Generation of temporally consistent video using synthetic images

Ayushi Arya
Department of Electronics and Communication Engineering
Netaji Subhas University of Technology
ayushi.arya.ec19@nsut.ac.in

Harshit Tangry
Department of Electronics and Communication Engineering
Netaji Subhas University of Technology
harshit.tangy.ec19@nsut.ac.in

Lakshay Ahuja
Department of Electronics and Communication Engineering
Netaji Subhas University of Technology
lakshay.ahuja.ec19@nsut.ac.in

Riya Agarwal
Department of Electronics and Communication Engineering
Netaji Subhas University of Technology
riya.ec19@nsut.ac.in

***Abstract*—Generation of synthetic videos have been a topic of research for a long time. It can form a basis for a broad number of applications, from path prediction to contribution in the entertainment industry. Multiple video generation models have been proposed but all of them had some shortcomings which this paper has aimed to overcome. 3D U-Net architecture has been used with diffusion techniques to generate high quality videos with promising initial results. Diffusion models have always given great results, and this paper shows training of data on both images and videos, to predict a fixed number of future frames. This has shown better performance than previously proposed models and overcomes basic issues like unstable training or narrow outcomes.**

In this paper, we propose to extend the U-Net architecture with gaussian diffusion techniques to produce high quality temporally consistent videos.

I. Introduction

Generating temporally consistent video is an important milestone in generative modeling research. Video Generation has a lot of applications in the modern world.

A lot of generative models have been proposed throughout these years, like GANs [1,2], VAEs [3,4,5,6], a combination of them [7] and flow based models [17]. These have shown great results, but each has some limitations. For realistic video generation, it is essential to learn which objects move, how they move, and how they interact with each other. Diffusion models are famously used for image generation, giving great results. Noising and de-noising of images is done to generate new images. This technique can be used in video generation too. Multiple synthetic frames can be generated to form a complete synthetic video.

We propose a diffusion model in combination of 3D U-Nets for video generation, which gives us great initial results.

3D U-Nets take input volume data, and give segmented 3D images. Temporal attention blocks will be used.

II. Background

Diffusion Models [8,9,10] are generative models that is they are used to generate data similar to the data on which they are trained. They work by destroying training data through the successive addition of Gaussian noise, and then learn to recover the data by reversing this noising process. After training, we just require random sample noise to recover a new image by repeating the denoising process. U-Net [15, 16] is a convolutional neural network that was developed for biomedical image segmentation at the Computer Science Department of the University of Freiburg.As the only requirement for diffusion model is that input and output dimensionality is identical so for the creation of high-resolution images using diffusion modelling we use U-Net model architecture.

III. IMPLEMENTATION

Diffusion models are becoming increasingly popular for generations of images and audio [11, 12, 13, 14]. We aim to generate synthetic videos using this model.

A. U-NET ARCHITECTURE

We use a standard diffusion model with a U-Net.The model is given a fixed number of frames of the video at the time training and using a 3-D Unet [18] over space and time. We modify the convolution layers from 2-D to 3-D.After the Unet structure we use 2 attention blocks. One for the spatial attention which treats the frame axis as the batch axis and the other for the temporal attention which treats the height and width axes as the batch axes.

B. THE DIFFUSION MODEL

The diffusion model [19,20,21] in continuous time [22,23,24,25] is described on the latents $z = \{z_t \mid t \in [0, 1]\}$ following the forward process as described by the markovian structure

$$q(z_t \mid x) \;=\; N(z_t \,;\, \alpha_t x \,,\, \sigma_t^2 I) \tag{1}$$

$$q(z_t \mid z_s) \;=\; N(z_t \mid (\alpha_t/\alpha_s) z_s \,,\, \sigma_{t/s}^2 I) \tag{2}$$

Where the noise schedule is given by the function $\lambda_t = log(\alpha_t^2/\sigma_t^2)$

C. TRAINING

In this section, we go over how to specify the reconstruction loss weight $w(\lambda_t)$ and parameterize the denoising model $\hat{x_t}$ We consider a standard variance-preserving diffusion process for which $\sigma_t^2 = 1 - \alpha_t^2$ The majority of of the work that follows chooses to parameterize the denoising model by directly predicting with a neural network $\hat{\epsilon_\theta}(z_t)$ that implicitly sets $\hat{x_\theta^2}(z_t) = \frac{1}{\alpha_t}(z_t - \sigma_t \hat{\epsilon_\theta}(z_t))$ We optimize the denoising model by training it with a loss function based on weighted mean squared error

$$L_\theta = \|\epsilon - \hat{\epsilon_\theta}(z_t)\|_2^2 = \|\frac{1}{\sigma_t}(z_t - \alpha_t x_\theta(z_t))\| = \frac{\hat{\alpha_t^2}}{\sigma_t^2}\|x - \hat{x_\theta(z_t)}\|_2^2 \quad (3)$$

This can also be interpreted as a reconstruction loss in x-space, but with weights assigned based on a weighting function $w(\lambda_t) = exp(\lambda_t)$ for log signal-to-noise ratio $\lambda_t = log[\alpha_t^2/\sigma_t^2]$ The reduction of generation to denoising can be achieved by optimizing a variational lower bound on the data log likelihood under the diffusion model, which is weighted, or by considering it as a form of denoising score matching[8,11,24,26].In practice, we use the -prediction parameterization, defined as

$$x_\theta(z_t) = (\hat{z_t} - \sigma_t \epsilon_\theta(z_t))/\alpha_t \quad (4)$$

Train the ϵ_θ model using a linear schedule to sample t, and compute the mean squared error in ϵ space.

D. SAMPLER

For this project we have used discrete time ancestral sampler[8].It is a type of MCMC sampler that can be used to simulate a diffusion process backward in time and can be used to generate samples from the posterior distribution of parameters in a Bayesian inference setting, where the parameters are governed by a diffusion process .This sampler is designed based on lower and upper bounds on the reverse process entropy, as described in research papers [27,8,28]. The forward process can be described in reverse by the conditional distribution.

$$\mu_{s|t}^{\sim}(z_t, x) = e^{\lambda_t - \lambda_s} z_t + (1 - e^{\lambda_t - \lambda_s})\alpha_s x \quad (5)$$

and,

$$\sigma_{s|t}^{\sim 2} = (1 - e^{\lambda_t - \lambda_s})\sigma_s^2 \quad (6)$$

It starts with z_1 drawn from a standard normal distribution N (0, 1). The sampler then follows a rule to generate the next sample, zs_s from the previous sample z_t conditioned on the observed data $\hat{x_\theta}(z_t)$.The rule is given by:

$$z_s = \mu_{s|t}^{\sim}(z_t, \hat{x_\theta}(z_t)) + \sqrt{(\sigma_{s|t}^{\sim 2})^{1-\gamma}(\sigma_{t|s}^2)^\gamma}\epsilon \quad (7)$$

where ϵ is standard Gaussian noise,γ is a hyperparameter that controls the stochasticity of the sampler, and s,t are drawn from a uniformly spaced sequence from 1 to 0. This sampler can be computationally intensive as it requires simulating the diffusion process multiple times, but it can provide accurate posterior samples.

IV. RESULTS

We trained our model on the Moving mnist dataset Some of the samples from the dataset are given below

Below table shows the results on various standard video generation metrics: FVD (Frechet Video Distance) [29] is the most popular metric for evaluation of videos which is built upon FID (Frechet Inception Distance)[30] a common metric for images. FID is extended to sequential data such as videos and captures both temporal coherence and the quality of frames.

SSIM (Structured Similarity Index Metric) [31], LPIPS (Learned Perpetual Image Patch Similarity) [32] and PSNR (Peak Signal to Noise Ratio) [33] are the metrics used for the quality of images, here it is applied to each frame of the video and the result is averaged out between all the frames. This gives us information about the quality of each frame.

TABLE I
EVALUATION METRICS FOR GENERATED VIDEOS

Frames	FVD	SSIM_{avg}	PSNR_{avg}	$\text{LPIPS}_{\text{avg}}$
2	-	0.6720	56.1626	0.2013
4	-	0.6717	56.1486	0.2023
5	161.8478	0.6719	56.1569	0.2020

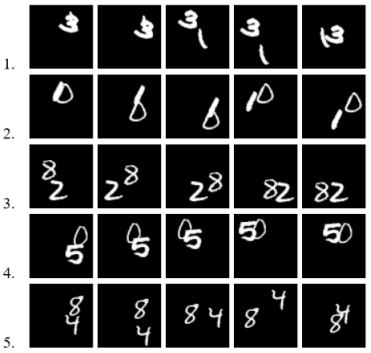

Fig. 1. Training samples

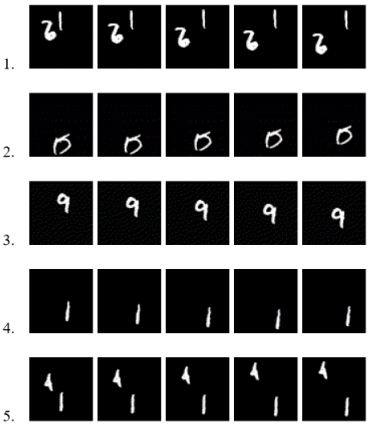

Fig. 2. Results from our model

V. Conclusion

Video Generation can be used to create videos and motions that are completely synthetic. It has a wide range of applications like in VFX, path planning and self driving cars, etc. Other models like GANs and VAEs have shown great results in the past but have their own limitations. Our diffusion model using a 3D Unet architecture attempts to overcome those limitations and provide better and more realistic results. There is a lot of scope for the future, as we can easily add on variational encoders to extend the number of video frames, and even add text embedding for contextual video generation. The current model produces a fixed number of frames. This can be extended to create longer videos with varying lengths autoregressively with guidance method. Another extension can be conditioning the video with text. Text gives context to the video and improves prediction. We can condition the diffusion model in the form of BERT-large embeddings.

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. (2014) 2672–2680

[2] M. Saito, E. Matsumoto, and S. Saito. Temporal generative adversarial nets with singular value clipping. In Proceedings of the IEEE international conference on computer vision, pages 2830–2839, 2017.

[3] Denton, E., Fergus, R.: Stochastic video generation with a learned prior. arXiv preprint arXiv:1802.07687 (2018)

[4] Walker, J., Doersch, C., Gupta, A., Hebert, M.: An uncertain future: Forecasting from static images using variational autoencoders. In: European Conference on Computer Vision (ECCV). (2016)

[5] Xue, T., Wu, J., Bouman, K., Freeman, B.: Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In: Neural Information Processing Systems (NIPS). (2016)

[6] Babaeizadeh, M., Finn, C., Erhan, D., Campbell, R.H., Levine, S.: Stochastic variational video prediction. In: International Conference on Learning Representations (ICLR). (2018)

[7] J.-Y. Zhu et al., "Toward multimodal image-to-image translation," in Advances in neural information processing systems, 2017, pp. 465- 476.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems, pages 6840–6851, 2020.

[9] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In International Conference on Machine Learning, pages 2256–2265, 2015.

[10] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In Advances in Neural Information Processing Systems, pages 11895–11907, 2019.

[11] Diederik P Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. arXiv preprint arXiv:2107.00630, 2021.

[12] Chitwan Saharia, William Chan, Huiwen Chang, Chris A Lee, Jonathan Ho, Tim Salimans, David J Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. arXiv preprint arXiv:2111.05826, 2021.

[13] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. arXiv preprint arXiv:2104.07636, 2021.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 234–241. Springer, 2015.

[15] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications. In International Conference on Learning Representations, 2017.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 234–241. Springer, 2015.

[17] Manoj Kumar, Mohammad Babaeizadeh, Dumitru Erhan, Chelsea Finn, Sergey Levine, Laurent Dinh, and Durk Kingma. VideoFlow: A flow-based generative model for video. arXiv preprint arXiv:1903.01434, 2019.

[18] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In International conference on medical image computing and computer-assisted intervention, pages 424–432. Springer, 2016.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems, pages 6840–6851, 2020.

[20] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In International Conference on Machine Learning, pages 2256–2265, 2015.

[21] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In Advances in Neural Information Processing Systems, pages 11895–11907, 2019.

[22] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. WaveGrad: Estimating gradients for waveform generation. International Conference on Learning Representations, 2021.

[23] Diederik P Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. arXiv preprint arXiv:2107.00630, 2021.

[24] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. International Conference on Learning Representations, 2021.

[25] Belinda Tzen and Maxim Raginsky. Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit. arXiv preprint arXiv:1905.09883, 2019.

[26] Belinda Tzen and Maxim Raginsky. Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit. arXiv preprint arXiv:1905.09883, 2019.

[27]] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, ICML, 2021.

[28] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In International Conference on Machine Learning, pages 2256–2265, 2015.

[29] T. Unterthiner et al., "FVD: A NEW METRIC FOR VIDEO GENERATION." Accessed: Apr. 25, 2023. [Online]. Available: https://openreview.net/pdf?id=rylgEULtdN

[30] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: https://doi.org/10.1109/tip.2003.819861.

[31] M. F. Naeem, S. J. Oh, Y. Uh, Y. Choi, and J. Yoo, "Reliable Fidelity and Diversity Metrics for Generative Models," arXiv:2002.09797 [cs, stat], Jun. 2020, Available: https://arxiv.org/abs/2002.09797

[32] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," arXiv:1801.03924 [cs], Apr. 2018, Available: https://arxiv.org/abs/1801.03924

[33] A. Horé and D. Ziou, "Image Quality Metrics: PSNR vs. SSIM," IEEE Xplore, Aug. 01, 2010. https://ieeexplore.ieee.org/document/5596999