was necessary for developing, training, and testing the model. It involved:

- Designing the U-Net-based architecture with the custom CloakBlock layer.
- Experimenting with various loss functions (e.g., Perceptual Loss, MSE) to achieve optimal balance between cloaking effectiveness and visual quality.
- Testing the model on a curated dataset of original and style-transferred images to validate its performance.

The experimental approach was essential for the iterative design and optimization of the model. By running controlled experiments, we were able to refine the CloakBlock framework, optimize feature integration, and tune hyperparameters to achieve the desired outcomes.

Comparative Investigation:
This technique was employed to benchmark the effectiveness of the cloaked images against the original images in real-world scenarios.

- Both original and cloaked images were passed through Stable Diffusion, and the results were analyzed using the Hugging Face ViT Art Classifier.
- Metrics like RMSE, PSNR, and SSIM were used to compare the structural and perceptual similarities between the original and cloaked images.

The comparative approach ensured that the model’s performance was validated in practical scenarios. Comparing the outputs of Stable Diffusion for original and

cloaked images demonstrated the success of the cloaking mechanism in disrupting AI-generated style mimicry.

The combination of these techniques allowed us to:

1. Build a robust model through controlled experimentation.
2. Validate its effectiveness in preventing style replication under realistic conditions.
3. Ensure reproducibility and reliability of results by systematically comparing outputs across different conditions.

This hybrid investigative approach provided a comprehensive framework to evaluate the functionality and real-world applicability of the Hemlock model

3.2 Proposed Solution

The solution that we proposed i.e. Hemlock, is a robust model designed to safeguard artists' original creations from unauthorized style replication by text-to -image diffusion based generative AI models. Our technique deceives the model and foil mimicry of art style. The main objective while building the model is to subtly alter an artist's original work in a way that it becomes resistant to style extraction and mimicry but preserves the visual integrity of the original art. This is achieved through a combination of deep learning, image processing techniques, and custom architectural components that blend features of alternate art styles in latent space of original artwork to address the unique challenges posed by AI-driven art mimicry.

1. Architecture Design

The project is inspired by the U-Net architecture , which is basically used for image segmentation and reconstruction.Each level of design is built in a layered manner to identify high as well as low level structural features of the image. The choice of this architecture was solely dependent on its encoder-decoder functionality which allows detailed feature extraction.

1.1 Encoder-Decoder Framework

The encoder consists of multiple convolutional layers to extract hierarchical features from high to low level while reducing spatial dimensions using Max-Pooling layer but increasing depth of feature maps.

- **ConvBlock_encoder**: Contains two convolutional layers with batch normalisation and ReLU activation.
- **EncoderBlock**: Combines ConvBlock_encoder with a max-pooling layer for down-sampling.

The decoder operates in reverse of encoder by reconstructing image from extracted features.Original spatial dimensions of the image are obtained by unsampling feature maps using deconvolutional layers

- **ConvBlock_decoder**: Similar to ConvBlock_encoder but uses Tanh activation in the final layer.

- **DecoderBlock**: Combines ConvBlock_decoder with a transposed convolution layer for up-sampling.

1.2 CloakBlock Framework

This refers to a tailored addition to the U-Net Architecture , it incorporates features from style-transferred image in the features maps making it difficult for diffusion models to extract art style.It operates as follows:

- **Feature Extraction :** Original and style transferred images are passed through respective encoder paths to obtain feature maps i.e. $\mathbf{F_{original}}$ **and** $\mathbf{F_{styled}}$ **respectively.**
- **Cosine Similarity Computation:** Cosine similarity between obtained features of original and style transferred model is calculated as:

$$S(fo, fs) = \frac{fo \cdot fs}{\|fo\| \|fs\|}$$

Where $\cdot$ signifies the dot product, and $\|\cdot\|$ stands for the Euclidean norm. The similarity scores are valuable means to detect the features in the original image to be replaced by the features from the style.

- **Feature Integration:** The CloakBlock then integrates features based on the cosine similarity computation.The process happens as the features that have similarity scores in the specified range (i.e., 0.12 to 0.45) are replaced by the features from the style-transferred image in the original image with the same corresponding score.

2. Loss Function and Optimization

As the aim is to maintain visual quality while cloaking we use loss functions to guide the model during the training process so as to introduce minimal perturbations.For this we use a combined loss function.

2.1 Perceptual Loss (LPIPS)

The Learned Perceptual Image Patch Similarity (LPIPS) metric evaluates the perceptual similarity between two images by comparing their high-level feature representations that are captured by a pre-trained deep network (e.g., VGG). This metric maintains visual quality of the original artwork, as seen by human.It is calculated as:

$$\Gamma_{LPIPS} = \sum_{i=1}^{L} \left\| \phi_i(\hat{y}) - \phi_i(y) \right\|_2^2$$

Where ϕ_i represents the activation of the i_{th} layer of a pre-trained network (such as VGG), $\hat{y}$ is the output image, and y is the original image.

2.2 Peak Signal-to-Noise Ratio (PSNR) Loss

In addition to perceptual loss, MSE is used to measure the pixel-level differences between the original and cloaked images, ensuring the changes introduced by the cloaking process are minimal at the pixel level.

$$MSE(x, y) = (1/n) * \sum_{i=1}^{n} (x_i - y_i)^2$$

Where:

- x_i and y_i are the pixel values of the original and cloaked images, respectively.
- n is the total number of pixels in the image.

2.3 Combined Loss

The combined loss is a weighted total of LPIPS AND MSE:

Total Loss=(1-a)×LPIPS+a×MSE

,where a represents the weight

3. Evaluation

To assess the effectiveness of Hemlock, we evaluate the cloaked images using a combination of perceptual and structural metrics. The key evaluation metrics include:

- RMSE: The Root Mean Square Error (RMSE) measures the average squared difference between corresponding pixels in the original and cloaked images.
- PSNR: Assesses the quality and fidelity of the reconstructed image by evaluating signal-to-noise ratio, indicating how well the cloaked image preserves the original.
- SSIM: The Structural Similarity Index (SSIM) is used to evaluate the structural similarity between the original and cloaked images, focusing on luminance, contrast, and texture.

Thus the proposed solution offers a comprehensive way to apply cloaking and foil art style mimicry by diffusion models.

3.3 WORK BREAKDOWN STRUCTURE

1. Project Management

1.1 Planning and Scheduling

- Define scope and objectives
- Develop project timeline
- Allocate resources and budget

1.2 Coordination and Communication

- Regular team meetings
- Progress tracking and reporting

2. Requirements Analysis

2.1 Requirements

- Gather and analyze requirements

2.2 Technical Specifications

- Define technical requirements and data needs

3. Design and Architecture

3.1 System Architecture

- Design overall architecture (UNet, CloakBlock)

3.2 Model Design

- Develop encoder, decoder, and CloakBlock modules

4. Data Preparation

4.1 Data Collection

- Acquire and generate images

4.2 Data Preprocessing

- Resize, normalize, and augment images

5. Model Development

5.1 Implementation

- Develop and integrate model components

5.2 Training

- Train model with images

5.3 Optimization

- Fine-tune model and perform hyperparameter tuning

6. Evaluation and Testing

6.1 Performance Metrics

- Evaluate using RMSE, PSNR, SSIM

3.4 Workable Modules

Model Architecture Module: Focuses on developing the UNet-inspired architecture and CloakBlock. It involves designing encoder and decoder blocks, and integrating the CloakBlock for feature cloaking.

Data Processing Module: Handles data collection, preprocessing, and augmentation. Ensures the quality and suitability of data for model training and evaluation.

Training and Optimization Module: Involves training the model, fine-tuning parameters, and optimizing performance based on loss functions and metrics.

Evaluation and Testing Module: Assesses the model's effectiveness and robustness through extensive testing and performance evaluation.

3.5 Tools And Technology

1. **Programming Language :** Python
2. **Machine Learning Frameworks :** PytorchVision
3. **Image Processing Libraries:** OpenCV ,PIL
4. **Evaluation Tools:** LPIPS Library

DESIGN SPECIFICATION

4.1 SYSTEM ARCHITECTURE

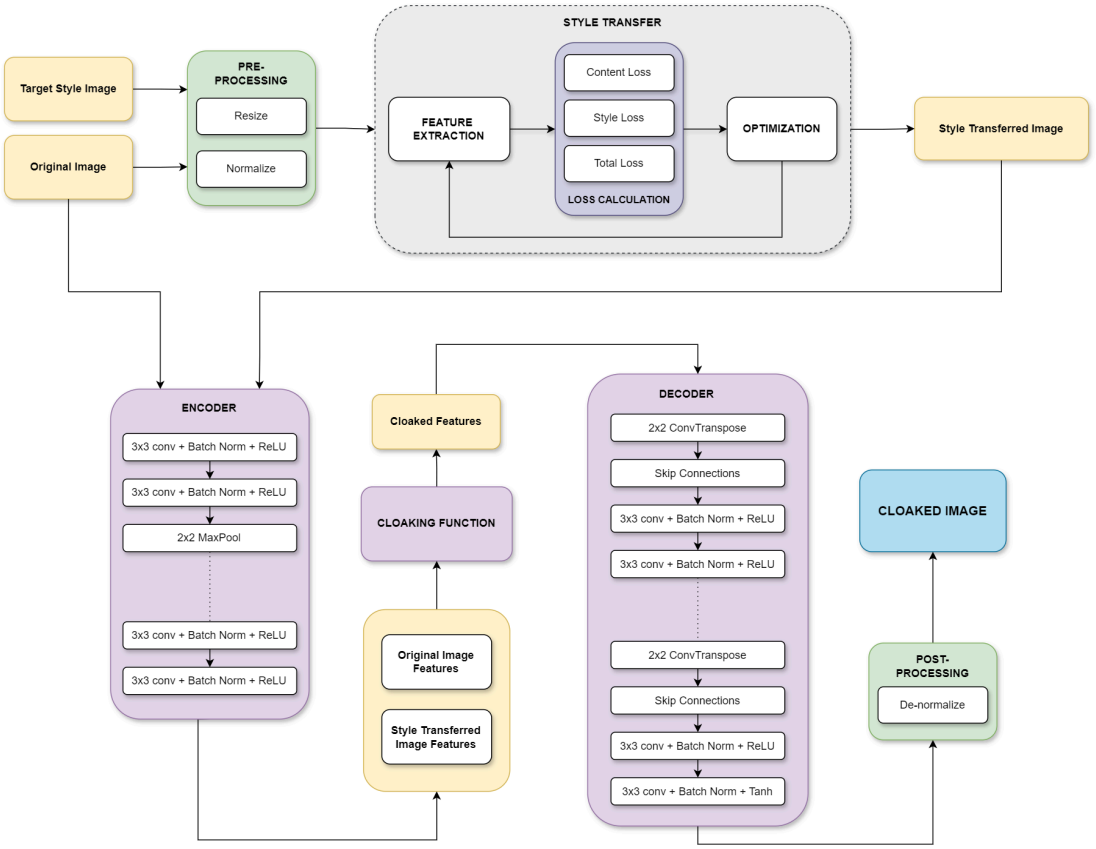

Figure 1: Block Diagram of Architecture

This block diagram represents a Hemlock-style cloaking model for images, featuring three main stages:

1. Style Transfer: Pre-processed target style and original images undergo feature extraction, loss calculation (content/style/total), and optimization to produce a style-transferred image.
2. Cloaking Function: An encoder-decoder framework integrates original and style-transferred features, generating cloaked features.
3. Post-Processing: The output cloaked image is de-normalized for final use.

The architecture uses convolutional layers, batch normalization, ReLU activations, and skip connections.

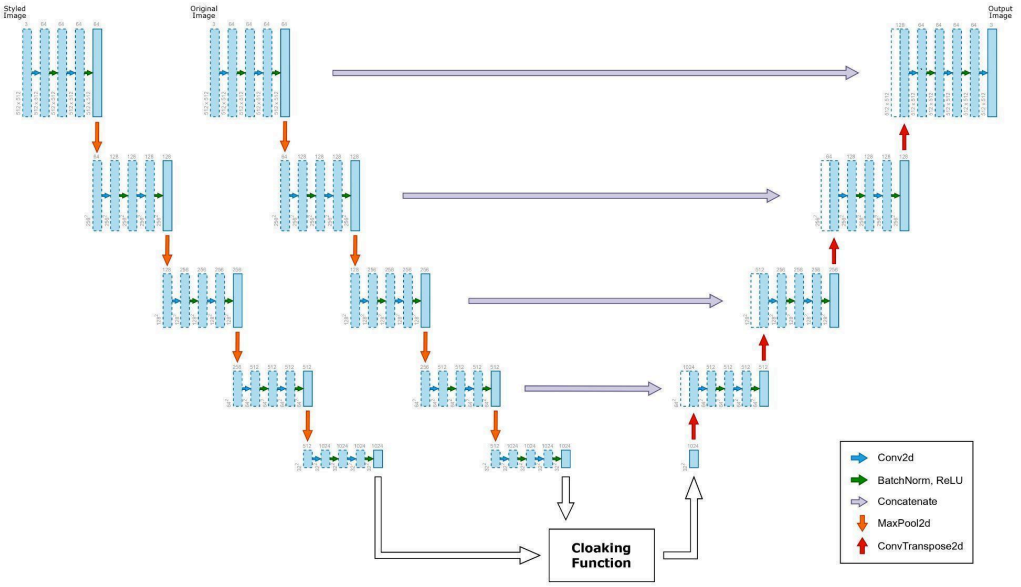

Figure 2: High-level Architecture

4.2 SNAPSHOTS OF WORKING PROTOTYPE

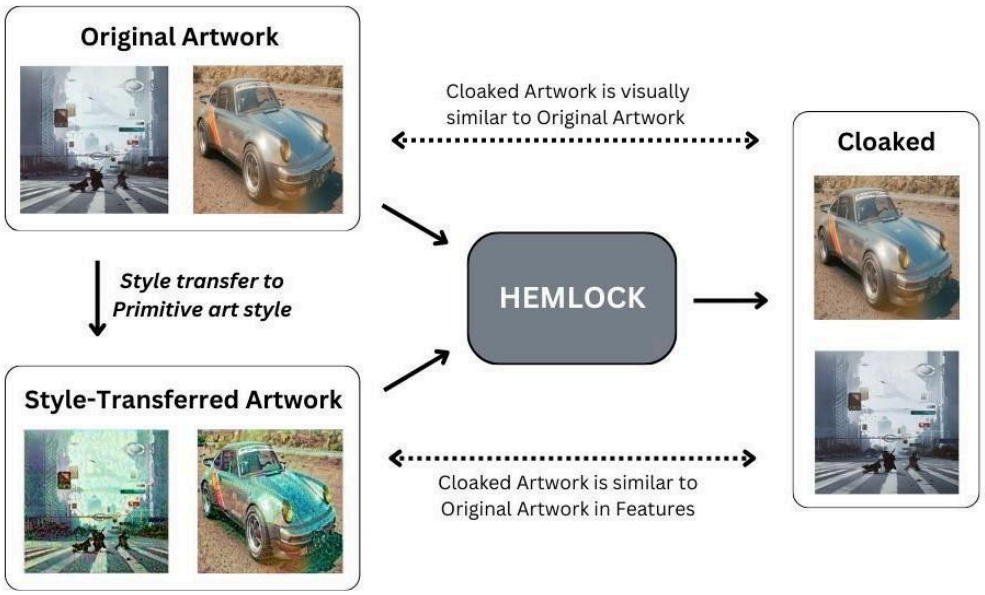

Figure 3: High level overview of how Hemlock style transfers the original artwork to a different style, it optimizes a cloak that makes the artwork's features representation match that of the style-transferred art, while constraining the amount of visible changes to the artwork.

Figure 4: Hemlock protection results. Column 1: artist's original artwork; column 2: mimicked artwork when artist does not use protection; column 3: style-transferred artwork used for cloak optimization and the target style; column 4: mimicked artwork when artist uses cloaking protection. All mimicry examples here use SD-based models.

IMPLEMENTATION AND EXPERIMENTAL RESULTS

5.1 Experimental Setup

The experimental setup for the Hemlock model involves configuring the environment, tools, and hardware to train, evaluate, and test the system effectively. The following configurations were used:

- **Hardware Configuration**:
    - Processor: Intel Core i5-9400KF or equivalent
    - GPU: NVIDIA GTX 1050 (4 GB VRAM)
    - RAM: 8 GB
    - Storage: 256 GB SSD
- **Software and Frameworks**:
    - Programming Language: **Python 3.x**
    - Deep Learning Libraries: **PyTorch Vision**
    - Pre-trained Models: **VGG19**
    - Metrics Libraries: **LPIPS Library**
    - Dataset Preprocessing: **Pandas** and **NumPy**
- **Dataset**: Artwork images sourced from **Fine Arts and Photography Society (FAPS)** and **4437 high-resolution images from "CyberVerse" dataset.**

The experimental simulation was conducted in a controlled environment to ensure consistent results during the model evaluation.

5.2 Experimental Analysis

5.2.1 Data

Data Sources:

- Artwork images were collected from FAPS and Echoes Club.
- Style-transferred versions of these images were generated using VGG19.

Data Cleaning:

- Removal of low-resolution and duplicate images to ensure a clean dataset by standardizing to a uniform resolution of 512x512 pixels.

Data Pruning:

- Only images with clear visual content and stylistic features were retained.
- Final dataset size: 4,437 high-quality cyberpunk-themed images.

Feature Extraction Workflow:

- Extracted high-level features from the images using the VGG19 model.
- Cosine similarity was computed between the features of the original and style-transferred images to identify replaceable components during cloaking.
- The final feature maps were used to train and validate the Hemlock model.

5.2.2 Performance Parameters

The following parameters were used to evaluate Hemlock's performance:

- **Root Mean Square Error (RMSE)**:
  Measures the pixel-level changes between the original and cloaked images. Lower RMSE indicates minimal visual distortion.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}$$

- **Peak Signal-to-Noise Ratio (PSNR)**:
  Evaluates the fidelity of the cloaked image compared to the original. Higher PSNR values indicate better visual quality.

$$PSNR = 10.log_{10}(\frac{MAX^2}{MSE})$$

- **Structural Similarity Index (SSIM)**:
  Measures perceptual similarity between the original and cloaked images based on luminance, contrast, and structure.

$$SSIM = \frac{(2\mu_X\mu_Y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)}$$

5.3 Working of the Project

5.3.1 Procedural Workflow

The following workflow explains how Hemlock operates:

1. **Input Image**: A high-resolution original artwork is provided.
2. **Style Transfer**: A style-transferred version of the artwork is generated using style transfer methods.
3. **Feature Extraction**: Both the original and style-transferred images are passed through the encoder to extract feature maps: $F_{original}$ and F_{styled} .
4. **.Cloaking Process**:
   - **Cosine similarity** is computed between $F_{original}$ and F_{styled}
   - Features with similarity scores in the range 0.12 to 0.45 are integrated into the original feature map using the CloakBlock.
5. **Reconstruction**: The combined feature map is passed through the decoder to reconstruct the cloaked image.
6. **Evaluation**: The cloaked image is assessed using RMSE, PSNR, and SSIM to ensure minimal visual distortion and maximum protection.

5.3.2 Algorithmic Approaches Used

```
Input: Original Image (I_original), Style-Transferred Image (I_style)
Output: Cloaked Image (I_cloaked)

1. Extract feature maps:
   F_original = Encoder(I_original)
   F_styled = Encoder(I_style)

2. Compute Cosine Similarity:
   For each feature pair (f_o, f_s) in F_original and F_styled:
      S(f_o, f_s) = (f_o · f_s) / (||f_o|| ||f_s||)

3. Integrate features:
   For each feature in F_original:
      If 0.12 <= S(f_o, f_s) <= 0.45:
         Replace f_o with blended feature (α * f_o + (1-α) * f_s)

4. Decode to reconstruct cloaked image:
   I_cloaked = Decoder(F_modified)

5. Return I_cloaked
```

Figure 5: Pseudocode

The algorithm uses cosine similarity to determine feature overlap and selectively integrates style-transferred features into the original image. The weighted blending ensures minimal perceptual changes while masking the style.

5.3.3 System Screenshots

```
class ConvBlock(nn.Module):
    def __init__(self, input_filters, output_filters):
        super(ConvBlock, self).__init__()
        self.conv1 = nn.Conv2d(input_filters, output_filters, 3, padding=1)
        self.bn1 = nn.BatchNorm2d(output_filters)
        self.conv2 = nn.Conv2d(output_filters, output_filters, 3, padding=1)
        self.bn2 = nn.BatchNorm2d(output_filters)
        self.activation = BipolarSigmoid()

    def forward(self, x):
        x = self.conv1(x)
        x = self.bn1(x)
        x = self.activation(x)
        x = self.conv2(x)
        x = self.bn2(x)
        x = self.activation(x)
        return x
```

```
class EncoderBlock(nn.Module):
    def __init__(self, input_filters, output_filters):
        super(EncoderBlock, self).__init__()
        self.conv_block = ConvBlock(input_filters, output_filters)
        self.pool = nn.MaxPool2d((2,2))

    def forward(self, x):
        x = self.conv_block(x)
        p = self.pool(x)
        return x, p

class DecoderBlock(nn.Module):
    def __init__(self, input_filters, output_filters):
        super(DecoderBlock, self).__init__()
        self.conv_transpose = nn.ConvTranspose2d(input_filters,
output_filters, 2, stride=2)
        self.conv_block = ConvBlock(output_filters+output_filters,
output_filters)

    def forward(self, x, skip_features):
        x = self.conv_transpose(x)
        x = torch.cat([x, skip_features], axis=1)
        x = self.conv_block(x)
        return x

class CloakBlock(nn.Module):
    def __init__(self):
        super(CloakBlock, self).__init__()

    def forward(self, original, styled):
        global og
        global st
        og = original.detach().cpu().numpy()
        st = styled.detach().cpu().numpy()
        similarity_scores = cosine_similarity_nd(og, st)
        cloaked = clk(og, st, similarity_scores)
        cloaked = torch.from_numpy(cloaked).to(device)
        return cloaked
```

```
class Architecture(nn.Module):
    def __init__(self):
        super(UNet, self).__init__()

        self.s1 = EncoderBlock(3,64)
        self.s2 = EncoderBlock(64,128)
        self.s3 = EncoderBlock(128,256)
```

```
        self.s4 = EncoderBlock(256,512)

        self.k1 = EncoderBlock(3,64)
        self.k2 = EncoderBlock(64,128)
        self.k3 = EncoderBlock(128,256)
        self.k4 = EncoderBlock(256,512)

        self.b1 = ConvBlock(512,1024)
        self.b2 = ConvBlock(512,1024)

        self.cloak_block = CloakBlock()
        self.d1 = DecoderBlock(1024,512)
        self.d2 = DecoderBlock(512,256)
        self.d3 = DecoderBlock(256,128)
        self.d4 = DecoderBlock(128,64)
        self.output = nn.Conv2d(64, 3, 1, padding=0)

    def forward(self, input1, input2):
        # Original image
        s1, p1 = self.s1(input1)
        s2, p2 = self.s2(p1)
        s3, p3 = self.s3(p2)
        s4, p4 = self.s4(p3)
        b1 = self.b1(p4)

        # Styled image
        k1, q1 = self.k1(input2)
        k2, q2 = self.k2(q1)
        k3, q3 = self.k3(q2)
        k4, q4 = self.k4(q3)
        b2 = self.b2(q4)

        # Cloaked features
        b3 = self.cloak_block(b1, b2)

        # Image reconstruction
        d1 = self.d1(b3, s4)
        d2 = self.d2(d1, s3)
        d3 = self.d3(d2, s2)
        d4 = self.d4(d3, s1)
        output = self.output(d4)
        return output
```

Figure 6: System Screenshots

5.4 Testing Process

To validate the effectiveness of the Hemlock model in protecting artwork from unauthorized style replication, we designed a comprehensive testing process. The goal of the testing phase is to verify that the cloaked images prevent generative models like Stable Diffusion from replicating the original style. We use the Hugging Face Art Classifier to evaluate the classification results of both the original and cloaked images and compare their behavior.

5.4.1 Test Plan

The testing process involves the following steps:

1. **Input**: Select an image from the dataset (original artwork).
2. **Cloaking**: Pass the original artwork through the Hemlock model to generate a cloaked version.
3. **Stable Diffusion**: Feed both the original image and the cloaked image into the Stable Diffusion model to observe the outputs (AI-generated results).
4. **Evaluation**: Use the **Hugging Face Art Classifier** to classify the generated images and compare how the classifier distinguishes between the original and cloaked styles.
5. **Metrics**: Analyze scores and results to confirm that the cloaked image sufficiently prevents style replication.

The test plan ensures that the cloaked images appear visually unchanged to humans while disrupting the style mimicry process of AI models.

5.4.2 Features to be tested

The following features are tested to ensure Hemlock's functionality:

1. Cloaking Effectiveness: Verify that the cloaked images prevent Stable Diffusion from replicating the original art style.
2. Visual Quality: Evaluate that the cloaked image retains visual integrity when compared to the original image.

5.4.3 Test Strategy

We adopt a comparative testing strategy to evaluate the Hemlock model's effectiveness in protecting artistic styles while maintaining visual quality. The strategy involves the following steps:

1. **Cloaking Process:**
    - An input artwork (original image) is processed through Hemlock to generate a cloaked image.
2. **Stable Diffusion Analysis:**
    - Both the original image and the cloaked image are passed into the Stable Diffusion model.
    - The outputs (generated images) are visually and structurally compared to observe how well the cloaked image disrupts the AI model's ability to replicate the style.
3. **Visual Integrity Evaluation:**
    - The visual integrity of the cloaked images is compared against the original images using the following perceptual and structural metrics:
        - Root Mean Square Error (RMSE): Measures pixel-level differences to ensure minimal perturbations.
        - Peak Signal-to-Noise Ratio (PSNR): Evaluates the fidelity of the cloaked image relative to the original.
        - Structural Similarity Index (SSIM): Assesses the similarity in luminance, contrast, and structure between the original and cloaked images.
4. **Verification Using Hugging Face Art Classifier:**
    - Both the original and cloaked images are processed through the Hugging Face Art Classifier.
    - The classifier's behavior is observed to confirm that the cloaked image disrupts the ability of AI to replicate or recognize the original artistic style.

5.4.4 Test Techniques

The following techniques are applied to test the Hemlock model:

1. **Functional Testing:** Verifies the core functionality of the CloakBlock framework and ensures that cloaking occurs without noticeable distortions.
2. **Comparative Analysis:** Both the original and cloaked images are analyzed after being processed through Stable Diffusion to evaluate differences in classification results.
3. **Classification Testing:** The Art Classifier determines whether the cloaked image disrupts style extraction, supporting results with numerical scores.

5.4.5 Test Cases

Test Case	Input	Process	Expected Outcome
Cloaking Effectiveness	Original Image	Generate a cloaked image using Hemlock and feed it to Stable Diffusion	Stable Diffusion produces outputs that differ from the original.
Classification Disruption	Cloaked Image vs Original	UseArt Classifier outputs from Stable Diffusion	Art Classifies the cloaked image as stylistically different
Visual Integrity Preservation	Original Image and Cloaked	Compare pixel-level changes using RMSE, PSNR and SSIM	Minimal Changes are detected (high PSNR, low RMSE,high SSIM)

Table 1: Test Cases

5.4.6 Test Results

The test results validate Hemlock's effectiveness in preventing style replication:

1. Cloaking Effectiveness:
    - The cloaked images, when passed into Stable Diffusion, produce outputs that significantly differ in style compared to outputs generated using the original image.

2. Classification Results (Hugging Face Art Classifier):
    - Original Image: Classified accurately
    - Cloaked Image: Classified as stylistically different

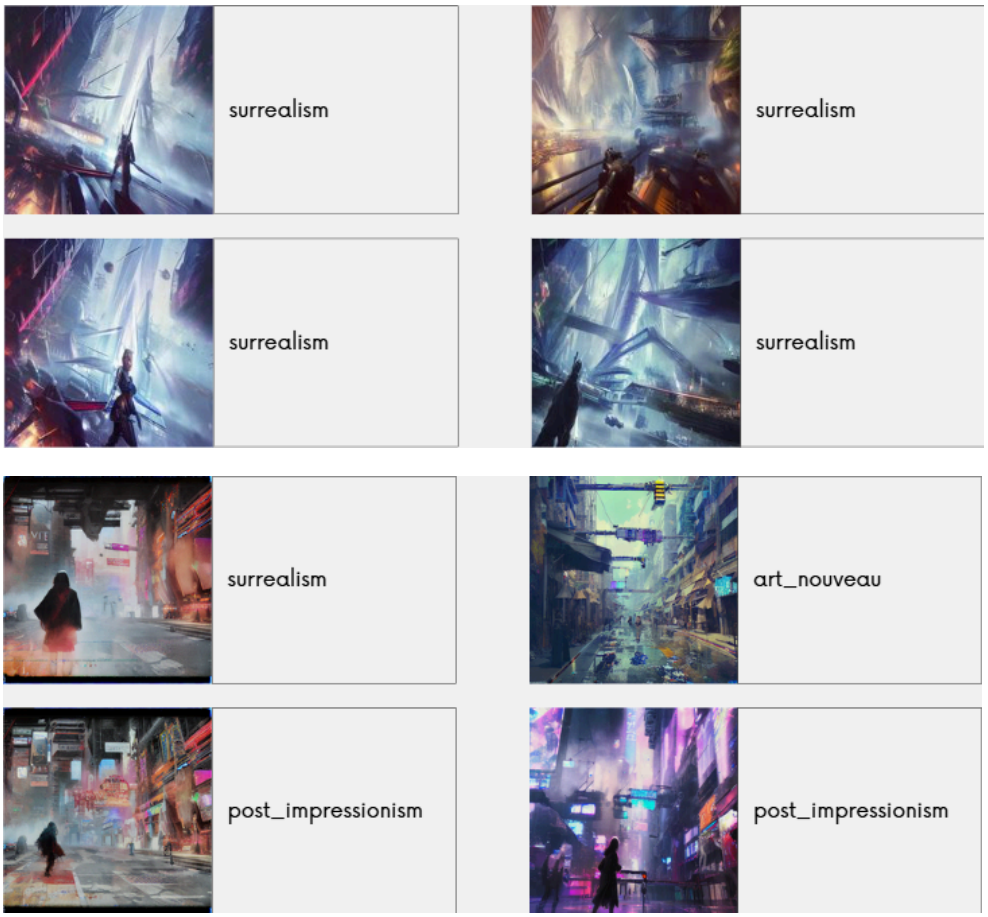

Figure 7: Comparative Analysis of Classification Results
The first image shows the classes assigned to the generated images when the original artwork was passed through Stable Diffusion, with classifications provided by the Hugging Face Art Classifier. The second image illustrates the classification results when the cloaked image was passed through the same model.

3. Visual Quality Metrics:

These results demonstrate that Hemlock successfully preserves visual integrity while effectively preventing unauthorized style replication.

Index	Image	RMSE	PSNR	SSIM
0	00bgv7gt.jpg	0.00194344029296189	53.9762677249997	0.997864198390648
1	13o09ong.jpg	0.00130077626090496	57.3797392186529	0.998402479536499
2	3ixgt5u5.png	0.0012355901999399	57.6797686743976	0.997421222603996
3	5ps9t2z5.jpg	0.00116225285455584	58.4816538556904	0.996824940875848
4	6t8vcaqm.jpg	0.00206777639687061	53.1055305479685	0.995493155191192
5	6tb4f28h.jpg	0.00174899713601917	54.7609668674563	0.995611962218897
6	78e7p6us.jpg	0.00147098244633525	56.3922318501946	0.996596634453357
7	89sum7zc.jpg	0.0020918280351907	53.2024511203906	0.998071184480384
8	8ej7x7t7.jpg	0.0012803558493033	57.6083212931567	0.998204840447008
9	8y10p9cm.jpg	0.00141105521470308	56.6633518386506	0.99864722534886
10	91yphwql.jpg	0.00202609878033399	53.586017788652	0.997492073859126
11	9npi2ngz.jpg	0.00113434821832925	58.5671981592786	0.99861493499339
12	cascuyfp.jpg	0.00136928306892514	57.1300834248718	0.998173126259249
13	dgqeinrv.jpg	0.00223078927956521	51.7036064824618	0.992665955526885
14	ebkh6ni1.jpg	0.00121469947043806	58.0609751095585	0.997188056899342
15	er4g9hds.jpg	0.001218251301907	58.018771395241	0.998403277026345
16	felk24li.jpg	0.00314706866629421	48.0596507196189	0.983990742051167
17	fr0981qj.jpg	0.0012432443909347	57.8187259910586	0.993369486368244
18	hdrrz1ys.jpg	0.00195657811127603	54.0589832863795	0.997758295116605
19	hk23r288.jpg	0.00155397376511245	55.9359870764431	0.997760928089488
20	ig7gdlxq.jpg	0.00173055881168693	54.925280711961	0.997924422290337
21	j9fvatre.jpg	0.00147843908052891	56.0927073681505	0.997600657567968
22	mk1uibzr.jpg	0.00149906065780669	56.0827128989415	0.998153758123861
23	nkpbgjjo.jpg	0.00113052933011204	58.6794866288441	0.998096270828174
24	nqqbwnoj.jpg	0.00213766261003911	52.7386592350598	0.993071699163168
25	orsn79ar.jpg	0.00201292452402412	53.3739654400045	0.989026653152886
26	p741ewu0.jpg	0.0008877664222382	60.8750098565966	0.996309340109318
27	pzwwp3i0.jpg	0.00187558389734476	54.4368645909194	0.997864035609131
28	r789wgiu.jpg	0.00312823173590004	49.7435909638443	0.995323226618715
29	rec3zgve.jpg	0.00223897793330252	52.5314463320957	0.99327753241299
30	rl2jhfnu.jpg	0.00118187570478767	57.7820816621187	0.996942439505021
31	rml0m9wa.jpg	0.00115558388642966	58.560107752549	0.998238621645058
32	sfx28oxe.jpg	0.00210435152985155	53.3053428425967	0.998065111607284
33	t61w1uvn.jpg	0.00234104692935943	52.4542332488615	0.997779192534328
34	tgevxz4e.jpg	0.00124297139700502	57.8935792238966	0.99630730587749
35	w25fc3i3.jpg	0.00108742562588304	58.6396321579758	0.998325634291302
36	w3utntba.jpg	0.00176292040850967	54.6373660504384	0.997854741748984
37	wmp9aodu.jpg	0.00102587754372507	59.5441854983711	0.998437606641549
38	xpnhsgy2.jpg	0.00225059315562248	52.1363913021972	0.990942523780782
39	xusf4q2m.jpg	0.00216196733526885	53.0690174234902	0.997167351226212
	Average	**0.00168104345648316**	**55.4922985903509**	**0.996381571111777**

Table 2: Original vs Cloaked Similarity Metrics

5.5 Results and Discussions

The performance of our model can be evaluated quantitatively by application of metrics like RMSE, PSNR and SSIM:

S.No	Metric	Average Values
1.	RMSE	0.00168104
2.	PSNR	55.492298
3.	SSIM	0.9963815

Table 3: Quantitative Test Results

The evaluation metrics demonstrate the effectiveness of the Hemlock model in maintaining the visual integrity of cloaked images while disrupting style replication.

1. **Root Mean Square Error (RMSE)**: The model achieved an average RMSE of **0.00168**, indicating minimal pixel-level differences between the original and cloaked images.
2. **Peak Signal-to-Noise Ratio (PSNR)**: An average PSNR of **55.49 dB** highlights the high fidelity of the cloaked images, ensuring that they remain perceptually similar to the original.
3. **Structural Similarity Index (SSIM)**: The SSIM score of **0.9964** reflects a near-identical structural similarity between the original and cloaked images, preserving luminance, contrast, and texture.

These results confirm that Hemlock introduces imperceptible perturbations to protect artistic styles while retaining the visual quality of the original artwork.

Visually the results of cloaking algorithm can be verified as follows :

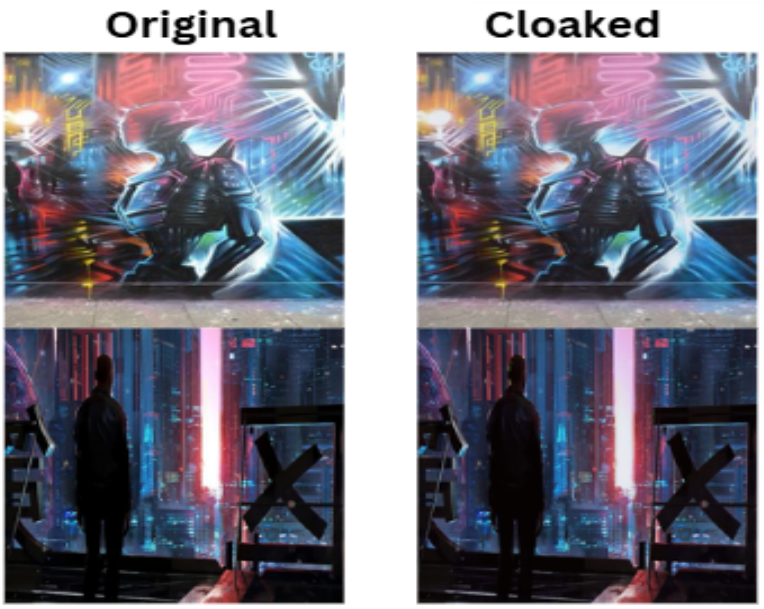

Figure 8: Column A shows the original image ; Column B shows the cloaked images generated by the Hemlock Model

5.6 Validation of Objectives

The objectives outlined for the Hemlock project were successfully validated as follows:

1. **Prevent Unauthorized Style Replication:**
    - Hemlock effectively cloaked artistic styles, making it difficult for Stable Diffusion and other AI models to mimic the original style. This was verified through differences in classification results from the Hugging Face ViT Art Classifier.
2. **Preserve Visual Integrity:**
    - The high PSNR and SSIM scores, along with a low RMSE, confirm that the cloaked images remain visually indistinguishable from the original, fulfilling the objective of maintaining aesthetic quality.
3. **Provide a Scalable and Practical Solution:**
    - Hemlock's architecture and processing workflow are computationally efficient and adaptable to various artistic styles, making it a practical tool for artists and art communities.
4. **Develop a Robust Cloaking Framework:**

- By leveraging cosine similarity-based feature integration and a combined loss function, Hemlock ensures robustness against style replication, achieving a significant advancement over existing methods like Fawkes and Nightshade.

The results conclusively demonstrate that Hemlock meets the stated objectives, offering an effective and innovative solution to protect artists' work in the era of generative AI.

CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

In this report, we presented Hemlock, a way to safeguard your artwork against text-to-image mimicry models like MidJourney and Stable Diffusion.The approach that we have adopted combines deep learning technique of neural network and a novel cloaking mechanism that has been incorporated in UNet Architecture.

Key Achievements:

1. The model we built effectively integrates the features from a style transferred image in the latent feature space of original artwork without compromising visual integrity but applying an effective cloak.
2. We have collected a diverse dataset from college societies and WikiArt for robust testing and fine-tuning of model parameters.
3. The cloaked image has been tested to ensure that it introduces minimal perturbations using a combined perceptual metric : LPIPS and RMSE.

While Hemlock provides a strong foundation for protecting artistic creations, ongoing improvements and adaptations will be essential to keep pace with evolving AI technologies. In conclusion, Hemlock represents a significant advancement in safeguarding artists' work in the digital age. Its technical innovation for artists seeking to protect their unique creative expressions from unauthorized imitation.

6.2 Reflections

Working on this project has been a transformative experience, fostering the development of various skills such as teamwork, time management, problem-solving, and technical decision-making. The collaborative nature of the project enhanced our ability to work cohesively as a team, ensuring the seamless integration of multiple components within the U-Net framework. Balancing technical challenges with real world applicability improved our multitasking and prioritization abilities, enabling us to deliver a robust and efficient solution.

This project has significantly enhanced our understanding of advanced machine

learning techniques, particularly in image processing and deep learning architectures. It provided invaluable insights into implementing depth-aware attention mechanisms, AutoEncoder -based networks, and optimizing frameworks.

On the academic front, this project allowed us to bridge the gap between theoretical knowledge and practical implementation. The iterative design, testing, and optimization processes reinforced the importance of meticulous experimentation and adaptive learning.

Looking ahead, this project inspires future aspirations to expand mimicry prevention tools.Potential avenues include enhancing the scalability of the web service for larger datasets, improving user security features, and refining the framework.

Overall, this project has not only achieved its objectives but has also left a lasting impact on our personal and professional growth. It serves as a stepping stone for future endeavours in image enhancement and related fields, equipping us with the confidence and skills to tackle complex challenges.

6.3 Future Plan

The Hemlock model has proven to be a robust and effective solution for safeguarding artistic styles against AI-generated mimicry while maintaining visual integrity. Looking ahead, several key directions will be explored to extend the capabilities and broaden the impact of the model.

Future efforts will focus on enhancing the efficiency of the model to make it even more practical for large-scale use. By streamlining the computational processes involved in the cloaking mechanism, the model can be optimized for faster performance and easier integration into diverse artistic workflows. Additionally, expanding the application of Hemlock beyond static images to video data will open new avenues for protecting animations and other dynamic media. This requires addressing unique challenges such as temporal consistency and processing efficiency, enabling the protection of a wider range of creative outputs.

Another critical area of exploration is the generalization of Hemlock across a broader spectrum of generative AI models, including popular systems like DALL·E and MidJourney. Evaluating and fine-tuning the model for compatibility with these platforms will ensure that artists are protected regardless of the AI tools being used. Furthermore, deploying Hemlock on scalable cloud platforms will make its functionality accessible to a larger audience, allowing artists to use the tool without the need for high-end hardware.

To complement these technical advancements, new avenues for evaluating the model's performance will also be explored. By incorporating advanced perceptual metrics, the evaluation process can provide deeper insights into the effectiveness of the cloaking mechanism while aligning with the evolving needs of artists and the creative industry.

These future plans aim to expand Hemlock's applicability and impact, ensuring it remains a valuable and accessible tool for artists to protect their unique styles in an ever-changing technological landscape.

PROJECT METRICS

7.1 Challenges Faced

During the development and testing of Hemlock, the following challenges were encountered:

1. **Dataset Curation**:
    - Collecting high-resolution artwork images and generating style-transferred versions required significant effort to ensure diversity and quality in the dataset.
2. **Feature Integration**:
    - Determining the optimal cosine similarity range for feature replacement to ensure effective cloaking without compromising the visual integrity was a key technical challenge.
3. **Computational Resources**:
    - The model's training and evaluation required GPUs with sufficient VRAM, making scalability difficult for larger datasets in constrained environments.
4. **Metric Optimization**:
    - Balancing perceptual similarity (LPIPS) with pixel-level fidelity (MSE, PSNR) during loss function optimization required careful tuning of hyperparameters.
5. **Testing Framework:**
    - Incorporating external tools such as Stable Diffusion and the Hugging Face Art Classifier into the testing workflow posed integration and compatibility challenges.

7.2 Relevant Subjects

The project draws from multiple subjects and technical domains, including:

- **Artificial Intelligence and Machine Learning:**Development of the deep learning architecture and implementation of loss functions.
- **Image Processing and Computer Vision**: Feature extraction, cosine similarity computation, and perceptual quality evaluation.

- **Data Science:** Dataset curation, preprocessing, and performance analysis.

7.3 Interdisciplinary Knowledge Sharing

The project leveraged expertise from various domains, including computer science, mathematics, and design, fostering a rich exchange of ideas:

- **Computer Science:** Team members specialized in machine learning and deep learning, contributing to the design and optimization of the model framework. Key roles included implementing U-Net,Cloak_Block and optimisation techniques.
- **Mathematics:** Applied statistical models, including PSNR, SSIM, and LPIPS metrics, to assess the quality, ensuring objective and quantifiable evaluation.

Task-Specific Knowledge Sharing:

- **Deep Learning Frameworks**: Shared knowledge on the use of U-Net architecture, VGG19, and other pre-trained models for feature extraction and reconstruction.
- **Adversarial Techniques**: Leveraged concepts from adversarial attacks (e.g., one-pixel attacks) to enhance the cloaking mechanism.

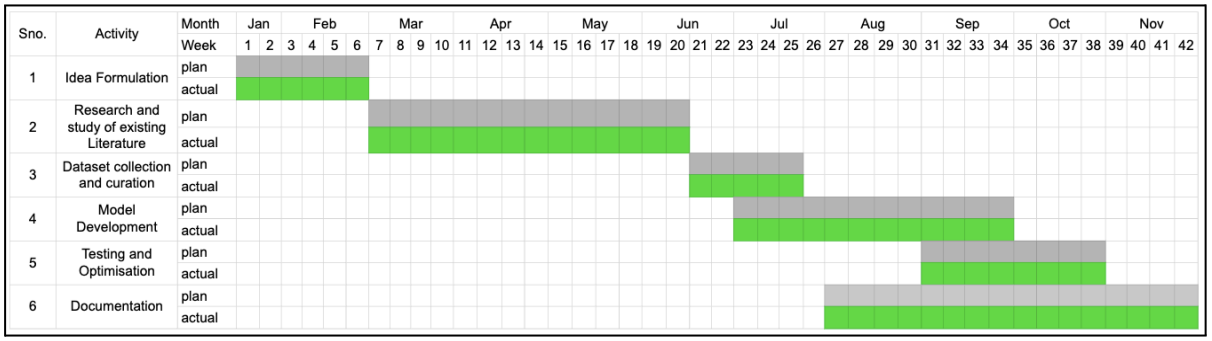

Figure 9: Gantt Chart

7.6 Student Outcomes Description and Performance Indicators

Outcome	Description	Performance Indicator
A	Apply knowledge of mathematics, science and engineering	Used metrics like MSE, RMSE, LPIPS, PSNR AND SSISM to evaluate model and applied deep learning principles to design hemlock
B	Collect and prune dataset	created a testing and training dataset by collecting artworks from college societies
C	Function on Multidisciplinary teams	Collaborated on tasks such as model development, dataset preparation and testing
D	Communicate effectively	Produced comprehensive reports and presentations
E	Use techniques, skills, and modern engineering tools necessary for engineering practise	Utilized tools like PyTorch, Kaggle and LPIPS Library for development, testing, a

Table 6: Student Outcomes Description and Performance Indicators (A-K Mapping)

7.7 Brief Analytical Assessment

Question 1: What sources of information did your team look into to come up with a list of potential project issues?

The team explored several sources to identify the core issues faced by artists in safeguarding their work against unauthorized replication by generative AI models. This included a detailed review of state-of-the-art research papers, articles, and documentation on diffusion models like Stable Diffusion, MidJourney, and DALL-E.

Existing solutions, such as Fawkes, Nightshade, and Glaze, were analyzed to understand their limitations in protecting artistic styles. Dataset sources such as artwork repositories and neural style transfer tools provided insights into the type of features generative AI models exploit. Regular consultations with our mentor also helped refine the scope and identify actionable project objectives.

Question 2: What analytical, computational, and/or experimental methodologies did your project team employ to find answers to the project's problems?

The project employed a combination of analytical, computational, and experimental methodologies.

- Analytical: We evaluated metrics such as RMSE, PSNR, and SSIM to assess the visual integrity and structural similarity of the cloaked images compared to the original artwork.
- Computational: The Hemlock model was built using a U-Net-inspired architecture with a custom CloakBlock layer, leveraging feature integration through cosine similarity. The model was trained on a high-resolution dataset of original and style-transferred images.
- Experimental: To test the robustness of cloaked images, both original and cloaked versions were fed into the Stable Diffusion model, and the outputs were analyzed using the Hugging Face Art Classifier to confirm disruption in style replication.

Question 3: Did the project demand demonstration of knowledge of fundamentals, scientific and/or engineering principles? If yes, how did you apply?

Yes, the project required a solid understanding of deep learning, image processing, and computer vision principles. Knowledge of convolutional networks, cosine similarity, and perceptual loss functions (e.g., LPIPS) was applied to design and implement the CloakBlock framework. Feature blending and reconstruction were achieved through encoder-decoder methods inspired by the U-Net architecture. Furthermore, metrics such as PSNR and SSIM were employed to evaluate visual quality, ensuring adherence to scientific standards. The project also demanded proficiency in software engineering principles to maintain modularity and scalability during model development.

Question 4: To manage design and production dependencies, how did your team share responsibilities and communicate scheduling information with others in the team?

Responsibilities were distributed based on individual expertise and aligned with project dependencies. Regular weekly meetings were conducted to discuss progress and synchronize efforts across modules. Tasks like dataset preparation, model development, testing, and documentation were tracked using project management tools such as Gantt charts. Team members maintained open communication channels to ensure timely delivery of milestones and to resolve cross-module dependencies effectively.

Question 5: For the duration of the project, what resources did you use to learn new materials that were not taught in class?

The team extensively referred to online resources, including research articles on diffusion models and adversarial techniques. The official documentation of libraries like PyTorch and tools like Stable Diffusion and Hugging Face Art Classifier provided technical guidance. Open-source platforms such as GitHub and Stack Overflow were instrumental in resolving development challenges. Insights from our mentor also played a crucial role, especially in refining the CloakBlock framework and tuning hyperparameters.

Question 6: Is the project making you grasp the need of utilizing engineering to address real-world issues, and might the project development be making you skilled with software development tools and environments?

This project highlighted the importance of engineering solutions to address real-world challenges, specifically protecting intellectual property and artistic integrity in the age of AI. The development of the Hemlock model demonstrated how deep learning and image processing techniques can be effectively applied to safeguard creative works. The project also provided hands-on experience with tools and environments such as PyTorch, Stable Diffusion, and the Hugging Face Art Classifier, enhancing our technical proficiency. Additionally, this work underscored the value of ethical engineering in creating systems that address critical societal issues.

APPENDIX A: REFERENCES

[1] Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., & Zhao, B. Y. (2023). Glaze: Protecting artists from style mimicry by text-to-image models. *arXiv preprint arXiv:2302.04222*

[2] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In the International *conference on machine learning* (pp. 8748-8763). PMLR.

[3] Zeng, Y., Pan, M., Just, H. A., Lyu, L., Qiu, M., & Jia, R. (2023, November). Narcissus: A practical clean-label backdoor attack with limited information. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security* (pp. 771-785).

[4] Shan, S., Ding, W., Passananti, J., Zheng, H., & Zhao, B. Y. (2023). Prompt-specific poisoning attacks on text-to-image generative models. *arXiv preprint arXiv:2310.13828*.

[5] Shan, S., Wenger, E., Zhang, J., Li, H., Zheng, H., & Zhao, B. Y. (2020). Fawkes: Protecting privacy against unauthorized deep learning models. In the 29th *USENIX security symposium (USENIX Security 20)* (pp. 1589-1604).

[6] Cherepanova, V., Goldblum, M., Foley, H., Duan, S., Dickerson, J., Taylor, G., & Goldstein, T. (2021). Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. *arXiv preprint arXiv:2101.07922*.

[7] Salman, H., Alaa, K., Leclerc, G., Ilyas, A., & Madry, A. (2023). Raising the Cost of Malicious AI-Powered Image Editing. Gradient science. 2022.

[8] Sara, U., Akter, M., & Uddin, M. S. (2019). Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study. *Journal of Computer and Communications*, *7*(3), 8-18.

[9] Ruta, D., Canet Tarrés, G., Gilbert, A., Shechtman, E., Kolkin, N., & Collomosse, J. (Year). DIFF-NST: Diffusion Interleaving For deFormable Neural Style Transfer. arXiv:2307.04157

[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (Year). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2103.00020