manGANime: Future Video Prediction with Conditional GANs

Stephen Huan, Luke Thistlethwaite

TJHSST Computer Systems Lab, 2020–2021

Introduction

A class of neural networks called generative adversarial networks (GANs) are now able to generate images of astounding quality and resolution [1]–[3]. However, these successes in unsupervised *image* generation have not necessarily translated to *video* generation [4]. Motivated by these shortcomings, we propose replacing the real world with the simplified domain of cartoon video produced in Japan, or *anime*.

Our reasons for focusing on anime are multifaceted: first, there is a large volume of unlabeled anime on the internet. Secondly, many assumptions existing algorithms [4], [5] make are justified in anime *a priori*, but not necessarily in the real world. Finally, there is a huge demand for anime but a shortage of labor, causing animators to work long hours for low pay. Machine learning systems like the ones already used by Sony can automate tedious work, allowing animators to focus on the creative aspects.

However, anime is not so abstract as to destroy any semblance of the real world. For example, the object detection model YOLOv4 can be directly applied to anime.

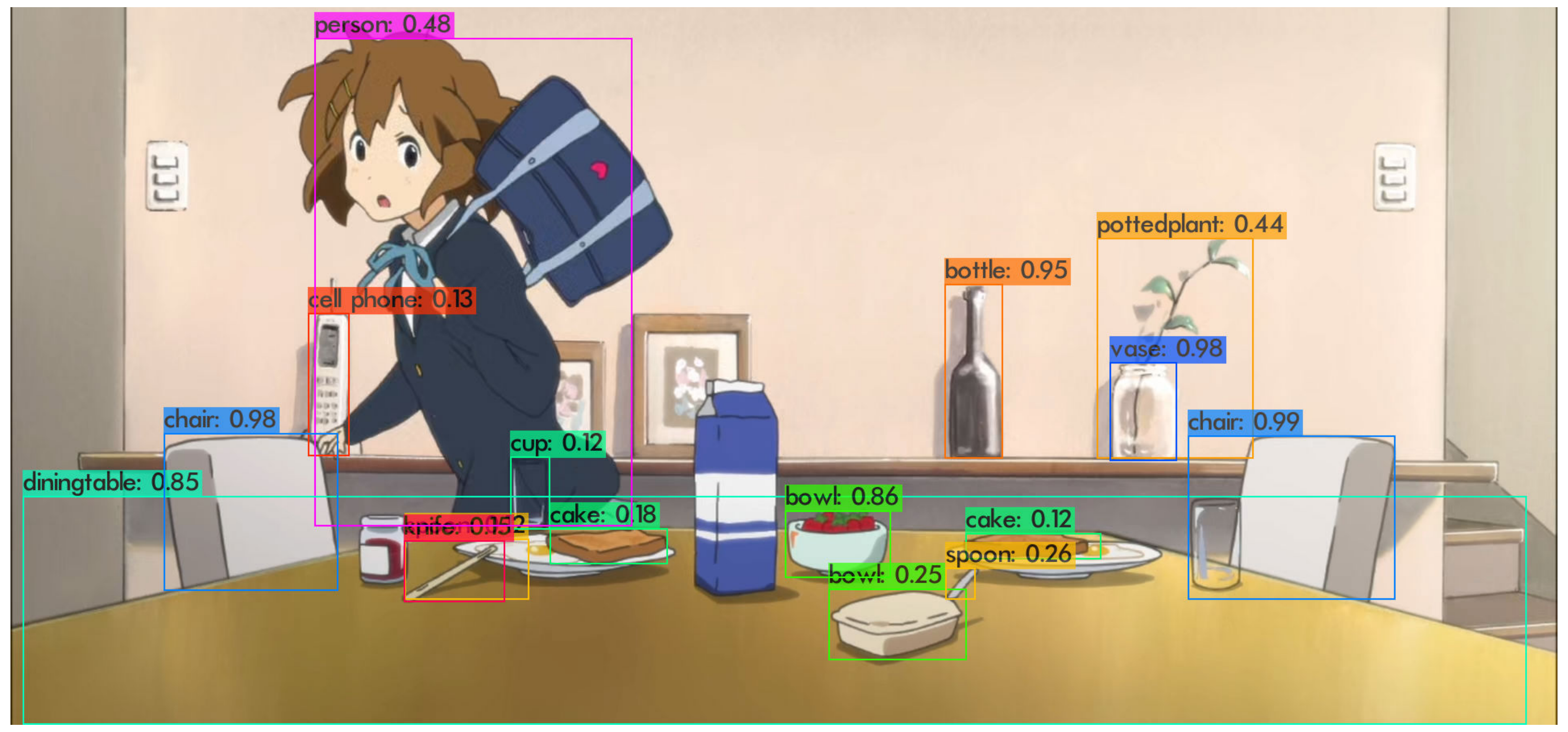


Figure 1: YOLOv4-CSP on the anime *K-On!*. See https://youtu.be/i8jCzh9nWc4 for a demonstration.

Animation is expensive, so anime is often produced with reference to a comic book, or *manga*. Certain anime will follow their source very precisely, as shown in Figure 2.

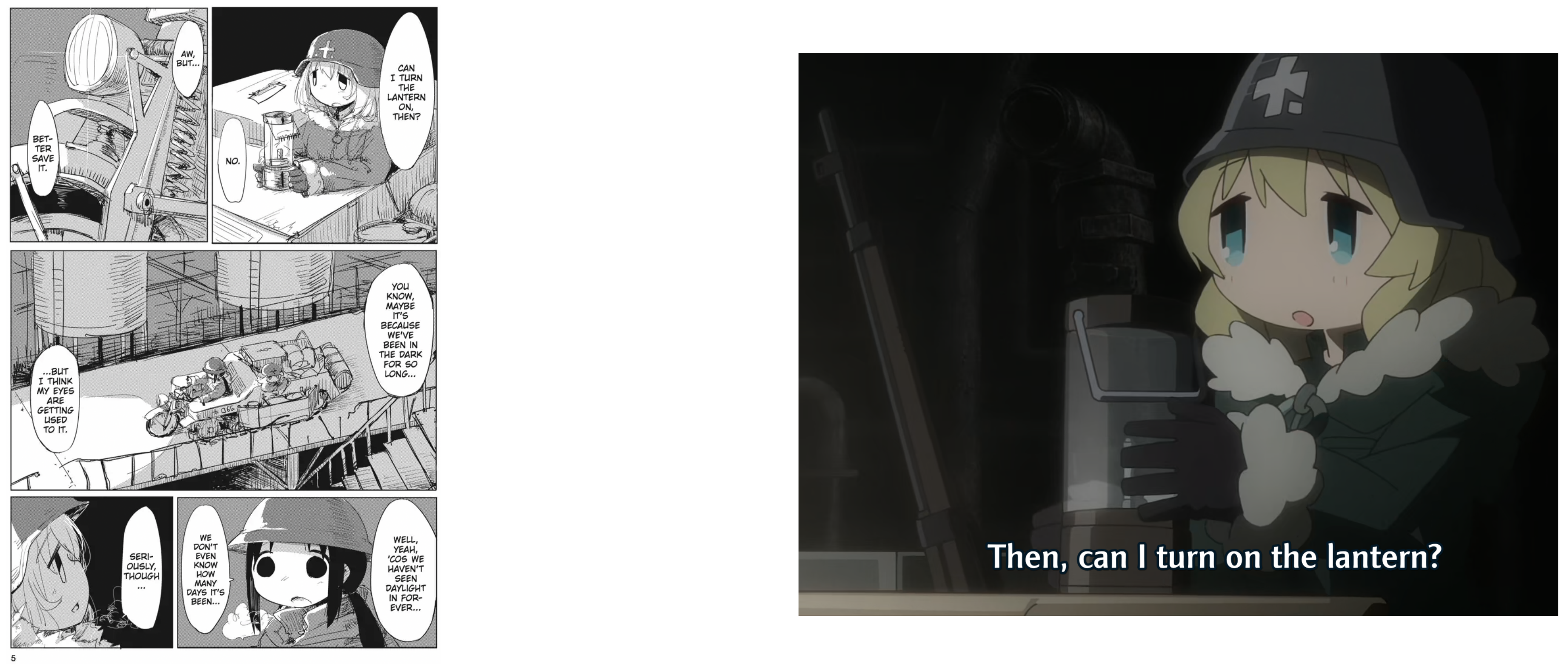
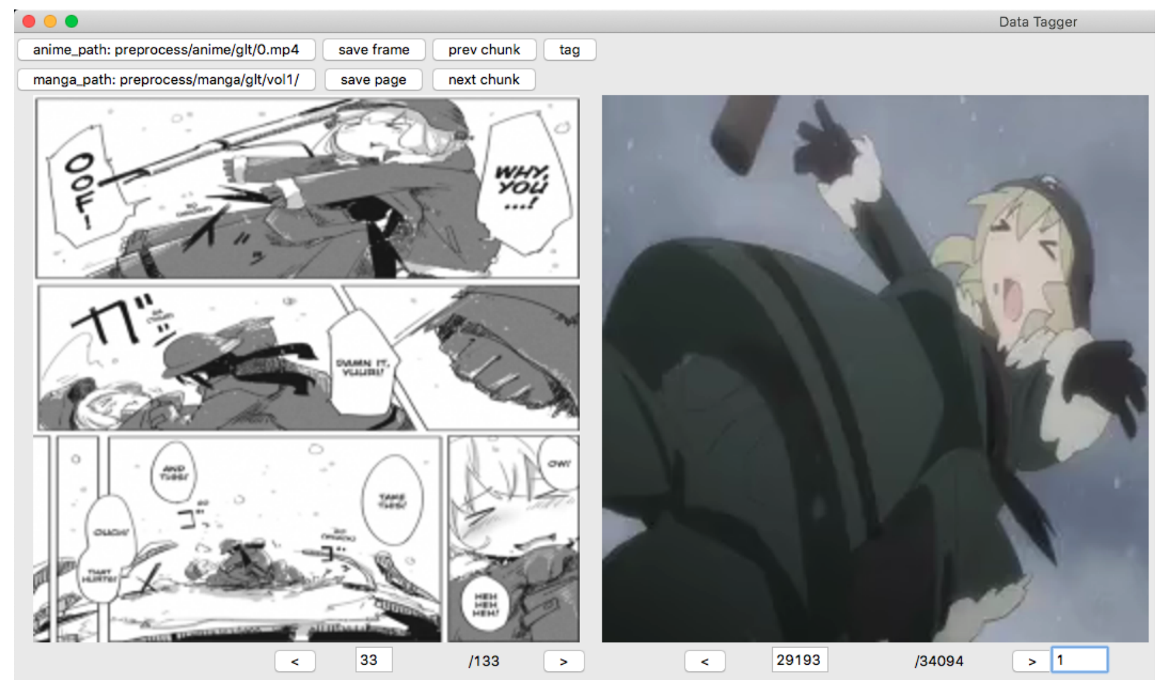

CAN I TURN THE LANTERN ON, THEN?
NO.
AW, BUT...
BETTER SAVE IT.
YOU KNOW, MAYBE IT'S BECAUSE WE'VE BEEN IN THE DARK FOR SO LONG...
...BUT I THINK MY EYES ARE GETTING USED TO IT.
WELL, YEAH, 'COS WE HAVEN'T SEEN DAYLIGHT IN FOREVER...
WE DON'T EVEN KNOW HOW MANY DAYS IT'S BEEN...
SERIOUSLY, THOUGH ...
5
Then, can I turn on the lantern?

Figure 2: The high correspondence between manga (left) and anime (right) in *Girls' Last Tour*.

For manga to be a useful conditioning signal, it should correspond well with its anime. Formally speaking, we define a *page function* $f : \mathbb{Z} \to \mathbb{Z}$ that maps an anime frame index to its corresponding manga page. The existence of f implies each frame is mapped, so we remove frames without support, for example, "intro" and "outro" songs. We further enforce that f is monotonic, i.e. $x \leq y$ implies $f(x) \leq f(y)$.

Making a Dataset

In order to construct f we tag each frame of the anime with its corresponding manga page. There are 174,792 total frames for *Girls' Last Tour* and 533 manga pages, which is too much to label by hand. But because f is monotonic, the anime "goes in the same order" as the manga; we only need to tag the *boundaries* when f changes. This is implemented in a graphical user interface (GUI) shown in Figure 3. However, we still need to find which frames mark a transition between manga pages, which can be detected by measuring if the similarity between adjacent frames is less than a certain cutoff.



anime_path: preprocess/anime/glt/0.mp4

manga_path: preprocess/manga/glt/vol1/

33 /133

29193 /34094

1

Figure 3: GUI for efficient manual data tagging.

We generalize the dot product similarity to the element-wise product of two tensors. Our similarity is then $\rho(A, B) = \frac{A \cdot B}{\|A\|\|B\|}$ where $\|X\| = \sqrt{X \cdot X}$, with a cutoff of 0.8.

Methods

We condition an unsupervised model for video generation by *replacing* the latent with our image data. The dimensionality of the latent $\boldsymbol{z}$ is 512 while we need to fit two 256x256 images (initial frame and manga page), so we greyscale, apply a 16x16 average pool, then concatenate to form $\boldsymbol{x}$. Finally, we use $\boldsymbol{z} = 2\boldsymbol{x} - 1$ in order to center at 0. $\boldsymbol{z}$ is passed to the generator G, which generates videos according to the recurrence:

$$F_{n+1} = G(F_n; M(n))$$

where F_n is the nth frame and $M(n)$ is the page function defined above. F_0 is the initial conditioning frame. We then optimize G with mean squared loss and gradient decent.

Latent Space Exploration

We first try to project an image, i.e. given an image X find a latent vector $\boldsymbol{w}$ such that $G(\boldsymbol{w})$ is as close to X as possible. Next, given two latents $\boldsymbol{w}_0$ and $\boldsymbol{w}_1$, we can linearly interpolate between the two with $\boldsymbol{w}(t) = \boldsymbol{w}_0 + t(\boldsymbol{w}_1 - \boldsymbol{w}_0)$, $t \in [0, 1]$.

Training StyleGAN2 with Adaptive Discriminator Augmentation (ADA) [3]

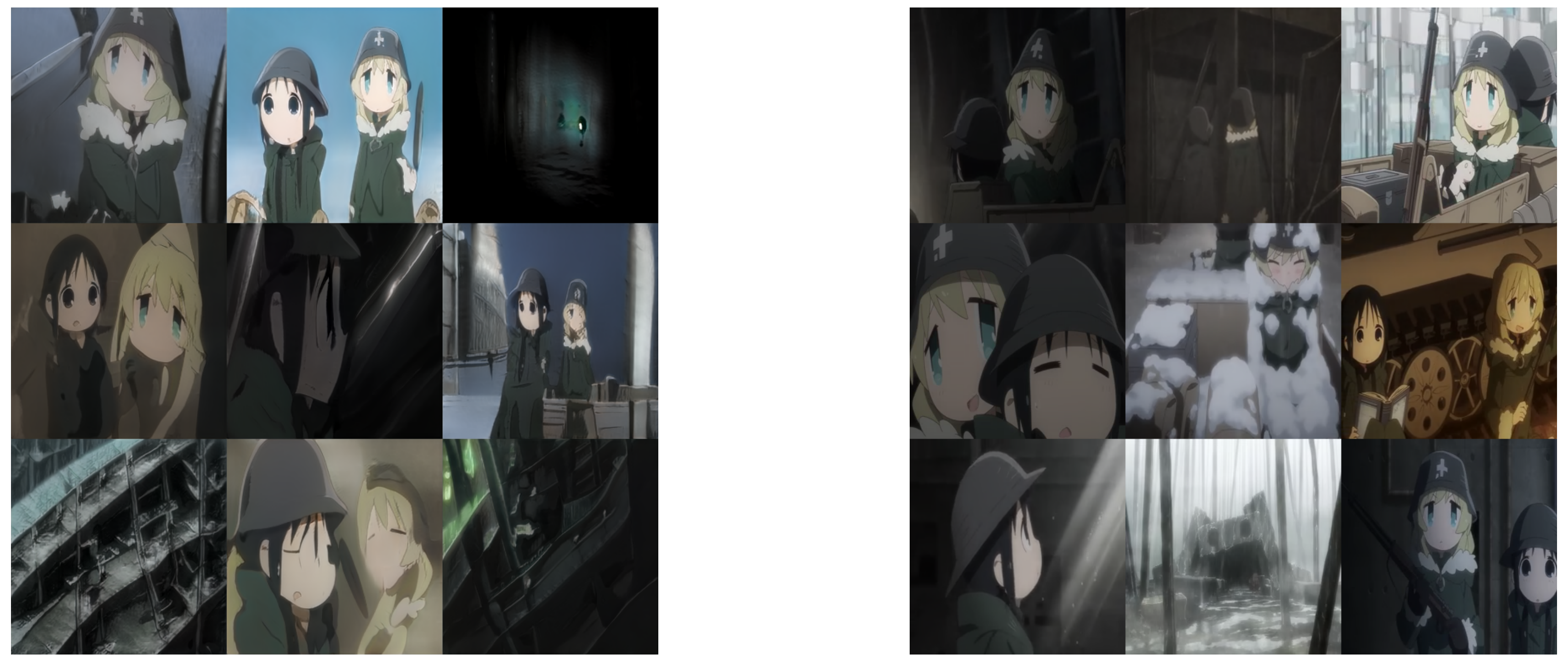


Figure 4: Although certain generated images (left) are strikingly similar to the ground truth (right) e.g. the top left of both, the generated images show clear distortions and are discernible from the real images.

We trained StyleGAN2-ADA [3], achieving a Fréchet inception distance (FID) of 25.722. FID is a common metric of generated image quality; this is relatively high compared to the FID of 3.88 achieved on the FFHQ dataset by [3] (lower is better). We suspect the high correlation between video frames decreases diversity, hindering training.

Conditioning StyleGAN

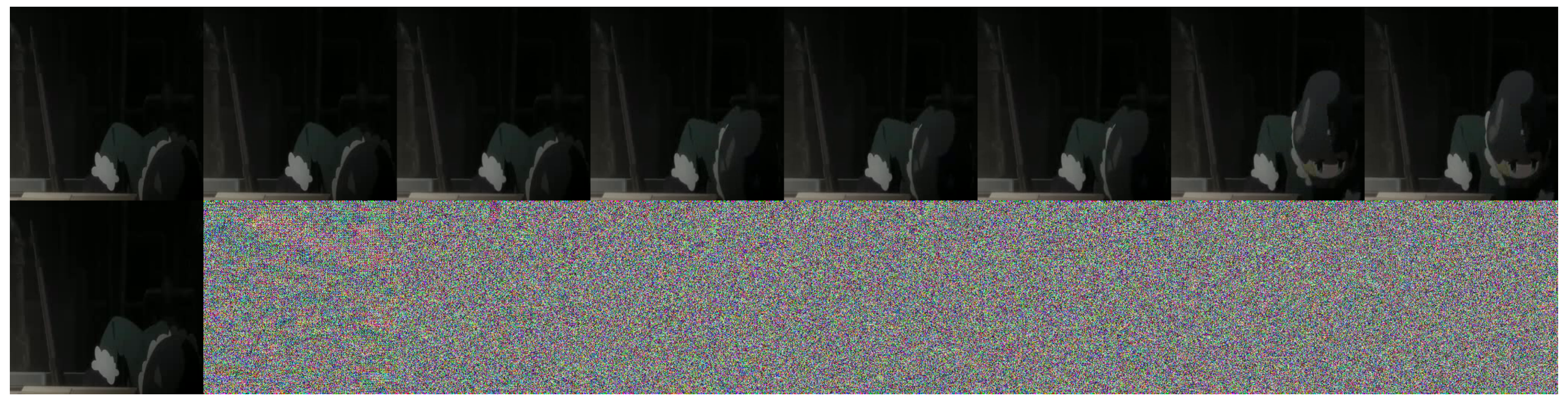


Figure 5: Top row is the ground truth and the bottom row is generated output.

We find that the generator outputs noise. We suspect a capacity issue, as $\boldsymbol{z}$'s dimensionality is too small to effectively encode the initial frame and manga page, so the generator loses the only benefit of conditioning, then is dominated by the discriminator.

Latent Space Exploration

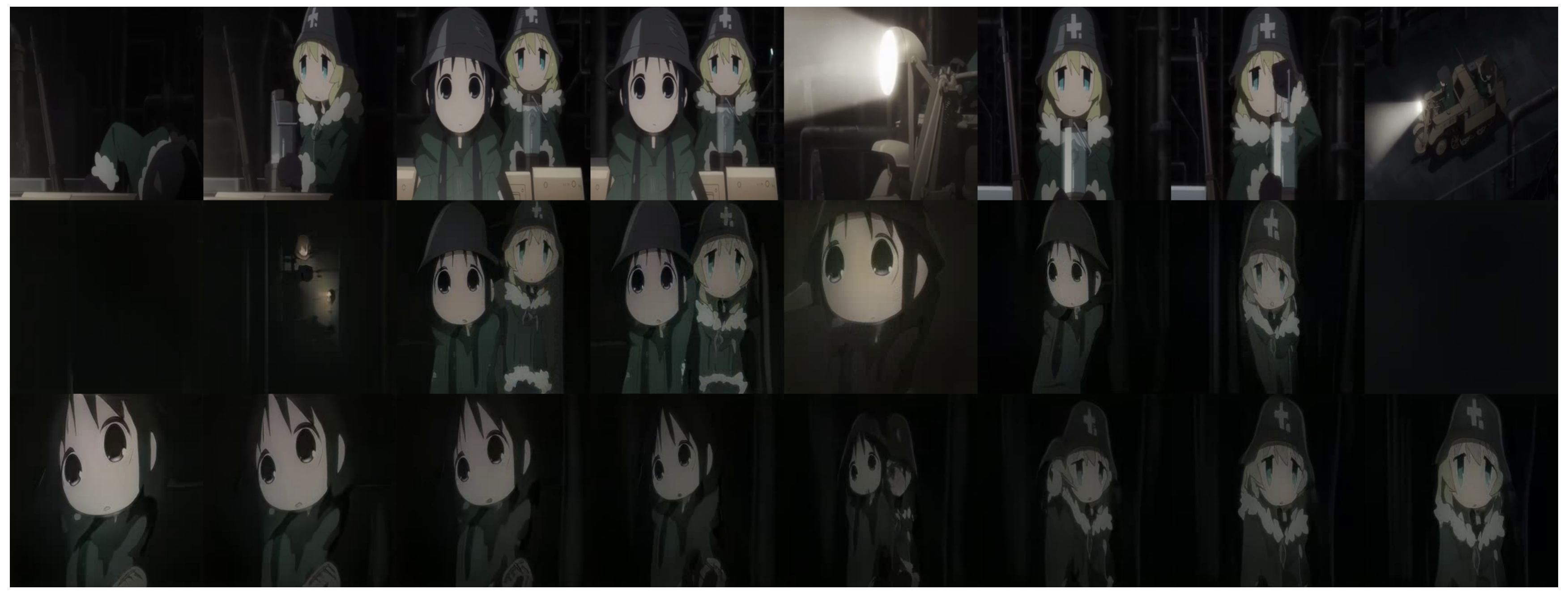


Figure 6: Top row is the ground truth, middle row is a projection of each frame into the latent space, and the bottom row is a linear interpolation. For the videos, see https://youtu.be/4J9BBHX-uNg.

For projection, frames 3, 4, and 7 are reconstructed well but others are not represented at all. We suspect only frequently occurring images can be convincingly represented.

For linear interpolation, we find that the video is "smooth" since the perceptual path length metric and path length regularization discussed in [1], [2] incentives smoothness by minimizing adjacent perceptual differences and by preserving path lengths.

Conclusion

Dismayed at the state of video generation, we propose the simplified domain of anime. We discuss important features of anime/manga for creating a dataset from unlabeled video. We then experiment with adjusting StyleGAN [3] for future video prediction with conditioning and with latent interpolation. Although our results are poor, we believe the avenues of research laid out in this work can be improved in future experiments; one approach may be to add a foreground/background mask [5] to DVD-GAN [4] and then upscale with a deep super-resolution algorithm like waifu2x or dandere2x.

References

[1] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *CoRR*, vol. abs/1812.04948, 2018. arXiv: 1812.04948. [Online]. Available: http://arxiv.org/abs/1812.04948.

[2] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," *CoRR*, vol. abs/1912.04958, 2019. arXiv: 1912.04958. [Online]. Available: http://arxiv.org/abs/1912.04958.

[3] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, *Training generative adversarial networks with limited data*, 2020. arXiv: 2006.06676 [cs.CV].

[4] A. Clark, J. Donahue, and K. Simonyan, "Efficient video generation on complex datasets," *CoRR*, vol. abs/1907.06571, 2019. arXiv: 1907.06571. [Online]. Available: http://arxiv.org/abs/1907.06571.

[5] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," *CoRR*, vol. abs/1609.02612, 2016. arXiv: 1609.02612. [Online]. Available: http://arxiv.org/abs/1609.02612.