CHAPTER 14

Generative adversarial network for video analytics

A. Sasithradevi[a], S. Mohamed Mansoor Roomi[b], and R. Sivaranjani[c]
[a]School of Electronics Engineering, VIT University, Chennai, India
[b]Department of Electronics and Communication Engineering, Thiagarajar College of Engineering, Madurai, India
[c]Department of Electronics and Communication Engineering, Sethu Institute of Technology, Madurai, India

14.1 Introduction

The objective of video analytics is to recognize the events in videos automatically. Video analytics can detect events such as a sudden burst of flames, suspicious movement of vehicles and pedestrians, abnormal movement of a vehicle not obeying traffic signs. A commonly known application in the research field of video analytics is video surveillance which has started evolving 50 years ago. The principle behind the video surveillance is to involve human operators to monitor the events occurring in a public area, room, or desired space. In general, an operator is given full responsibility for several cameras and studies have shown that increasing the number of cameras to be monitored per operator degrades the performance of the operator. Hence, video analysis software aims to provide a better trade-off between accurate event detection and huge video information [1–3]. Machine learning, in particular, its descendant namely deep learning has prompted the research in the video analytics domain. The fundamental purpose of deep learning is to identify the sophisticated model that signifies the probability distributions over the different samples of videos which need analytics.

Generative adversarial network (GAN) provides an efficient way to learn deep representations with minimal training data. GAN is an evolving technique for generating and representing the samples using both unsupervised and semisupervised learning methods. It is accomplished through the implicit modeling of high-dimensional data distribution. The underlying working principle of GAN is to train the pair of networks in competition with each other. Among these networks, one acts like an imitator and the other as a skillful specialist. From the formal description of GAN, the generator creates fake data mimicking the realistic one and the discriminator is an expert trained to distinguish the real samples from the forger ones. Both the networks are trained simultaneously in competition with each other. This generic framework for GAN is shown in Fig. 14.1.

Both the generator and the discriminator are the neural networks where the former generates new instances and the latter assesses whether the instances belong to the dataset.

Generative Adversarial Networks for Image-to-Image Translation
https://doi.org/10.1016/B978-0-12-823519-5.00008-7

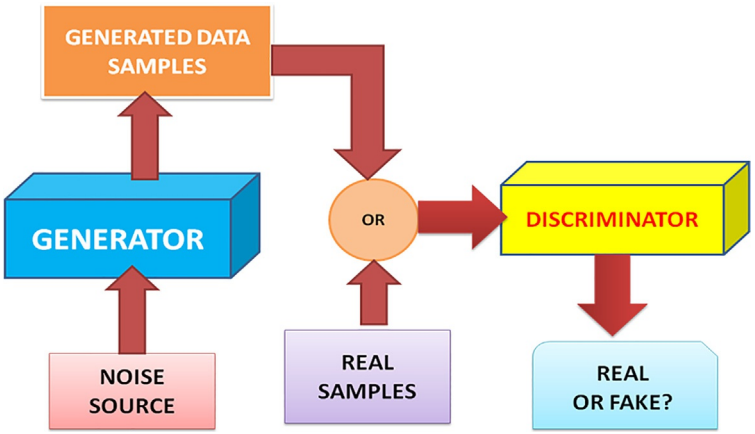

Fig. 14.1 Generic framework for generative adversarial networks.

For the purpose of classification, the discriminator plays the role of a classifier to distinguish the real from the fake. To build a GAN, one needs to have a training dataset and a clear idea about the desired output. Initially, GAN learns from simple distribution of 2D data, later GAN could be able to mimic high-dimensional data distribution along with eventual training. During the training phase, both the competing networks get the attributes regarding the distribution of data. The data samples generated by the generator along with the real data samples are used to train the discriminator. After sufficient training, the generator is trained against the discriminator. Thus the generator learns to map any random data samples. Consider the scenario as Fig. 14.1, where a D-dimensional noise vector obtained from the latent space is fed into the generator which converts them into new data samples. The discriminator then processes both the real and fake samples for classifying it. The main advantage of GAN relies on its randomness which aids it to create new data samples rather than the exact replica of the real data. Another crucial advantage of GAN over Autoencoders [4] and Boltzmann machine [5] is that GAN does not rely on Markov chain for the purpose of generating training models. GANs were designed to eliminate the high complexity associated with Markov chains. Also, the generator function undergoes a minimum restriction compared to Boltzman machines. Owing to these advantages, GANs have been attracted toward a variety of applications and the craving to utilize it in numerous areas is increasing. They have been effectively used in a wide variety of tasks like image to image translation, obtaining high-resolution images from low-resolution images, deciding the drugs for treating desired diseases, retrieving images, object recognition, text-image translation, intelligent video analysis [6], and so on. In this article, we present an overview of the working principle of GANs and its variants available for video analytics. We also emphasize the pros, cons, and the challenges for the fruitful implementation of GANs in different video analytic problems.

The remainder of this chapter is organized as follows: Section 14.2 provides the building blocks of GANs, its driving factor called objective functions and the challenging issues of GANs. Section 14.3 highlights the variants of GANs emerged for the problem of video analytics in past years. Section 14.4 discusses the possible future works in the area of video analytics based on GAN. Section 14.5 concludes this chapter.

14.2 Building blocks of GAN

This section describes the basic building blocks of GAN and the different objective functions used for training the GAN architectures.

14.2.1 Training process

The training process involving the objective or cost function is the basic building block for GANs. Training of GAN is a dual process which includes choosing the parameters for a generator that confuses the discriminator with fake data and discriminator that maximize the accuracy for any given application. The algorithm involved in the training process is described as follows:

Algorithm 14.1
Step 1: Update parameters of discriminator "θ_D"**:**
Input: "m" samples from real frames and "m" samples from noise data.
Do: Compute the expected Gradient $\nabla_{\theta_D} = f\{\mathrm{J}_{\theta_D}(\theta_D; \theta_G)\}$
Update: $\boldsymbol{\theta_D} \leftarrow (\boldsymbol{\theta_D}, \boldsymbol{\nabla_{\theta_D}})$
Step 2: Update parameters of Generator "θ_G"**:**
Input: "m" samples from noise data and $\boldsymbol{\theta_D}$.
Do: Compute the expected Gradient $\nabla_{\theta_G} = f\{\mathrm{J}_{\theta_G}(\theta_G; \theta_D)\}$
Update: $\boldsymbol{\theta_G} \leftarrow (\theta_G, \nabla_{\theta_G})$.

The objective or the cost function $\boldsymbol{V(G, D)}$ for the training depends on the two competing networks. The training process includes both maximization and minimization as

$$\underset{\boldsymbol{D}}{\boldsymbol{max}} \; \underset{\boldsymbol{G}}{\boldsymbol{min}} \; \boldsymbol{V(G, D)} \tag{14.1}$$

where $V(G, D) = f_{pdata(x)} \log D(x) + f_{pg(x)} \log(1 - D(x))$.

As illustrated in the Algorithm 14.1, one of the model parameters are updated, while the other is fixed. An exclusive discriminator $D'(x) = \frac{p_{data(x)}}{p_{data(x)} + p_{g(x)}}$ is available for any fixed generator G [7]. The generator is also optimal when $P_g(x) = P_{data}(x)$ and it shows that the generator reaches an optimal point only when the discriminator is totally confused in

discriminating the real data from fakes. The discriminator is not trained completely until the generator reaches the optimum value. But the generator is updated simultaneously with the discriminator. An alternate cost function typically used for updating the generator is $\boldsymbol{max_G \log D(G(Z))}$ instead of $\boldsymbol{min_G \log (1 - D(G(Z)))}$.

14.2.2 Objective functions

The main objective of generative models is to make $P_g(x)$ equivalent to the real data distribution $P_{data}(x)$. Hence, the underlying fact for training the generator is to reduce the dissimilarity between the two distributions [8]. In recent years, researchers have attempted to utilize various dissimilarity measures to upgrade the performance of GAN. This section describes the difference in computation using various measures and objective functions.

f-Divergence: It is a dissimilarity measure between two distribution functions that are convex in nature. The f-Divergence between the two convex functions [8] namely $P_g(x)$ and $P_{data}(x)$ is written as

$$D_f\left(P_{data}||P_g\right) = \int_{x1}^{x2} P_g(x) f\left(\frac{P_{data}(x)}{P_g(x)}\right) dx \tag{14.2}$$

Integral probability metric: It provides the maximal dissimilarity measure between two arbitrary functions [8]. Consider the data space $X \in R$ with probability distribution function defined as $P(X)$. The IPM distance metric between the distributions P_{data}, $P_g \in P(X)$ is defined as

$$d_F\left(P_{data}, P_g\right) = Sup_{f \in F} E_{x \sim P_{data}}\left[f(x) - E_{x \sim P_g}(f(x))\right] \tag{14.3}$$

Auxiliary object functions: The auxiliary functions that are related to the adversarial objective functions are reconstruction and classification objective function.

- *Reconstruction objective function*: The goal of reconstruction objective function is to minimize the difference between the output image of the neural network and the real image provided as input to the neural network [9, 10]. This type of objective function aids the generator to preserve the content of the real image data and use the autoencoder architecture for GAN's discriminator [11, 12]. The discrepancy value evaluated using reconstruction objective function mostly involves L_1 norm measure.
- *Classification objective function*: Discriminator network can also be used as classifier [13, 14] when cross entropy loss is employed as an objective function in the discriminator. Cross entropy loss is widely used in many GAN applications for semisupervised learning and domain adaptation. This objective function can also be used to train the generator and discriminator jointly for the classification purpose.

14.3 GAN variations for video analytics

In recent years, intelligent video analytics has become an emerging technology and research field in academics and industry. The scenes in videos are recorded by cameras that aids for invigilation of happenings that occur in the area where human ability fails. Recently, a huge number of cameras are utilized for useful purposes [6] like fire detection, person detection and tracking, vehicle detection, smoke detection, unknown object and crime detection in country borders, shopping malls, airports, sports stadiums, underground stations, residential areas, and academic campuses and so on. The manually monitoring the videos is really cumbersome due to the obstacles like drowsiness of the operator, diversion due to increased responsibilities, etc. This prompts the need for semisupervised approaches for analyzing the events in videos [15, 16]. Hence, intelligent video analytics is one challenging problem in the field of computer vision where deep networks have not succeeded classical handcrafted attributes. To date, video analytics has traveled a long journey from holistic features such as motion history image [17] (MHI), motion energy image [18] (MEI), action banks [19] up to local feature-based approaches like HOG3D [20], spatiotemporal histogram of radon projection (STHRP) [21], histogram of optical flow [22], and tracking approaches. One efficient approach is to employ deep networks for learning and analyzing the videos without the knowledge of class labels but with the sequential organization of frames termed as "weak supervision." This technique also requires a little supervision in strategies for providing input to deep neural networks such as sampling, encoding, and organizing methods. Unlike deep networks, generative models called GANs [23, 24] have been successfully implemented in the field of video analytics without human intervention in labeling the videos for applications such as future video frame prediction [25]. Over a period of time, the architectures of GAN is modified for various applications like video generation, video prediction, action recognition, video summarization, video understanding, and so on as listed in Table 14.1.

14.3.1 GAN variations for video generation and prediction

Recent progress in generative models [26] has attracted the researchers to examine image synthesis. In particular, GANs have been employed to synthesize images from random data distribution, through nonlinear transformation from prime image to synthesized one or generate the synthesized images from the source domain. This enhanced advances in image synthesis have gained the confidence to utilize GANs for generating video sequence. One of the challenging issues in using GANs for generating and predicting videos is that the output of the GAN architectures must provide meaningful video responses. This challenge has added huge responsibility to GAN which includes understanding both the spatial and the temporal content of the video. One such extension of GAN is MoCoGAN [27], which is used for generating videos with no prior knowledge

Table 14.1 GAN variations.

S. no	GAN variations	Application
1	MoCoGAN	Video generation
2	VGGAN	
3	LGGAN	
4	TGANs	
5	Dynamic transfer GAN	
6	FTGAN	
7	DMGAN	Video prediction
8	AMCGAN	
9	Discrimnet	Action recognition
10	HiGAN	Video recognition
11	DCycle GAN	Face translation between images and videos
12	PoseGAN	Human pose estimation
13	Recycle GAN	Video retargeting
14	DTRGAN	Video Summarization

about priming image. This variant of GAN architecture partitions the input data distribution into two subspaces namely content and motion subspace. The content subspace sampling follows Gaussian distribution sampling whereas motion subspace sampling was accomplished by RNN. These two subspaces form the two discriminators called content and motion discriminators. Even though MoCoGAN could generate videos of variable lengths, the motion discriminator was designed only to handle the frames in limited number. As shown in Fig. 14.2, the spatial content generation was performed for different instances of appearance but the motion was fixed at the same expression.

Another useful variant of GAN is dynamic transfer GAN [28]. It attempts to generate the video sequence by transferring the dynamics of temporal motion available in the

Fig. 14.2 Example frames generated by MoCoGAN [27].

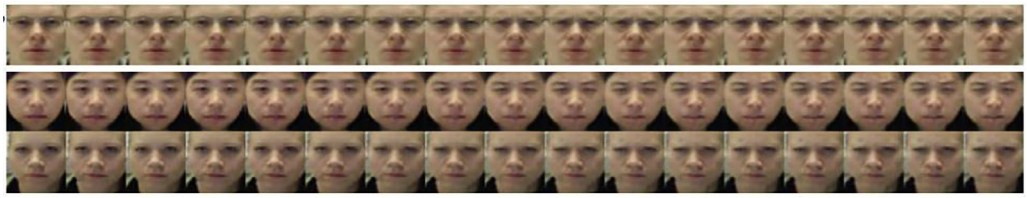

Fig. 14.3 Frames generated by dynamic GAN [28] for anger expression.

source video sequence onto a prime target image. This target image contains the spatial content of the video data and the dynamic information is obtained from the arbitrary motion.

RNN is used for spatiotemporal encoding. This dynamic GAN can generate video sequences of variable length using the competition between a generator and two discriminator networks. Among the two discriminator networks, one act as spatial discriminator to monitor the fidelity of the generated video sequence and other acts as a dynamic discriminator to maintain the integrity of the entire video sequence. They have provided visualization to demonstrate the ability of the dynamic GAN in encoding the enriched dynamics from source videos by suppressing the appearance features. Fig. 14.3 shows an example of frames generated using dynamic GAN for anger expression.

Ohnishi et al. developed flow and texture generative adversarial network (FTGAN) model [29] used to generate hierarchical video from orthogonal information. FTGAN comprises two networks namely FlowGAN and TextureGAN. This variation in the GAN architecture is proposed to explore the representation and generate videos without enormous annotation cost. Flow GAN is used to generate optical flow which provides the edge and motion information for the video to be generated. The RGB videos are generated from optical flow using Texture GAN. The generic framework for FTGAN is shown in Fig. 14.4.

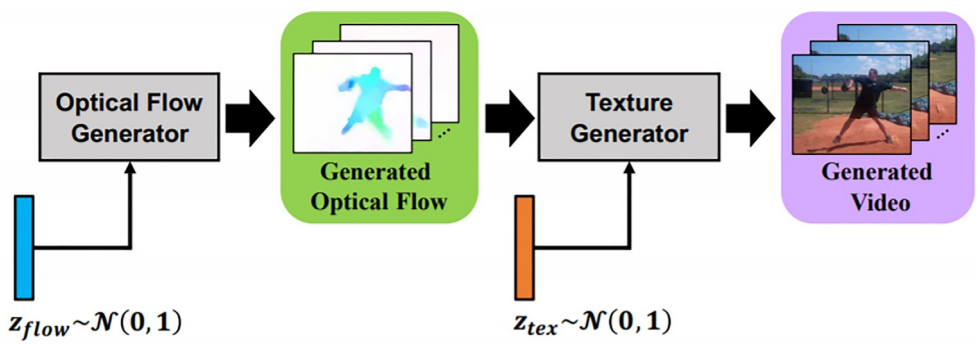

Fig. 14.4 Generic framework for FTGAN [29].

TextureGAN preserves the consistency in the foreground and the scenes while accumulating texture information with the generated optical flow. This model provides a progressive advance in generating more realistic videos without label data. The prime advantage of FTGAN is that both GANs share complementary information about the video content. The authors used both real and computer graphics (CG) videos for training the texture and FlowGANs. The real-world dataset namely Penn Action contains 2326 videos of 15 different classes whereas the CG human video dataset namely SURREAL consists of 67,582 videos. The TextureGAN and FlowGAN are trained on these dataset for 60 k iterations. The accuracy obtained on SURREAL dataset is 44% and 54% while using textureGAN and flowGAN respectively. On Penn Action dataset, the accuracy obtained through textureGAN is 72% and flowGAN is 58%.

A multistage dynamic generative adversarial network (MSDGAN) was proposed for generating time-lapse videos of high resolution. The process involved in MSDGAN [30] is twofolds: at the initial stage, realistic information is generated for each frame in the video. The next stage prunes the videos generated by the first stage through the use of motion dynamics which could make the videos closer to the real one. The authors had used a large-scale time-lapse dataset to test the videos. This model generates realistic videos of up to 128×128 resolution for 32 frames. They had collected over 5000 time-lapse videos from YouTube and short clips are created manually from it. After that, the short video clips are partitioned into frames and MSDGAN is used to generate clips. A short video clip can be generated from continuous 32 frames. Fig. 14.5 shows the frames generated by the MSDGAN and the red circle indicates motion between adjacent frames.

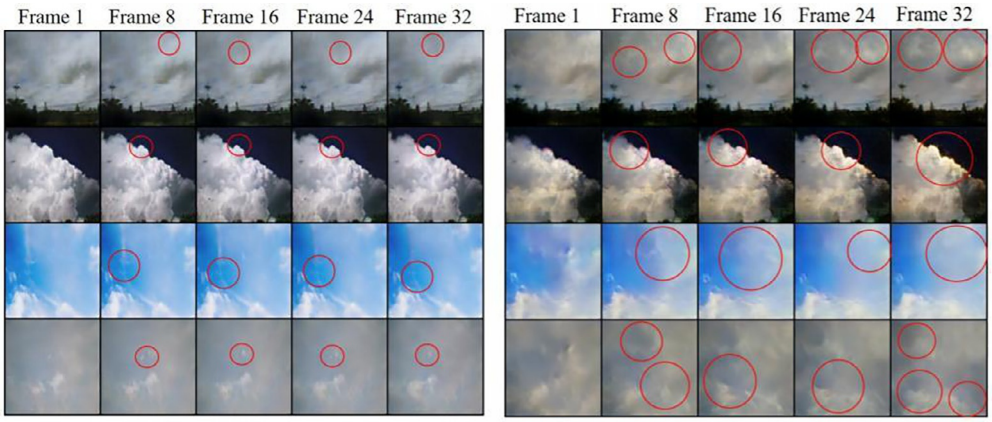

Fig. 14.5 Frames generated by MSDGAN [30], given the start frame 1.

A robust one-stream video generation architecture which is an extension of Wasserstein GAN architecture known as improved video generative adversarial network (iVGAN) is another variation of GAN. This model generates the whole video clip without separating foreground from background. Similar to classical GAN, iVGAN [31] model has two networks called a generator and a critic/discriminator network. The aim of using a generator network is to create videos from a low-dimensional latent code. Critic network discriminates the real and fake data and updates in competence with the generator. This iVGAN architecture tackles the challenging issues in video analytics such as future frame prediction, video colonization, and in painting. The authors used different dataset such as stabilized videos collected from YouTube and Airplanes dataset. This model works by constantly filling the damaged holes to reconstruct the spatial and temporal information of videos. Fig. 14.6 depicts the example video frames generated by iVGAN.

One of the useful efforts for generating the videos for the given description/caption has been taken by Pan et al. [32]. These kinds of video generation from the text description are attracted toward real-time applications. It is attained through the efficient extension of GAN architecture termed as temporal GAN (TGAN). TGAN consists of a generator and three discriminator networks. The input to the generator network is the combination of noise vector and the encoded sentences derived from LSTM network. The generator produces the frames of video sequences using 3D convolution operator. Three discriminators are utilized in TGANs for purposes such as video, frames, and motion discrimination. Among the discriminators, the function of two networks is to distinguish the real and fake videos or frames formed by the generator. In addition to this, these discriminator networks discriminate the semantically matched and mismatched video or frame text description pairs. The need for last discriminator network is to improve the temporal coherence between the real and generated frames. The whole TGAN architecture has undergone end to end learning. This GAN variant is evaluated over datasets like SMBG, TBMG, and MSVD to generate videos from captions. The coherence metric implies readability and temporal coherence of videos. The coherence metric of 1.86 is reported for TGANs. Table 14.2 enumerates the collection of dataset used for evaluating the GAN variants proposed for video generation.

14.3.2 GAN variations for video recognition

Video recognition is usually done via large number of labeled videos during the training session. For a new test task, many videos are unlabeled and annotation is needed there. It also requires human help to annotate every video. It's a tedious process to annotate a large set of data. In order to overcome this Yu et al. proposed a novel approach called hierarchical generative adversarial networks (HiGAN) [3] where the fully labeled images are utilized to recognize those unlabeled videos. The idea behind HiGAN model is

Fig. 14.6 Video frames generated using iVGAN [31].

Table 14.2 List of dataset available to validate video generation techniques.

S. no	Model	Dataset
1	MoCoGAN	MUG facial expression dataset, YouTube videos, Weizmann action dataset and UCF101
2	TGANs	SMBG,TBMG, MSVD
3	Dynamic transfer GAN	CASIA
4	FTGAN	Penn Action, SURREAL
5	iVGAN	Tiny videos, Airplane dataset
6	MSDGAN	YouTube videos, Beach dataset, Golf dataset

combining low-level conditional GAN and high-level conditional GAN and utilizing the adversarial learning from them. Also, this method provides domain invariant feature representation between labeled images and unlabeled video. The performance is evaluated by conducting experiments on two complex video datasets UCF101 [33] and HMDB51 [34]. In this work, each target video is split into 16-frame clips without any overlap and it constructs a video clip domain by combining all the target video frame. In each video clip, the deep feature that is 512D feature vector from pool 5 layer of 3D ConvNets is extracted and are used to train large-scale video dataset. The HiGAN comparatively outperforms in terms of recognition rate in both datasets compared to the approach C3D [35]. HiGAN recognition rate is observed as 4% improvement in UCF101 and 10% improvement in HMDB51 dataset compared to C3D technique.

Human behavior understanding in video is still a challenging task. It requires an accurate model to handle both the pixel-wise and global level prediction. Spampinato et al. [36] demonstrated an adversarial GAN-based framework to learn video representation through unsupervised learning to perform both local and global prediction of human behavior in videos. In this approach, first the video is synthesized by factorizing the process in to the generation of static visual content and motion and secondly enforcing spatiotemporal coherency of object trajectories and finally incorporates motion estimation and pixel-wise dense prediction. So, the self-supervised way of learning provides an effective feature set which further used for video object segmentation and video action tasks [37]. Also, the new segmentation network proposed is able to integrate into any another segmentation model for supervision. This provides a strong model for object motion. The wide range of experimental evaluation showed that VOS-GAN performance on modeling object motion better than the existing video generation methods such as VGAN, TGAN, and MoCoGAN.

In the previous researches the following approaches were implemented for video retargeting [38]: The first one is specifically performed domain wise which is not applicable for other domains. And the second one is implemented across the domain which needs manual supervision for labeling and alignment of information and the last approach is unsupervised and unpaired image translation where learning is mutually done in different domains which is also shown insufficient information for processing. Bansal et al. [38] propose a new unsupervised data-driven approach for effective video retargeting which incorporates spatiotemporal information with conditional generative adversarial networks (GANs). It combines both spatial and temporal information along with adversarial loses for translating content and also preserving style. The publicly available Viper dataset is used for experimentation for image-to-labels and labels-to-image to evaluate the results of video retargeting. The performance measures such as mean pixel accuracy (M), average class accuracy (AC), and intersection over union (IoU) provides comparatively better results for the combination of cycle GAN and recycle-GAN

Jang and Kim [39] developed appearance and motion conditions generative adversarial network (AMC-GAN) which consists of a generator, two discriminators, and perceptual ranking module. The two discriminators monitor the appearance and motion features. They used a new conditioning scheme that helps the training by varying appearance and motion conditions. The perceptual ranking module enables AMCGAN for understanding the events in the video. AMCGAN model is evaluated on MUG facial expressions and NATOPS human action dataset. The MUG dataset consists of 931 video clips which contain six basic emotions like anger, disgust, fear, happy, sad, and surprise. It is preprocessed to get 32 frames of resolution 64×64 pixels. The NATOPS human action dataset has 9600 videos containing 24 different actions.

In unsupervised video representation future frame prediction is a challenging task. Existing methods operate directly on pixels which result blurry prediction of the future frame. Liang et al. [26] proposed a dual motion generative adversarial net (GAN) architecture to predict future frame in video sequence through dual learning mechanism. The future frame prediction and dual future flow prediction form a close loop. It achieves better video prediction by generating informative feedback signals to each other. This dual motion GAN has fully differentiable network architecture for video prediction. Extensive experiments on video frame prediction, flow prediction, and unsupervised video representation learning demonstrate the contributions of Dual Motion GAN to motion encoding and predictive learning. Caltech and YouTube Clips are taken for future frame analysis to show the performance of video recognition using dual motion GAN compared to other existing approaches in the KITTI dataset. The performance evaluation metrics such as mean square error (MSE), peak signal-to-noise ratio (PSNR), and structural similarity index metrics (SSIM) are used to evaluate the image quality of future frame prediction. Higher PSNR and SSIM are achieved via dual motion GAN. The implementations are based on the public Torch7 platform on a single NVIDIA GeForce GTX 1080. Dual motion GAN takes around 300 ms to predict one future frame.

14.3.3 GAN variations for video summarization

Due to the availability of the huge amount of multimedia data produced by the progressive growth of video capturing devices, video summarization [12,40–42] plays a crucial role in video analytics problem. Video summarization [43] extracts the representative and useful content from the video for data analysis and it is highly useful in large scale video analysis. One of the efficient approaches in video summarization is that deriving the suitable key frames from the entire video and those set of key frames are enough to portray the story of the video. To enhance the quality of summarization, there exists some challenges need to be tackled by the summarization techniques. The first challenge is to choose a fine key frame selection strategy which takes into account the temporal relation

of the frames within the video and the importance of the key frames. The next challenge is to devise a mechanism to assess the preciseness and completeness of the selected key frames. To address these issues, several models have been introduced so far like feature-based approaches [12], Long-short-term memory (LSTM)-based models [1,44] and determinantal point process (DPP)-based techniques [45]. Owing to the memory problems that arise in LSTM as well as DPP and redundant key frames issue in feature-based approaches, GAN has attracted the researchers in this community because of its regularization ability. One of the GAN variants proposed for video summarization namely dilated temporal relational generative adversarial network (DTRGAN) [46] is shown in Fig. 14.7.

The generator contains two units namely dilated temporal relational (DTR) and bidirectional LSTM (Bi-LSTM). The generator gets the video and the real summary of the respective video as input. The DTR unit aims to tackle the first challenge. The inputs to the discriminator are real, generated, and random summary pairs and the purpose of the discriminator is to optimize the player losses at the time of training. A supervised generator loss term is introduced to attain the completeness and preciseness nature of the key frames.

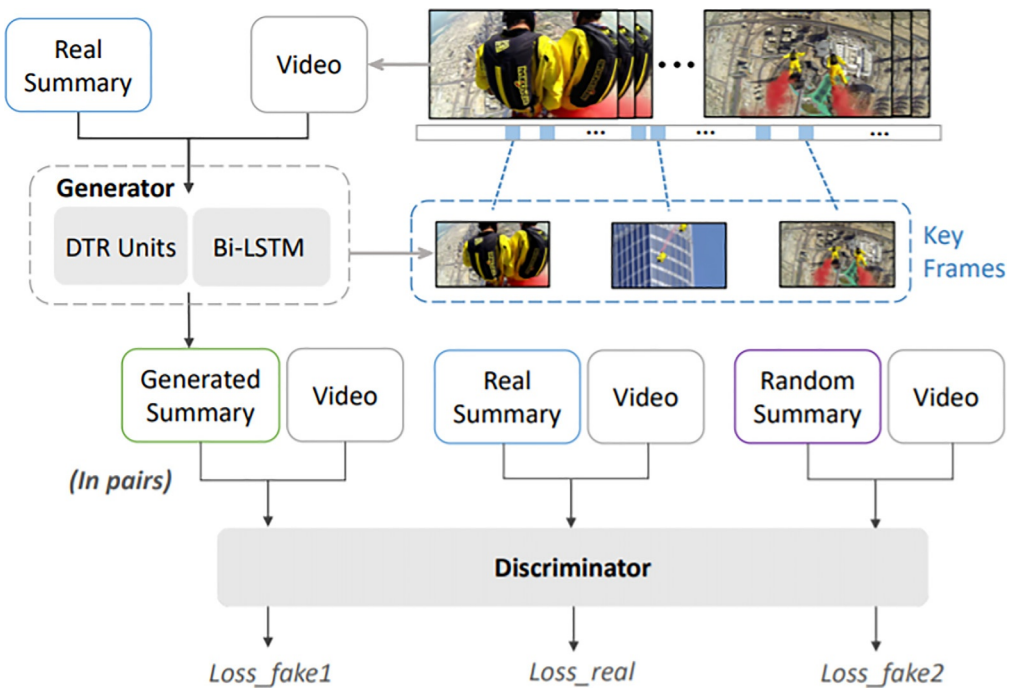

Fig. 14.7 Architecture of DTRGAN [46].

14.3.4 PoseGAN

Walker et al. [25] have developed a video forecasting technique by generating future pose using generative adversarial network (GAN) and variational autoencoders (VAEs). In this approach, video forecasting is attained by generating videos directly in pixel space. This approach models the whole structure of the videos including the scene dynamics conjointly under unconstrained environment. The authors divided the video forecasting problem into two stages: The first stage handles the high-level features of video like human scenes and uses VAE to predict the future actions of a human. The authors used UCF101 dataset for evaluating the poseGAN architecture in predicting the future poses of human.

14.4 Discussion

14.4.1 Advantages of GAN

One of the major advantages of GAN is that it does not require knowledge about the shape of the generator's probability distribution model. Hence, GANs avoids the need for the determined density shapes for representing high complex and high-density data distribution.

Reduced time complexity: The sampling of generated data can be parallelized in GANs and it makes them pretty faster than PixelRNN [47], Wavenet [48], and PixelCNN [49]. In the future frame prediction problem [39], the autoregressive models rely on the value of the previous frame's pixel value for the prediction of the probability distribution of the future frame's pixel. Hence, the generation of the future frame is too slow and the time consumption is even worse for high-dimensional data. But GANs uses a simple feed-forward neural network strategy for mapping in the generator. The generator creates all the future frame pixels at the same time itself rather than pixel by pixel approach followed by autoregressive models. This pace of GAN processing attracted many researchers in various fields.

Accurate results: Form the study of different GAN variants it is evident that GAN can produce astounding results for video analytics problems. Also, the performance is far better than variational autoencoder (VAE), one of the generator models which assume the probability distribution of pixels as a normal distribution. As GAN can master in capturing the high-frequency parts of the data, the generator develops to guide the high-frequency parts to betray the discriminator.

Lack of assumptions: Even though VAE attempts to maximize likelihood through variational lower bound, it needs assumptions on the prior and posterior probability distributions of data. On the other hand, GANs do not need any strong assumptions about the probability distribution.

14.4.2 Disadvantages of GAN

Trade-off between discriminator and generator: The imbalance occurs between generator and discriminator because of nonconvergence and mode collapse. Mode collapse is a commonly occurring and difficult issue in GAN models. It happens in the case when the generator is offered with images that look similar. Also, when the generator is trained extensively without updating any information to the discriminator, the mode collapses. Owing to this mode collapse, the generator will converge to an optimal data which fools discriminator the most and it is the best realistic image from the perspective of the discriminator. A partial mode collapse occurs in GANs frequently than a complete mode collapse. Thus, the training process involved in GAN is heuristic in nature.

Hyperparameters and training: The need for suitable hyperparameters to attain the cost function is a major concern in GANs. The tuning of these parameters is also a time-consuming process.

14.5 Conclusion

GAN is growing as an efficient generative model through the generation of real-like data using random latent spaces. The underlying fact in the GAN process is that it does not need the understanding of real data samples and high-level mathematical foundations. This merit allowed the GANs to be extensively used in various academic and engineering fields. In this chapter, we introduced the basics and working principle of GAN, several variations of GAN available for various applications like video generation, video prediction, action recognition, and video summarization in the area of video analytics. The enormous growth of GAN in the video analytics domain is not only due to its ability to learn the deep representation and nonlinear mapping but due to its potential to use the enormous amount of unlabeled video data. There are huge openings in the development of algorithms and architectures of GAN for using it in different application domains apart from video analytics, such as prediction, superresolution, generating new human poses, and face frontal view generation. The future scope in video recognition includes exploiting large-scale web images for video recognition which will further improve the recognition accuracy. Video retargeting can be accomplished more precisely using spatiotemporal generative models and further, it can be extended to multiple source domain adaptation. Also, the spatiotemporal neural network architecture can be applied for video retargeting in future. The real-world videos with complex motion interactions can be attempted for video recognition through the modeling of multiagent dependencies. Also, the alternative can be made for loss function, evaluation metrics, RNN, and synthetically generated videos to improve the performance of video recognition system. Generative adversarial neural networks can be the next step in deep learning evolution and while they provide better results across several application domains.