CV + GAN Project

One-Shot Image to Video Translation with Facial Motion Capture

FastAnim8

Team Members: Sahil Nagaralu,
Sanskar Jadhav, & Roshan Yadav (074, 076, 130)

Introduction

Our algorithm animates a given image with motion captured from a human speaking

Problem Statement

Our aim is to transform an anime character face image into a video with human motion video exhibiting similar movements and emotions as best as possible in an custom background. The applications of this project would be massive, as it could completely alter the manner in which animation studios around the world produce animated sequences. With our model, there would no longer be any need of drawing each frame of animation.

Currently similar products in the market focus on motion of the entire body (eg. ControlNet, WarpFusion) with errors of inconsistency in movement, and blurring of face details because facial landmarks on animated faces do not share the same mesh structure as on humans. Our solution involves working with animated character faces with modifications on dlib face landmarks so as to replicate human face movement and motion.

“It mimics the effects of green screen and motion capture, while being inexpensive.”

Our Objectives

Stage 1

Replace the Background

Image segmentation of character outline to remove background

Stage 2

Replace the Human

AI character in same pose using dlib 81 landmarks, Super Resolution using Custom SRGAN

Stage 3

Finetune the Results

Scaling warped face to match original, Matching empty spaces around face to background

Existing Products in Field

Video to AI Animation Tutorial For Beginners: Stable WarpFusion + Controlnet | MDMZ - YouTube

WarpFusion
+
ControlNet

Literature Review

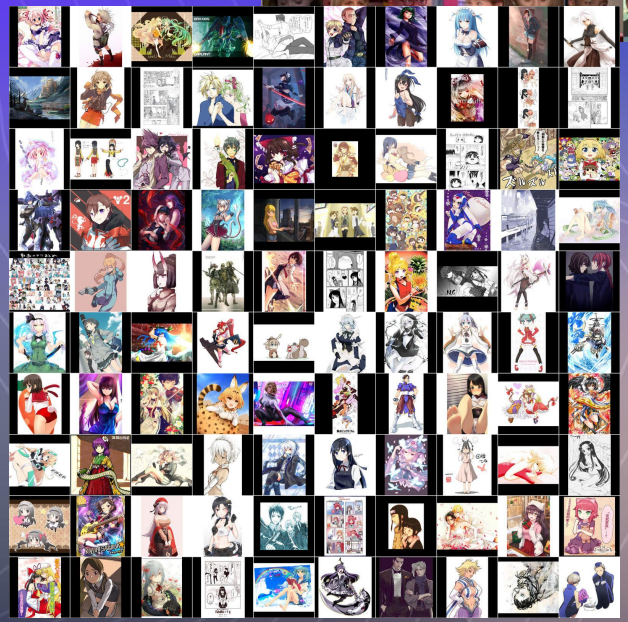
Works	Methodology	Pros	Cons	Links
DeepVideoPaint	Uses Generative Adversarial Networks (GANs) to replace objects in real-time video.	Achieves high-quality results, good at maintaining temporal coherence.	Limited object variety, struggles with complex object interactions.	https://arxiv.org/pdf/2303.09290
Spatio-Temporal Object Manipulation in Videos	Leverages attention mechanisms and transformers to manipulate object appearance and motion in videos.	Handles complex object manipulations, allows for interactive control.	Requires large training datasets, high computational cost.	https://arxiv.org/pdf/2308.04549
Neural Object Rewinding	Employs a recurrent neural network (RNN) to generate frames where objects appear to rewind their motion.	Offers a unique visual effect, good for creative applications.	Limited object types, works best with simple motions.	https://arxiv.org/pdf/2212.07626
Deep Exemplar-Based Image Animation	Utilizes exemplar images to guide the animation of objects within a video sequence.	Enables control over specific object animations, handles partial occlusions.	Requires high-quality exemplar images, can be computationally expensive for long videos.	https://arxiv.org/abs/2211.13227
Video in Painting (VIP)	Transforms real-time video into artistic styles using convolutional neural networks (CNNs).	Provides artistic flair to live video, allows for various style options.	Limited control over specific objects, may introduce artifacts in complex scenes.	https://arxiv.org/abs/1905.02884
Semantic Image Editing with Image Inpainting	Applies image inpainting techniques for object manipulation in static images, potentially adaptable to video.	Proven effective for basic object removal/replacement, conceptually applicable to live video.	Not specifically designed for video, may struggle with frame-to-frame coherence.	https://arxiv.org/abs/1607.07539

Deep Video Inpainting	Investigates deep learning methods for video inpainting, potentially applicable to object manipulation.	Offers a foundation for future video object manipulation techniques.	Primarily focused on inpainting missing regions, requires further development for object manipulation.	https://arxiv.org/abs/1905.01639
Learning to Segment Moving Objects in Videos	Develops algorithms for segmenting objects in videos, a crucial step for object manipulation.	Improves object segmentation accuracy, aiding future manipulation techniques.	Focused on segmentation, not manipulation itself, further research needed.	https://arxiv.org/abs/1712.01127
Mask R-CNN	Proposes a deep learning model for object detection and segmentation, useful for identifying objects in live video.	Offers real-time object detection capabilities, valuable for pre-processing in manipulation.	Not specifically designed for manipulation, requires additional steps for animation.	Mask R-CNN: https://arxiv.org/abs/1703.06870
FlowNet 2.0: Deep Learning for Optical Flow Estimation	Introduces a deep learning architecture for estimating optical flow, essential for understanding object motion in video.	Improves accuracy of optical flow estimation, facilitating realistic object animation.	Focused on flow estimation, not manipulation itself, needs integration with other methods.	https://arxiv.org/abs/1612.01925
Thin-Plate Spline Motion Model for Image Animation	Thin-Plate Splines & Multi-Resolution Masks for Unsupervised Image Animation	Enables flexible animation, handles occlusions, learns without needing labels	May struggle with large pose differences, increases model complexity, requires more training data.	https://arxiv.org/abs/2203.14367
Animating Arbitrary Objects via Deep Motion Transfer	Deep Motion Transfer with Keypoint Detection & Heatmaps	Versatile object application, Motion capture, Independent motion control	Training data dependence, Computational cost, Black box nature	https://arxiv.org/abs/1812.08861
Latent Image Animator: Learning to Animate Images via Latent Space Navigation	Variational Autoencoder (VAE) for style transfer between images and videos	Transfers animation style from videos to images, potentially applicable for creating anime-styled animations	May struggle with complex scenes and maintaining object consistency during style transfer	https://arxiv.org/abs/2203.09043
Structure-aware Video Style Transfer with Map Art	Deep Neural Network with attention mechanism	The "map art" style transfer can provide a base for building anime-style visuals. Elements like flat colors and sharp lines are common in anime aesthetics.	While the technique alters the visual style, it doesn't directly address character animation, a crucial aspect of anime.	http://graphics.csie.ncku.edu.tw/Tony/papers/ACM_TOMM_MapArt_Video_Accepted.pdf
Anime-Like Motion Transfer with Optimal Viewpoints	Anime-Like Motion Transfer with Optimal Viewpoints	Effectively extracts suitable poses for lower frame rates, Relieves redundancy in motions due to physical speed constraints	Inconsistent emphasis on speed, Difficulty in identifying character positions due to monotonous background	https://dl.acm.org/doi/pdf/10.1145/3550082.3564212

Datasets

Vox-Celeb Dataset - VoxCeleb is an audio-visual dataset consisting of short clips of human speech, extracted from interview videos uploaded to YouTube.

Danbooru2021: A Large-Scale Crowdsourced & Tagged Anime Illustration Dataset - Only Selected ones for the specific needs



Methodology

01

Background Substitution

- Preprocessing - Histogram Equalization
- Canny Edge Detector - To outline character
- Contour - Made Contours (Mask) of the character
- Applied Inverted Mask for new background

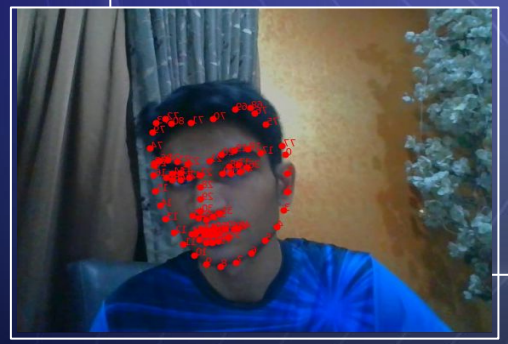


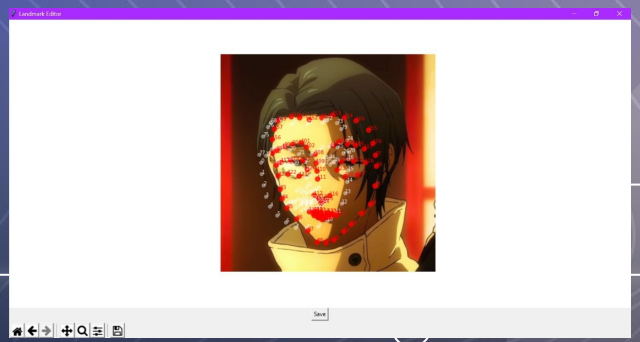

Methodology

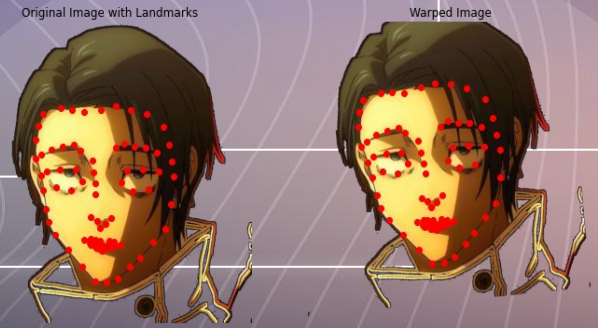
02

Face Landmark Detection and Mapping

- Preprocessing - Histogram Equalization
- Canny Edge Detector - To outline character
- Made contours (mask) of the character
- Applied Inverted Mask for new background
- Custom GUI for annotating landmarks on character face
- Warping and aligning face to new landmark positions







Methodology

03

Aligning of Frames for Conversion into Video

- Placing transparent moved character face onto original
- Inpainting of black regions left when character moves
- Blurring of edges for capturing motion blur
- Testing movement of individual facial features
- Setting a boundary box region of interest near the face

```
{'encoding': 'ascii', 'confidence': 1.0, 'language': ''}
Processed image saved as /kaggle/working/0001.jpg
{'encoding': None, 'confidence': 0.0, 'language': None}
Processed image saved as /kaggle/working/0002.jpg
{'encoding': 'ascii', 'confidence': 1.0, 'language': ''}
Processed image saved as /kaggle/working/0003.jpg
{'encoding': None, 'confidence': 0.0, 'language': None}
Processed image saved as /kaggle/working/0004.jpg
{'encoding': 'ascii', 'confidence': 1.0, 'language': ''}
Processed image saved as /kaggle/working/0005.jpg
{'encoding': None, 'confidence': 0.0, 'language': None}
Processed image saved as /kaggle/working/0006.jpg
{'encoding': 'ascii', 'confidence': 1.0, 'language': ''}
Processed image saved as /kaggle/working/0007.jpg
{'encoding': None, 'confidence': 0.0, 'language': None}
```

Original Image

Cropped Boundary Box

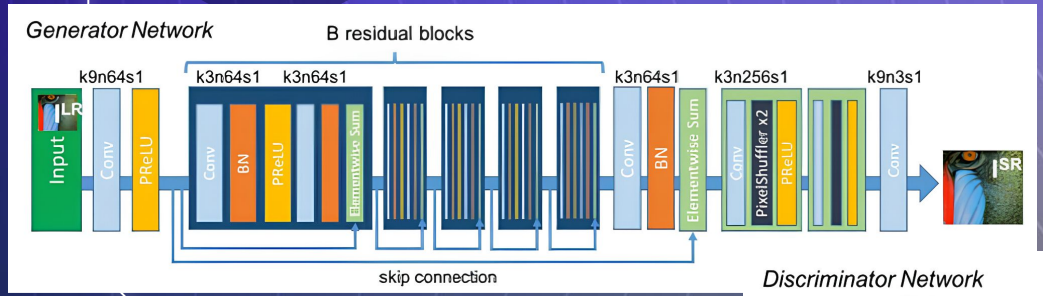
Resulting Image

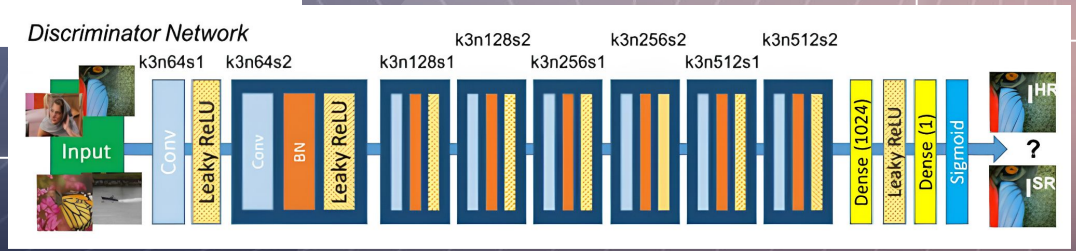
Methodology

04

Super-Resolution of Each Frame in Video

- Custom Super Resolution Wasserstein GAN specially trained on selected Anime images.
- Increase the resolution of the frames then compress them into a video.





Final Experimental Results

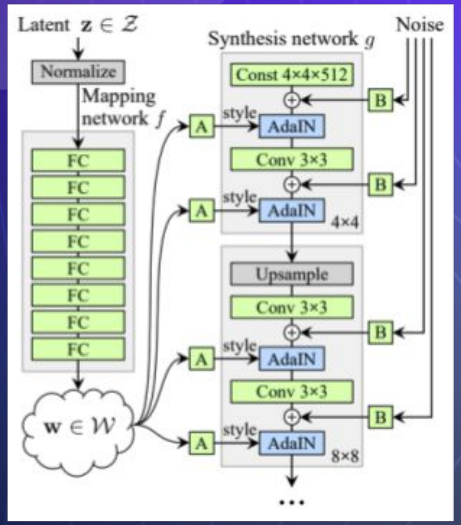
Runs with any human, any custom background,
and any character!

Final Experimental Results

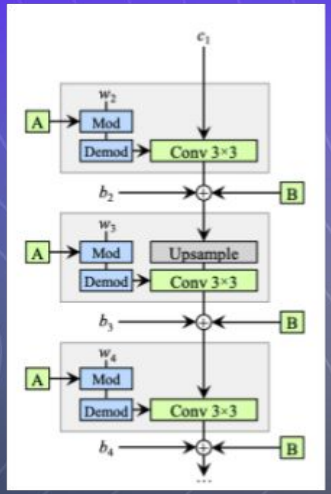
Before and After applying SR GAN

Final Experimental Results

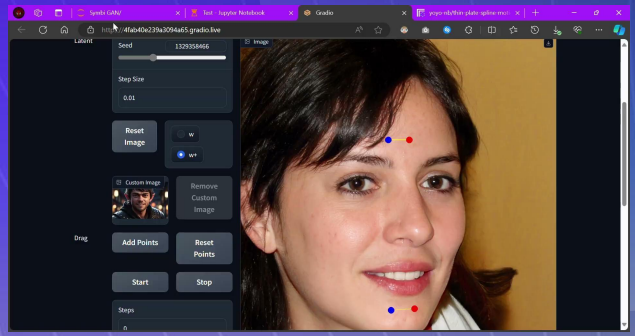
Downloading and Working with Drag GAN
https://arxiv.org/abs/2305.10973



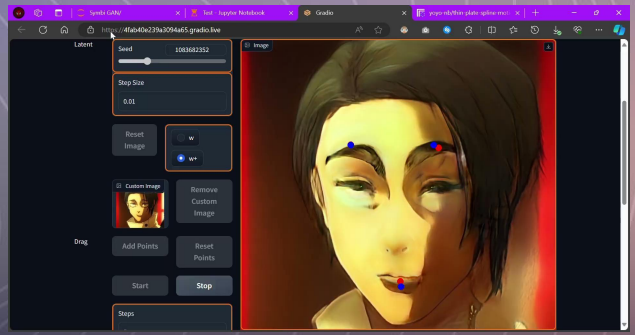
StyleGan



StyleGan-2's Synthesis network



Works well on given training dataset of real faces
Not so well on custom input of animated faces

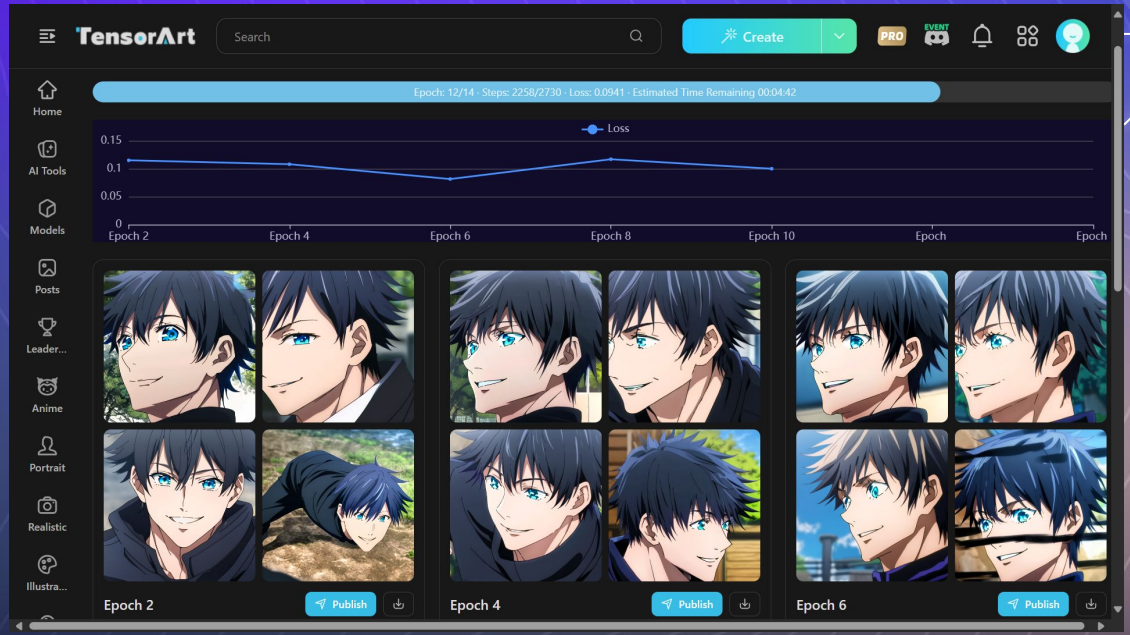


Final Experimental Results

Working with Diffusion Models (Anything v5) and Custom Training of LoRA for generating anime characters

14 epochs
Cosine Learning Rate Scheduler (changes every 2 epochs)

AdamW optimizer
0.0005 initial learning rate



Prompt:

1boy, solo, smile, short hair, blue eyes, male focus, black jacket, outdoors

Final Experimental Results

Final Output from Training of LoRA

Goal was to match Jujutsu Kaisen’s art style without an explicit mention

Final Experimental Results

*New: Using ControlNet to recreate human faces from animated

Conclusion & Future Work

Our goal is to contribute by creating datasets which have annotated detailed face landmarks on character faces that share dynamic features (e.g. larger eyes, different shaped hair), as well as generating more characters with face poses and varying perspectives.

Another future scope is to add audio to the video using AI tools for voice synthesizing, wherein from a sample of the character's voice (from the TV show or movie), we can replicate their manner of speaking to say any custom text we input.

Tools Used

Drag Your GAN: Colab Support for Installing and Deploying Online
https://github.com/camenduru/DragGAN-colab

Training of Custom LoRA and ControlNet Img2Img Generation
Tensor.art : Online service for hosting models and image generation

Custom built GUI for landmark annotation
dlib_face_landmarks_81.dat file used

References

1. DeepVideoPaint
2. Spatio-Temporal Object Manipulation in Videos
3. Neural Object Rewinding
4. Deep Exemplar-Based Image Animation
5. Video in Painting (VIP)
6. Semantic Image Editing with Image Inpainting
7. Deep Video Inpainting
8. Learning to Segment Moving Objects in Videos
9. Mask R-CNN
10. FlowNet 2.0: Deep Learning for Optical Flow Estimation
11. Thin-Plate Spline Motion Model for Image Animation
12. Animating Arbitrary Objects via Deep Motion Transfer
13. Latent Image Animator: Learning to Animate Images via Latent Space Navigation
14. Structure-aware Video Style Transfer with Map Art
15. Anime-Like Motion Transfer with Optimal Viewpoints