



Advancements In GAN-Based Video Generation: A Comprehensive Survey

Mariya, Adiba Muskaan, Anandam Nagalaxmi, Ms. Nadia Anjum

¹⁻³ BE, AI&DS, VIII Sem, SCETW, Hyderabad, INDIA, ⁴Associate Professor, ADCE dept, SCETW, OU,
Hyderabad, INDIA

Stanley College of Engineering and Technology for Women

Abstract: In recent times, Generative Adversarial Networks (GANs) have become a potent deep-learning framework. This framework is used for synthetic data generation and has wide-ranging applications, including image synthesis, text-to-image translation, and video generation. GANs were first introduced by Goodfellow et al.

They have been adapted to handle intricate video generation challenges. The challenges are related to the synthesizing of temporally consistent realistic video frames. Video generation comes with its own set of hurdles. One of these is maintaining temporal coherence. Another is modeling complex motion patterns. Video generation is a highly active research area. It is also a demanding field.

The goal of this survey is to give an overview. The overview will detail the state-of-the-art GAN-based techniques. These techniques are used in video generation. The survey includes different elements. These elements are architecture training methodologies, evaluation metrics and recent advancements. The advancements are in text-to-video models. These models can generate video from textual descriptions.

Keywords: Generative Adversarial Networks (GANs), Video Generation, Deep Learning, Temporal Coherence, Transformer-based GANs, Text-to-Video Synthesis, Video Prediction

1. Introduction

Generative Adversarial Networks (GANs) have transformed the field of artificial intelligence (AI) by enabling the synthesis of high-quality images, text, and video. Originally introduced by Goodfellow et al. [1], GANs operate through a competitive framework between a generator (which creates synthetic samples) and a discriminator (which evaluates their realism). Over the past decade, GANs have evolved significantly, influencing applications in computer vision, deep learning, and multimedia generation.

Video generation remains a challenging task due to the high dimensionality and temporal dependencies involved in video sequences. Unlike static image synthesis, video generation requires models to ensure temporal coherence, motion consistency, and scene realism over multiple frames. Traditional methods relied on autoencoders and recurrent neural networks (RNNs), but these approaches often produced blurry and unrealistic videos. The emergence of GAN-based video synthesis has led to significant breakthroughs in generating high-fidelity, temporally consistent videos.

Motivation and Importance of GANs in Video Synthesis

GANs are highly suitable for video synthesis. Very much the reason behind this turns out to be their adversarial training framework. This structure encourages realistic data creation with continuous refinement. These players engender the refinement between a discriminator and a generator. There are many varieties of GAN models. There is MoCoGAN, StyleGAN-V, and even VideoGPT. Each has adapted the key ideas of GAN to video synthesis. In doing so, some unique challenges that videos pose have been relatively effectively met. Such challenges may include temporal coherence and motion-content disentanglement. MoCoGAN provides a unique decomposition method that helps in separating the motion from the content. This increases its effectiveness in modeling static and dynamic aspects of video sequences. The conditioning of StyleGAN on video is StyleGAN-V. It allows for high-quality, coherent video generation since it has control over the features both in space and time. Other active domains of GAN research in the realm of video generation include the application of GANs to super-resolution and video prediction. This paper presents an example designed for super-resolution: the Temporally Coherent GAN, which tries to enhance video quality without necessarily affecting the temporal coherence among frames. The other model presents the Stochastic Adversarial Video Prediction. This model adds some form of stochasticity to handle the intrinsic indeterminism of future frames. This can make it more adaptable to different conditions for motion.

Contributions of This Survey

The main contribution of this survey is to collect and assess the GAN-based video generation elements from technical and practical perspectives. We classify and compare various GAN structures, training strategies, and evaluation standards. A panoramic view of the current GAN-based video generation landscape is provided. Promising research directions such as hybrid model integration and improving computational efficiency are also identified.

Continuing, we take a look at GAN structures that apply to video generation. We analyze methodologies that improve them. We study improvements in text-to-video techniques. Finally, we take a look at the most recent advancements in evaluation and dataset standards.

This structured review hopes to help researchers and professionals know more about this developing domain. The domain of which is GANs for video generation.

2. Background and Key Concepts

Key concepts are described as below:

1. Generative Adversarial Networks (GANs) have emerged as a powerful framework for image and video synthesis, demonstrating remarkable success in text-to-image, image-to-image translation, and video generation. Introduced by Goodfellow et al. (2014) [1], GANs consist of two competing neural networks:

- A generator (G) that synthesizes fake samples.
- A discriminator (D) that distinguishes between real and generated samples.

2. Problems Specific to Video The Temporal Coherence Issue and Spatiotemporal Consistency- Various problems GANs encounter.

They should maintain temporal coherence. This is about continuity in time. It must present smooth transitions between frames. They should also maintain spatiotemporal consistency. This means realism both in space and in time. An example exists in Temporally Coherent GANs [2]. This addresses continuity in time.

Another one is MoCoGAN [11]. It is a motion-content splitting model. The splitting of these allows one to have more control over them. An additional technique is Stochastic Adversarial Video Prediction [3]. It has introduced randomness. Such an innovation aims at capturing more than one future frame. That is why it suits more real applications.

3. Video Generation Evaluation Metrics:

The evaluation of generated videos measures metrics like:

- Fréchet Video Distance (FVD) This measures the similarity of distributions between produced and real videos. It captures spatial and temporal aspects.
- Inception Score (IS) IS assesses frame quality. It is based on the distinctiveness of objects.
- Peak Signal-to-Noise Ratio (PSNR) PSNR is usually found in super-resolution. It assesses frame similarity. Unfortunately, it does not provide insight into temporal aspects.

These metrics are highly informative. They give a strong basis for evaluating the quality of frames and the flow of frames across video sequences.

4. Common Metrics for Video GAN Training and Testing:

Video GANs need vast datasets. Some popular ones are:

- UCF101: Suggests 101 classes of actions. Useful in motion-focused applications .
- Kinetics-600: Has 600 classes of action. Useful for scene generalization .
- YouTube-8M: A video content mix. Useful for complex motion and scene changes.

These datasets are basic. They support model training for varied realistic video generation. The options are tailored based on the needs of the task. Needs such as action or scene diversity.

3. Techniques and Architectures in GAN-based Video Generation

3.1 Motion and Content Decomposition GANs (MoCoGAN)

MoCoGAN [1] decouples motion and content encoding in videos.

The content module encodes the spatial information of the static frames.

The motion module acquires temporal variations and motion patterns.

This strategy empowers the controlled video synthesis via the motion vectors modification and maintenance of the content.

3.2 VideoGPT (Autoregressive GANs for Video)

VideoGPT-2 is a transformer-based GAN architecture proposed to capture temporal dependencies in videos. It deploys Vector Quantization (VQ) for the purpose of encoding video frames into discrete tokens. Transformer model creates consistent sequences by learning spatiotemporal correlations. Generates video samples that are of higher quality than GANs.

3.3 StyleGAN-V (Style-Based Video Generation)

The StyleGAN-V algorithm extends the capabilities of GANs from synthesizing still images to video synthesis, by incorporating spatio-temporal coherence. Utilizes adaptive instance normalization for texture variations control. Enhanced temporal coherence is achieved through training with the addition of motion constraints. Crafting diverse, highly-realistic video samples with fine control is within the program's capabilities.

4. Methodology for GAN

A discriminator (D) assesses the realism of the synthetic video frames produced by a generator (G) in a two-network adversarial learning approach used in GAN-based video generation. By minimizing a loss function that is derived from adversarial feedback, the generator learns to create temporally coherent video sequences.

GAN-Based Video Generation Pipeline

The following steps make up the standard GAN-based video generation methodology:

The first step is data preprocessing, which involves extracting frames from actual videos and normalizing them for training.

2. Feature Extraction: Convolutional and recurrent layers are used to learn temporal and spatial representations.

3. Adversarial Training: The discriminator determines whether video frames are real or fake, while the generator creates them.

4. Loss Optimization: Adversarial, perceptual, and temporal loss functions are used to train the model iteratively.

5. Post-processing: The finished video clip is polished.

Video generation counts heavily on GANs. Diverse approaches have been fashioned. These approaches tackle specific problems tied to data with temporal attributes. An approach that stands out is the Two-Time-Scale Update Rule. It goes by the name TTUR.

TTUR technique stabilizes the training process of GANs. This is due to its unique characteristics. TTUR updates the discriminator and generator at a fluctuating rate. This fluctuation leads to higher-quality outputs. Additionally, Temporal GANs achieve significance. Their goal is to uphold the coherence of time. They guarantee the synchrony of created frames in time.

Randomness is present in stochastic video prediction models. The generation process carries some kind of randomness. This model produces different potential future frames. Uncertainties like these are part of a dynamic scene.

Methods like self-supervised learning get used regularly. They improve temporal coherence. The fascinating aspect is that large labeled datasets are not needed. Rely on implicit supervision instead.

Cross-domain adaptation is now discussed. It is related to GANs creating diverse content. The content comes in a myriad of styles and domains. This action aids in enhancing the model's versatility. It also boosts generalization capabilities.

Datasets Video GAN Training and Evaluation.

GAN-based video generation model performance is robustly influenced by datasets used for training and evaluation. Dataset choice can directly influence the model's learning ability. It can impact the diversity of motion patterns and object interactions. It can also alter temporal coherence. Below are some of the datasets commonly used in the domain of video generation. We demonstrate their characteristics. We can also find relevance to various applications.

1. UCF101: UCF101 contains 13,320 videos. These videos are spread over 101 action categories. It is widely applied to the benchmark of tasks on action recognition and video generation. The actions are diversified, and the context also varies. This can be helpful for training the model. Models can include varied motion dynamics and different compositions of scenes.

2. Kinetics-600: Kinetics-600 is a very large dataset. Kinetics-600 includes almost 600 action categories. There are more than 500,000 video clips derived from YouTube. This dataset proves beneficial. It can train models on a wide range of human activities and interactions. A rich source of varied motion sequences is provided.

3.YouTube-8M: This dataset features millions of labeled video segments. Topics and content are diverse. Its annotations are extensive. They facilitate the training of models. The models are for tasks like video classification and generation. It helps to improve understanding of the contextual relationships in video data.

4. LSUN (Large-scale Scene Understanding): Focusing mainly on images LSUN offers labeled scene categories. These can be expanded to video generation. They do this by constructing synthetic video clips from static scenes. This dataset is key for tasks that involve scene consistency over time.

5.Hollywood2: It includes 1,707 video clips from movies of various genres. This dataset specifically caters to action recognition. It can be useful too for tasks involving video generation. These tasks require understanding of intricate narrative structures and character interactions.

5.Results

Evaluation Metrics and Performance Comparison.

For comparing the efficiency of GAN video generation models, robust metrics are a must to measure. Such metrics should ideally assess the quality of the video produced. These metrics are supposed to handle realism, diversity, and coherence aspects. Let's now highlight the prominent evaluation metrics that are most frequently used under video generation tasks. So, descriptions for each metric are presented in the coming lines.

1. Fréchet Video Distance (FVD): This has been derived as an extension of Fréchet Inception Distance (FID). This measure applies to image creation. The measure stretches across from the image to the video domain. FVD calculates the distance between the distribution of generated videos and those real videos. It operates the calculation in the space of features. It accounts for temporal coherence and is also a feature of visual quality. These are vital to this case of VFD. The low FVD case signifies that generated videos do belong closer to real videos, hence closer by the point of distribution.

2. IS: Inception Score .It was designed to score images. Today, it can score the quality of any video created. It depends upon any pre-trained Inception model's output. This score measures both the quality of content as well as the diversity across different generated samples. That is why it is significant. Sadly, IS does not entertain any explicit consideration of temporal coherence either.

3. Peak Signal-to-Noise Ratio (PSNR): PSNR is used vastly in image and video compression. It measures the ratio between the maximum power of the signal and the power of corrupting noise. High values of PSNR indicate quality. They imply the produced video is similar to the original. However, PSNR may not always correspond to perceived visual quality.

4. Structural Similarity Index (SSIM): SSIM measures the similarity between two images or videos. It is determined through luminance contrast and structure. SSIM is supposed to be much more sensitive to visual quality evaluation as compared to PSNR. It computes its results from perceived differences. The range of its value ranges between -1 and 1. In case there is complete similarity, its value should be 1.

5. Temporal Coherence Metrics: These measure smoothness. Also, they measure the consistency of motion between consecutive frames. In an AI-generated video. One of the metrics is optical flow consistency. The other is motion smoothness. These metrics prove useful for evaluation.

6.Conclusion

This review paper offers a thorough summary of advances in Generative Adversarial Networks known as GANs for video creation. It mirrors the swift evolution of techniques in this dynamic field. This has been shown by the study of 15 leading research papers. GANs have shown significant progress in overcoming the unique challenges of video synthesis. This applies especially to maintaining temporal coherence and achieving consistency in space and time. The examination of chief techniques such as MoCoGAN and VideoGPT stands out. As does the study of other methods for text-to-video generation. These show the innovative ways researchers are using to increase quality and realism in video generation. MoCoGAN for example divides motion and content. This allows for more organized video creation. VideoGPT also uses vector quantization and autoregressive models. It generates coherent video sequences from nothing. These techniques show how structural improvements can directly tackle video data complexities. The result is a

more precise output. This output captures both dynamic and static elements effectively. It is hard to overstate the importance of datasets. Examples are UCF101 Kinetics-600 and YouTube-8M. They serve as crucial resources for GAN model training and evaluation. These datasets cover a broad range of actions and scenarios. This allows for solid model training and generalization over various video contexts. Discussion of evaluation metrics is also important. Fréchet Video Distance (FVD) and Peak Signal-to-Noise Ratio (PSNR) are such metrics. They play a critical role in assessing video performance. Through these metrics, researchers can measure the quality of video synthesis. Despite advancements, challenges persist in GAN-based video generation. Achieving high-resolution outputs is a challenge. This is especially true if we want to keep temporal coherence. Many models have difficulties. They struggle with managing complex motion patterns. They also have issues with diverse content types. These issues can result in artifacts or inconsistencies in generated videos. Another critical barrier is the computational demands of training these models. This barrier often necessitates extensive resources and time to overcome. We look ahead to a few promising yet challenging paths for forthcoming research. Hybrid methodologies that blend the strengths of varied architectures may amplify model efficiency and firmness. Attempts to refine generalization across different datasets could lead to more adaptable models. These models can adjust to various applications. Moreover, there is a growing demand for research. The focus is on the ethical implications of GAN-originated content. There is a special emphasis on areas such as deep fakes, misinformation, and privacy issues. GAN technology is advancing rapidly. Its possible uses are vast and diverse. GANs are set to overhaul how we generate and interact with video content. They are applicable in entertainment gaming education and healthcare. Their capacity to spawn lifelike animations and simulate real-world scenarios is unparalleled. GANs can assist in content creation processes. This makes them invaluable in the digital sphere. Moreover, the integration of GANs with AI technologies is noteworthy. Technologies such as natural language processing and reinforcement learning are included. This could boost their prowess. This would lead to more sophisticated and clever systems. Lastly, the continually progressing research on GANs in video creation emphasizes impressive strides in the area. It also underscores the significance of sustained research and innovation. We venture deeper into the intricacies of video synthesis. The knowledge gained from this field will not only push forward our understanding of artificial intelligence. It will also pave the way for revolutionary applications. These applications can reform our interaction with visual media.

7. References

- [1] I. Goodfellow, et al., "Generative Adversarial Nets," Conference on Neural Information Processing Systems (NeurIPS), 2014.
- [2] Y. Kwon and J. Lee, "Temporally Coherent GANs for Video Super-Resolution," IEEE International Conference on Computer Vision (ICCV), 2019.
- [3] C. Vondrick, et al., "Stochastic Adversarial Video Prediction," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
- [4] Y. Chen, et al., "Temporal Generative Adversarial Networks for Video Coherence," AAAI Conference on Artificial Intelligence, 2020.
- [5] S. Zhou, et al., "Learning Temporal Coherence via Self-Supervision," European Conference on Computer Vision (ECCV), 2018.
- [6] Z. Gong, et al., "TiVGAN: Text to Image to Video Generation," IEEE Transactions on Multimedia, 2020.
- [7] N. Aafaq, et al., "Video Generation from Text," International Journal of Computer Vision (IJCV), 2017.
- [8] L. Huang, et al., "Variable Length Video Generation from Text," ACM Multimedia Conference (ACM MM), 2022.
- [9] C. Guan, et al., "CogVideo: Large-Scale Pre Training for Text-to-Video Generation via Transformers," Conference on Neural Information Processing Systems (NeurIPS), 2022.

- [10] M. Babaeizadeh, et al., "Stochastic Latent Residual Video Prediction," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2019.
- [11] S. Tulyakov, et al., "MoCoGAN: Decomposing Motion and Content for Video Generation," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
- [12] Y. Chen, et al., "VideoGPT: Video Generation with VQ-VAE," IEEE Transactions on Image Processing (TIP), 2021.
- [13] T. Karras, et al., "StyleGAN-V: A Style-Based Generator Architecture for Video," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.
- [14] J. Ho, et al., "Video Diffusion Models," Advances in Neural Information Processing Systems (NeurIPS), 2021.
- [15] C. Zhang, et al., "Efficient Video GANs," IEEE Transactions on Multimedia, 2020.

