Unpaired Style Transfer from 3D Renders to Anime using Temporal-Aware GANs

Cole Feuer

April 23, 2025

Abstract

I present a novel unpaired style transfer method to convert 3D-rendered animation frames into anime-style images while preserving temporal stability and stylistic cohesion. My pipeline uses an 8-channel representation incorporating edge, depth, color, and blurred prior-frame information to enforce feature consistency. Initially trained via perceptual, pixel-wise, and style losses on anime datasets, my generator is later adapted using a CycleGAN-style dual-discriminator system to match stylistic features without paired ground truth. I highlight the motivations, implementation, and preliminary results of this approach, emphasizing its importance in practical anime-style animation production.

1 Introduction

The demand for high-quality, efficient animation production techniques has never been greater, particularly within the anime industry. Traditional anime production requires a substantial amount of manual labor, including hand-drawn keyframes and in-between frames that contribute to the expressive and stylized aesthetic of the medium. However, as production scales up and timelines tighten, studios increasingly rely on 3D rendering pipelines to streamline certain aspects of animation such as character movement, camera work, and environmental consistency.

3D renders are used in 2D animation not just for convenience but for their inherent strengths. They enable artists and studios to maintain precise control over lighting, camera angles, spatial positioning, and motion, allowing for the creation of reusable assets and the efficient production of complex scenes. However, these renders often lack the stylistic nuances that define anime — such as flat shading, exaggerated contours, and textured hatching — which makes them visually jarring when integrated into otherwise hand-drawn sequences. This motivates the need for a system capable of automatically transforming 3D-rendered frames into stylistically coherent 2D anime visuals without sacrificing the benefits of 3D pipeline efficiencies.

In this paper, I present a deep learning-based style transfer method designed to convert 3D-rendered frames into anime-style images. My method uniquely addresses two critical challenges in this space: the lack of paired training data and the need for temporal coherence across frames. By employing an 8-channel input representation and a two-stage training pipeline involving both perceptual training and adversarial adaptation, I aim to deliver temporally stable, visually faithful stylizations of 3D animation sequences.

2 Related Work and Background

Neural style transfer was first introduced by Gatys et al. (2016), who demonstrated that convolutional neural networks (CNNs) could be used to separate and recombine content and style from images. While visually striking, their method was computationally expensive and limited to static images. Subsequent works sought to improve efficiency and extend stylization to video frames. Johnson et al. proposed feed-forward networks for faster stylization, but struggled with temporal coherence.

The introduction of CycleGAN by Zhu et al. (2017) enabled unpaired image-to-image translation through the use of cycle-consistency and adversarial losses. This innovation was critical for tasks where paired training

data are unavailable, such as transforming real-world photos into artwork. However, direct application of CycleGAN to sequential frames often resulted in temporal instability, as each frame is processed independently without any awareness of adjacent frames.

Other methods attempted to tackle temporal consistency through architectural or loss-based means. ReCoNet introduced temporal losses using optical flow and occlusion maps, while approaches like vid2vid adopted generative frameworks that explicitly modeled temporal dependencies. Despite progress, many of these techniques either require paired data, are not specific to the anime domain, or do not generalize well to 3D render inputs.

Current anime stylization methods have often failed to achieve a balance between visual authenticity and temporal coherence. Many rely on static image techniques that result in noticeable flicker or inconsistency when applied to video, making them unsuitable for production-quality animation. Others achieve stylistic fidelity only by using paired datasets, which are rare or expensive to obtain for anime. Additionally, models trained on general artistic styles frequently lack the nuance and structured abstraction characteristic of anime, resulting in oversaturated or incoherent outputs.

3 Methodology

My method consists of two key stages: a supervised pretraining phase for style and content learning, followed by an adversarial adaptation phase for domain translation. Central to both phases is my use of an enriched 8-channel input that captures scene structure and temporal context.

3.1 Multi-Channel Input Representation

To address the need for both spatial detail and temporal coherence, I transform each 3D render frame into an 8-channel input tensor. This includes three color channels (RGB), a depth map normalized to scene scale, an edge map extracted using a Canny detector or learned filters, and a temporally blurred prior output frame. Including the prior frame allows the model to learn transition patterns and reinforces temporal stability, especially under dynamic motion and lighting conditions.

This design encourages the generator to preserve scene semantics and motion continuity across frames. The depth and edge channels encode structural boundaries and spatial relationships, which are critical for accurate contour stylization and occlusion management. The blurred prior frame acts as a soft memory, gently nudging the generator toward temporal coherence without enforcing hard constraints that may limit stylization.

3.2 Stage 1: Perceptual Pretraining of Generator

The first stage focuses on training a U-Net-based generator to learn style mappings from individual frames. Using anime datasets composed of stylized frames, I train the generator to produce outputs that match the statistical properties of real anime art.

I use several losses in this stage:

Pixel Loss: An L_1 loss between the generated output and target anime frame to enforce low-level accuracy.

Perceptual Loss: A feature-level loss using VGG-19 activations from intermediate layers. This loss guides the generator to align with the high-level semantics of anime imagery.

Style Loss: Computed as the Gram matrix difference between feature maps of the generated and target images. This loss captures stylistic attributes such as texture, brush stroke patterns, and color harmony.

This pretraining ensures that the generator is initialized with strong stylistic priors before being adapted to unpaired 3D data.

3.3 Stage 2: Unpaired Adaptation with Cycle-Consistent GAN

Once the generator is pretrained, I adapt it for use with 3D renders using a CycleGAN-inspired architecture. I introduce two PatchGAN discriminators: one to distinguish between real and generated anime frames, and the other to enforce reconstruction fidelity by mapping back to the original render domain.

The core components of this stage include:

Cycle-Consistency Loss: This loss ensures that when an input is mapped to the target domain and then back again, it retains the original content. It prevents mode collapse and promotes semantic consistency.

Identity Loss: This encourages the generator to produce identity outputs when the input is already in the target domain. It helps stabilize training and reduces over-stylization.

Adversarial Loss: I train the generator to fool the discriminator, while the discriminator learns to distinguish fake anime outputs from real ones. This loss sharpens textures and improves realism.

This unpaired training framework allows me to leverage large anime image corpora and 3D render datasets without requiring alignment or correspondence.

4 Results and Analysis

As of the current stage of development, I have completed the perceptual pretraining phase. Visual inspection of stylized outputs indicates a high level of temporal coherence, with minimal flickering across sequential frames. The incorporation of the blurred prior frame as an input channel appears to play a central role in this stability.

However, the degree of stylistic transformation remains conservative. While some anime-style textures and color adjustments are visible, the outputs lack the bold stylistic transformations characteristic of fully stylized anime. I believe this limitation stems from the absence of strong distribution-matching forces, which I expect to address in the second training stage.

The CycleGAN-based adversarial adaptation phase is currently in progress. I expect this phase to significantly enhance the visual fidelity and stylistic alignment of the outputs. Particular attention will be given to assessing whether adversarial training compromises the temporal smoothness established during pretraining.

5 Discussion

This method introduces several important contributions to the field of style transfer for animation. First, I demonstrate that a carefully constructed 8-channel input representation can guide neural networks to retain semantic and temporal continuity without explicitly modeling temporal loss. This is particularly valuable in animation settings where the absence of flickering is essential.

Second, by relying on unpaired training data, I make the method far more scalable than supervised alternatives, which require laborious frame-by-frame labeling. By adopting a cycle-consistent adversarial framework, I bridge the gap between the stylistic domain of anime and the structural domain of 3D renders.

There are limitations, however. The memory cost of processing 8-channel inputs is non-trivial, especially for high-resolution sequences. Additionally, the blurred prior frame approach, while effective at reducing flicker, may suppress fine details and result in overly smoothed transitions. Finally, as with most GAN-based methods, training can be unstable and sensitive to hyperparameter tuning.

6 Conclusion and Future Work

I have presented an innovative method for unpaired anime-style transfer from 3D-rendered animation frames, emphasizing temporal coherence and stylistic authenticity. Preliminary results validate the use of structural and temporal input representations to stabilize frame-to-frame transitions, while future adversarial training is expected to improve stylistic strength.

Future work will focus on several directions. I plan to integrate explicit temporal consistency losses, such as those based on optical flow or feature tracking. I also aim to quantitatively evaluate flicker levels and style accuracy using perceptual metrics and user studies. Finally, I will experiment with multiple anime style datasets to enable broader generalization and possibly introduce a conditional mechanism to control the style output.

Ultimately, I hope this work contributes to the vision of an automated, scalable pipeline for high-quality anime-style animation generation from 3D content, blending the strengths of both domains through deep learning.