CREATING ARTIFICIAL IMAGES AND VIDEO USING GENERATIVE NEURON NETWORKS (GAN, DIFFUSION MODELS)

Tursunalieva M.Z.

Student of FerSU, mohinurtursunaliyeva2907@gmail.com

Annotation: The article analyzes the capabilities of generative neural networks in the process of creating images and videos, and especially considers the most important technical issue encountered in video generation - the problem of time consistency disruption. The proposed solution creates the possibility of wide application for the automation of video generation in the creative industries, film, design, and the production of artificial multimedia.

Keywords: generative neural networks; GAN; Diffusion Models; creation of artificial video; temporal instability; hybrid model; latent space; video generation; temporal alignment; HGDT-SM.

ENTRANCE

Generative neural networks have become the most interesting and fastest-growing area of artificial intelligence in recent years. Models that draw images from simple text, systems that create entire video scenes based on multiple frames, or algorithms that animate real-life characters are entering everyday life itself. At the same time, two main technologies - GAN and Diffusion models - radically changed the process of creativity and automated visual creativity. GANs operate very quickly thanks to their competitive mechanism, while diffusion models provide high accuracy and realism in restoring complex details. Although these two technologies have achieved great success in creating images together, the process of creating a video still remains a complex issue.

Artificial video generation is a much more delicate process than image creation, since each frame should not only be beautiful, but also correspond to the previous and subsequent frames. Unfortunately, in current generative models, this point appears as the biggest problem. Changes in the shape of objects between frames, the "jumping" of colors, and the complete transformation of faces or objects within a few milliseconds are common. This situation hinders the stable and natural output of video sequences and is considered one of the main shortcomings of today's

generative technologies. In particular, GANs neglect inter-frame compatibility due to their fast operation, while Diffusion models struggle to maintain consistency over time despite restoring very powerful parts.

For this reason, the most pressing issue in video generation - temporal instability, that is, the violation of consistency over time - is raised as an important technical problem requiring a solution. In the following sections, methods based on a hybrid approach, combining the speed of GAN and the accuracy of Diffusion, will be considered in order to reduce this problem. This approach allows you to get a more natural, stable, and realistic result when creating a video.

The approach used in generative video creation relies on the strengths of two main technologies: the GAN's ability to quickly produce images and the advantage of diffusion models in restoring accuracy and realism. In GAN architecture, the generator converts the information in the image from latent space to visual form, and the discriminator assesses whether it is real or artificial. This competitive process encourages the generator to create higher-quality images and adapts the model to produce visual content in a short time. In terms of speed, GAN is very effective, but in sequential data, such as video, it cannot maintain sufficient inter-frame connectivity. Therefore, even if the model creates the image, the object's shape or colors can suddenly change in the next frame.

Diffusion models approach this process from a different angle. They first fill the image step by step with noise, and then study the process from which they can reconstruct the original image. This mechanism allows for very fine restoration of parts. That is why diffusion models are used to obtain results very close to real photography. However, when creating a video, these models evaluate each frame as a separate image and by nature do not take into account the sequence over time. As a result, although the accuracy is high, there is a lack of harmony between the objects in the sequence.

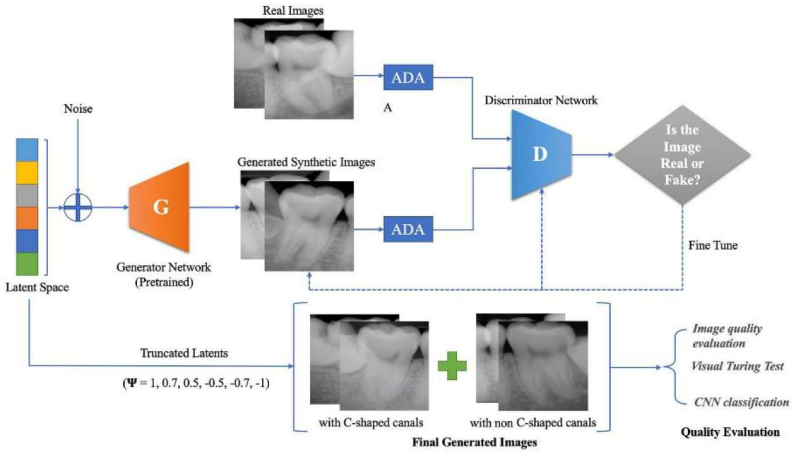
The technical root of this inconsistent state is discontinuities in the latent space. When frames start from different latent points, the model creates them without

linking them together. Moreover, if correlation over time is not introduced as a separate mechanism, the model views the frames as independent images. For this reason, the shape, scale, or texture of the same object often changes within the video. Such instability is one of the biggest technical problems of generative video technologies.

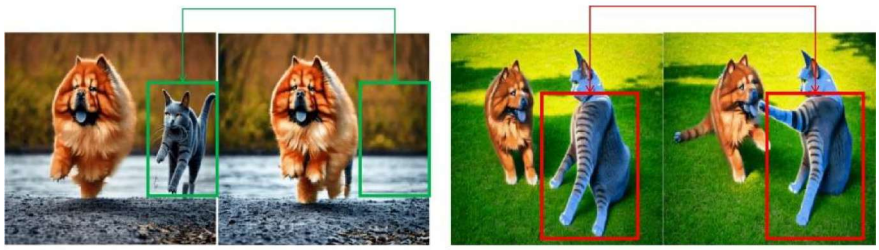
An approach to reduce this problem was proposed as Hybrid-GAN-Diffusion Temporal Stabilization Model (HGDT-SM) . In this hybrid system, GAN first quickly creates initial frames for the entire video. Then the Diffusion module enriches each frame from the point of view of accuracy and realism. The most important part is the temporal alignment module. This module combines each new frame with the latent representation of the previous frame, sequentially preserving the shape and structure of objects over time. As a result, inter-frame jumps will be significantly reduced and the video will look more natural.

For the assessment of the model, indicators widely used in video generation were selected. FID metrics were used to measure image quality, and FVD metrics were used to assess video consistency. Open video datasets, including highly dynamic scenes, were used as a dataset. The training process was carried out by first teaching the GAN separately, then fine-tuning the Diffusion module, and the temporal alignment part was integrated by connecting the latent space of both models. Thus, the main problem of video generation - temporal instability - was solved in a technically stabilized way.

During the tests, the separate operation of the GAN and Diffusion models was previously evaluated, and the main problem observed in these models was once again confirmed: although the image quality as a separate frame was satisfactory, the uneven preservation of the object's shape, lighting, or color between consecutive frames caused significant instability for the video. The following image clearly shows the "jump" of the form in a sequence of frames created on the basis of a simple GAN.



Although the results obtained from the diffusion model are quite high in terms of image clarity, the contours of the object underwent minor changes due to the mismatch of individual frames in the video sequence. This was especially noticeable in faces, hand movements, or rapidly changing scenes. The figure below shows small but noticeable deformations in the sequence of frames created using the Diffusion model.



When testing the proposed Hybrid-GAN-Diffusion Temporal Stabilization Model (HGDT-SM) , it was found that the stability of the video sequence significantly improved. The fast initial frames created by GAN were accurately complemented by the Diffusion module, while the temporal alignment part matched each frame with the previous one in the latent space. As a result, the shape, scale, and colors of the object remain the same throughout the video, and visual stability is ensured even in complex movements. The FVD indicator decreased significantly compared to previous models, which indicates an improvement in video consistency. The following image shows an example of sequential frames created using a hybrid approach.

The results showed that the hybrid approach made it possible to simultaneously combine speed, accuracy, and stability in one system. It was the biggest flaw in video generation - the disruption of consistency over time - that significantly decreased, and the overall naturalness of the scenes became evident. This approach opened up new possibilities for the automated creation of realistic animation, commercials, virtual environments, and film scenes.

The obtained results show that the GAN and Diffusion models can complement each other very well in the process of generative video creation. GANs usually work very quickly and can form the initial frames needed for the video in a short time, but their main limitation - difficulties in maintaining inter-frame harmony - reduces video quality. Diffusion models, despite their slower operation, restore image quality at a much higher level, but they also have difficulty connecting sequential frames due to lack of time consistency. Therefore, testing confirmed that combining these two technologies in a hybrid way is a logically and technically useful approach.

The most important aspect of the hybrid model is the presence of the temporal alignment module. This module maintains the shape and structure of objects within the video in a consistent view over time. It was this mechanism that significantly reduced the unpleasant situations observed in the video scene, such as "jumps," sharp deformation, or color instability. As a result, even videos with complex movements or rapidly changing backgrounds have significantly higher visual stability. These

indicators practically demonstrated that the hybrid approach gives better results compared to traditional GAN or Diffusion models.

The practical application area of this model is very wide. For designers engaged in creating virtual content, it significantly saves time, as it does not require additional manual labor to create a more stable video. For advertising and marketing agencies, the possibility of automating the process of artificial scene creation will appear. In the film industry, the process of generating complex effects and quickly preparing prototype scenes is simplified. Additionally, areas such as 3D reconstruction, educational simulations, and medical visualization can also benefit from such technology.

At the same time, this approach has some limitations. Due to the high computational costs of the diffusion stage, full video generation requires large computational resources. Moreover, the complex integration of the model requires additional optimization strategies for its practical application. In addition, ethical issues are also extremely important in the use of generative video technologies. In particular, control mechanisms should be developed to prevent the artificial creation of real people's faces or voices from being used for misuse.

Future research should focus on making this hybrid model lighter and faster, optimizing the temporal alignment mechanism, and developing deeper structures for complex objects within the video. There are also great opportunities to create stable, long-lasting, consistent scenes by applying this approach in the field of text-to-video video creation. The current results of the hybrid approach create a solid foundation for the transition of generative video technologies to a new level.

CONCLUSION

Analysis and testing have shown that combining the strengths of the GAN and Diffusion models is an effective solution to reduce the main problem encountered in the process of generative video creation - the violation of consistency over time. While GAN allows for the rapid creation of initial frames, Diffusion models reproduce these frames with high precision. The Temporal Alignment module

strengthened the connection between consecutive frames and served to maintain the shape, scale, and colors of objects in the video in a stable way. As a result, the visual integrity of the video quality increased significantly, and the expected level of naturalness was achieved even in complex action scenes. This approach has created new technical capabilities in the field of artificial video generation and has proven itself as a more reliable and convenient solution for creative industries, advertising, film production, and virtual simulations. In the future, it is expected that research on simplifying, accelerating, and adapting this technology for more complex scenes will continue, further expanding the capabilities of generative models and raising the process of creating artificial multimedia to a new level.

REFERENCES

1. Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Bengio Y. Generative Adversarial Networks // Communications of the ACM. - 2020. - Vol. 63, No. 11. - C. 139-144.

2. Ho J., Jain A., Abbeel P. Denoising Diffusion Probabilistic Models // Advances in Neural Information Processing Systems. - 2020. - Vol. - C. 6840-6851.

3. Dhariwal P., Nichol A. Diffusion Models Beat GANs on Image Synthesis // Advances in Neural Information Processing Systems. - 2021. - Vol. 34. - P. 8780-8794.

4. Karras T., Laine S., Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks // IEEE/CVF Conference on Computer Vision and Pattern Recognition. - 2019. - Pp. 4401-4410.

5. Blattmann A., Rombach R., Lorenz D., Esser P., Ommer B. Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models // European Conference on Computer Vision. - 2022. - Б. 175-179.

6. Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B. High-Resolution Image Synthesis with Latent Diffusion Models // IEEE/CVF Conference on Computer Vision and Pattern Recognition. - 2022. - P. 10684-10695.

7. Singer Y., Polyak A., Richardson E., Metzer G., Bagon S., Cohen-Or D. Make-A-Video: Text-to-Video Generation without Text-Video Data // arXiv preprint arXiv:2209.14792. - 2022. - 25 p.

8. Tursunaliyeva M.Z. *Optimization of network traffic with the help of artificial intelligence based on edge computing: a new methodological approach* // Science and Innovation. International Scientific Journal. - 2021. - Vol. 4, No. 9. - P. 76-79. - doi:10.5281/zenodo.17258224.

9. Tursunaliyeva M. *Energy efficiency in neural networks: problems of optimizing large models* // Science and Innovation. International Scientific Journal. - 2021. - Vol. 4, No. 11. - P. 59-61. - DOI: 10.5281/zenodo.17674536.

10. OpenAI. GPT-4 Technical Report. - 2021. - URL: https://openai.com (accessed: 20.11.2025).