
Thinking Between the Lines: Guided 2D Animation with Generative Adversarial Networks

Evan Casey
evan.casey92@gmail.com

Harry Teitelman
hst472@gmail.com

Abstract

In this work we present an approach for generating colorized 2D animated video from line sketches and a single reference image. We pose this problem as per-frame image-to-image translation problem conditioned on both the line sketches and a reference image with spatio-temporal smoothing. We demonstrate the effectiveness of our approach on a dataset of professional hand-drawn animation data containing diverse textures, color mappings, and shapes. Our video demo can be found at <https://vimeo.com/297648691>.

1 Introduction

The process of creating traditional hand-drawn animation is two-fold. First, the animator draws each frame in the video as a simple line sketch to ensure the overall shape and movement across frames is correct. Second, using the line sketches as a guide, the animator fills each frame with the color and texture, yielding the final result. While hand-drawn animation produces stunning results, this process is labor intensive and requires many years of experience to do well.

Our work seeks to offer a more efficient way of producing 2D animation in which a generative model is used to assist the artist by automatically translating the line sketches into filled frames. The network receives as input the hand-drawn line sketches and a *single* reference frame conveying the style of the desired filled images and adds the desired color and texture.

2 Related Work

Recently, significant progress has been made in photorealistic image translation and video translation. In particular, Isola et al. (1) showed success in translating edge maps to handbags and other objects. Chan et al. (4) showed results on translating pose detections to photorealistic renderings. Jang et al. (7) attempted to use an reference image and motion information for conditional video prediction. However, our work is the first of our knowledge to combine the use of separate conditional inputs for fill style (the reference image) and overall structure (the line sketch) for video translation. Moreover, this work is the first, to our knowledge, to apply a generative model to 2D animation.

3 Model Description and Dataset

Given a source sequence of line sketches and a single filled reference image, our goal is to generate a new sequence of frames containing filled images in the style of the reference image that correspond with the shape and structure of the line sketches. Since the desired style may change many times across the entire animation, we split the training data into multiple scenes with a single reference frame chosen to represent the style for the entire scene. The network must learn to use the information in the reference frame as a guide to find the correct mapping from line sketch to filled color and texture, of which many different variations exist in the data.

Our dataset consists of 600 frames from roughly 2 minutes of animation from the film “Cats Dogs & Rats”¹ animated by Harry Teitelman. We split the frames across 28 training scenes and 14 test scenes, each containing a variable number of images between 5 and 40. We augment the data using random cropping and flipping. The image resolution used at test time is 128 x 256.

3.1 Generator

The generator has three conditional inputs: the line frame x_t and the reference frame y_r which is held constant for the entire scene, and the generated image at the last timestep $G(x_{t-1})$. We employ a similar generator architecture to the one proposed in CycleGAN (3) except that each conditional input is passed through 3 separate convolutional layers before being concatenated along the channel dimension and passed to the remaining layers. We train the network using a combination of L1 loss on the ground truth filled images and a learned loss function comprised of the style and smoothness discriminators (described below). We use the LSGAN learning objective as described in (5). Full training details for all networks can be found in the Appendix.

3.2 Style Discriminator

The style discriminator learns to discriminate between generated and real images based on the image’s adherence to the style of the reference image and structure of the line sketch. It accepts the following images as input: the generated filled image $G(x_t)$, the line sketch x_t , and the reference image y_r . Both discriminators follow the design of the 256 × 256 discriminator in pix2pix (2) with the exception that each conditional input is passed through 3 separate convolutional layers before being concatenated along the channel dimension and passed to the remaining layers. Since the reference image has a different global structure from the line sketch and generated output, we have the discriminator look at the whole image instead of image patches.

3.3 Smoothness Discriminator

The smoothness discriminator evaluates the generated filled image on the basis of its temporal coherence with the generated image at the last timestep. It accepts the following images as input: the generated filled image at the current timestep $G(x_t)$ and the generated filled image at the previous timestep $G(x_{t-1})$. The network architecture is identical to the style discriminator.

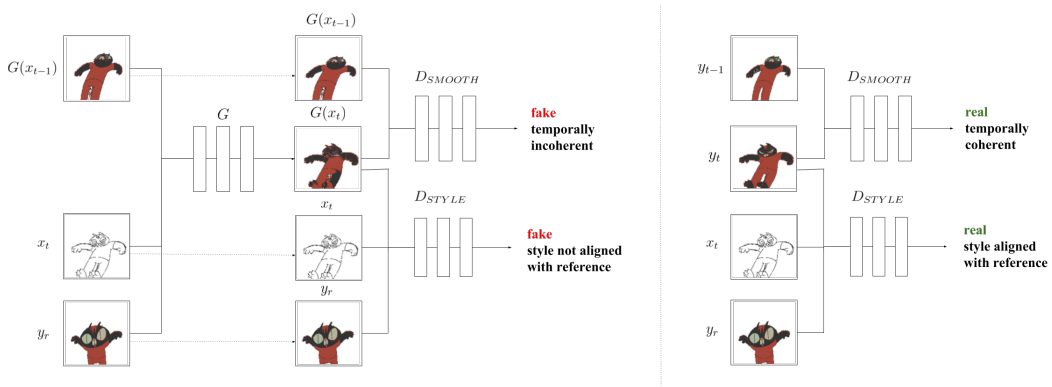


Figure 1: Flow diagram of our network.

4 Discussion

In this work, we propose a generative model for translating line sketches to filled 2D animations, guided by a single reference image. We show that this enables the model to generalize to many different possible fill styles in a single model. As an creative tool, this could enable animators to create work much faster while still retaining the aesthetic of hand-drawn animation.

¹The original version of the video can be viewed at <https://www.youtube.com/watch?v=JhEOImokCcM>.

References

- [1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio. Generative Adversarial Networks, 2014; arXiv:1406.2661.
- [2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou and Alexei A. Efros. Image-to-Image Translation with Conditional Adversarial Networks, 2016, CVPR 2017; arXiv:1611.07004.
- [3] Jun-Yan Zhu, Taesung Park, Phillip Isola and Alexei A. Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, 2017; arXiv:1703.10593.
- [4] Caroline Chan, Shiry Ginosar, Tinghui Zhou and Alexei A. Efros. Everybody Dance Now, 2018; arXiv:1808.07371.
- [5] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang and Stephen Paul Smolley. Least Squares Generative Adversarial Networks, 2016; arXiv:1611.04076.
- [6] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz and Bryan Catanzaro. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs, 2017; arXiv:1711.11585.
- [7] Yunseok Jang, Gunhee Kim and Yale Song. Video Prediction with Appearance and Motion Conditions, 2018; arXiv:1807.02635.

