Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 8358314, 8 pages
https://doi.org/10.1155/2021/8358314

Research Article

Conditional Deep 3D-Convolutional Generative Adversarial Nets for RGB-D Generation

Richa Sharma,[1] **Manoj Sharma**,[2] **Ankit Shukla**,[2] **and Santanu Chaudhury**[3]

[1]*IIT Delhi, New Delhi, India*
[2]*ECE Department of Bennet University, Greater Noida, India*
[3]*Department of Electrical Engineering, IIT Delhi and Director of IIT Jodhpur, New Delhi, India*

Correspondence should be addressed to Richa Sharma; richa.sharma@ee.iitd.ac.in

Received 6 September 2021; Revised 3 October 2021; Accepted 12 October 2021; Published 11 November 2021

Academic Editor: Vijay Kumar

Generation of synthetic data is a challenging task. There are only a few significant works on RGB video generation and no pertinent works on RGB-D data generation. In the present work, we focus our attention on synthesizing RGB-D data which can further be used as dataset for various applications like object tracking, gesture recognition, and action recognition. This paper has put forward a proposal for a novel architecture that uses conditional deep 3D-convolutional generative adversarial networks to synthesize RGB-D data by exploiting 3D spatio-temporal convolutional framework. The proposed architecture can be used to generate virtually unlimited data. In this work, we have presented the architecture to generate RGB-D data conditioned on class labels. In the architecture, two parallel paths were used, one to generate RGB data and the second to synthesize depth map. The output from the two parallel paths is combined to generate RGB-D data. The proposed model is used for video generation at 30 fps (frames per second). The frame referred here is an RGB-D with the spatial resolution of 512×512.

1. Introduction

Deep learning requires a huge volume of data to train the networks. Collection of data by physically creating is a daunting task. While capturing images or videos physically, there will be some issues like a foreground objects shadow, background clutter, change in illumination, the effect of moving background objects, and viewpoint of the scene. These issues evoke the need for depth information in data. With the addition of depth as an extra dimension, useful information about the scene is gathered which is insensitive to variation of illumination. Additionally, combining the depth map with RGB gives a rich 3D scene which is close to real life experience and is very useful in various applications. Despite this requirement and with the availability of a vast variety of sensors, RGB-D data acquisition is a challenge.

For a particular application such as gesture recognition and activity recognition, till now we have two largest datasets, namely, ChaLearn gesture challenge [1, 2] and NTU RGB + D [3]. This gives rise to the need for a generation of synthetic data with or without a little intervention (to acquire reference frame) of any RGB-D sensor. There are a few networks which can generate RGB images and videos from random number.

Synthetic data created are used in training purpose for varied applications related to computer vision and also in the machine learning domain which includes scene reconstruction, camera and object tracking, pose identification, action/gesture recognition, and many more. Using these networks, we can generate scenes which are difficult to capture in real life. From the literature, we can see that the quality of synthetic data generated through generative adversarial networks has been much better than the previously used methods. GAN and its variants are one of the potentially important breakthroughs in deep learning. Though GAN has been used successfully to generate RGB videos, very little attention has been devoted to RGB + D generation.

Motivated by the importance of depth data in many applications, we are proposing a new framework for

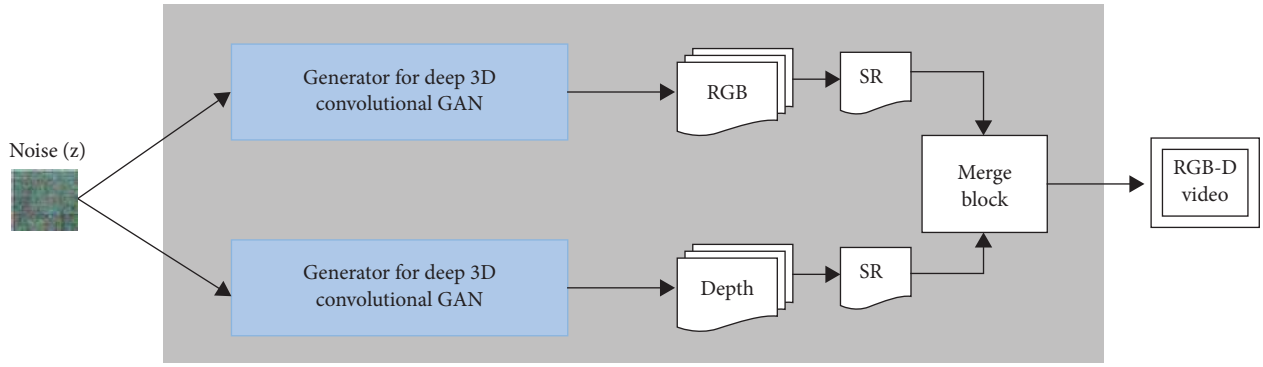

Figure 1: Block diagram of the framework.

RGB + D data generation which uses conditional deep 3D-convolutional generative adversarial network for RGB-D data generation. The proposed framework has two parallel paths. Here, each path is having conditional deep 3D-convolutional GAN, one is for generating RGB video and second one is to synthesize depth map and combine them to generate RGB-D video. Two generators are fed with a noise vector sampled from normal distribution to generate RGB-D data by two-stream conditional deep 3D-convolutional GAN. Discriminator learns to differentiate between generated synthesized videos and real videos (which are the videos from NTU RGB + D [3] dataset) for both RGB and depth videos.

The remaining paper is organized as follows. Section 2 discusses the related work. Proposed methodology is discussed in Section 3. Section 4 presents the experimental results. Section 5 concludes the paper.

2. Related Work

Ian Goodfellow gave the sophisticated architecture of GANs (generative adversarial networks) [4] for the purpose of generating data. Since then, several adaptations have been applied for various applications [5]. Early research on data synthesis was dominated by synthesis of image data using GAN and its variants. Durugkar et al. [6] developed an architecture known as generative multiadversarial network in which multiple discriminators and a single generator were used to model images. A similar system called generative adversarial parallelization was created by Daniel et al. [7, 8] in which multiple GAN pairs were used that were interchangeable during training. Initially, it resulted in an unstable result but later modified and improved and various variants of the architecture have been used since then. Radford et al. [9] developed an architecture called "Deep Convolutional Generative Adversarial Networks" acronymic to DCGANs for unsupervised representational learning of images by combining GAN and CNN architectures which was later modified, and its variants were used by many researchers. The outcome of these architectures is better than that of conventional GAN. Much of the research on GAN was confined to image generation [10], and very little has been done for video sequences.

Vondrick et al. [11] developed an architecture combining GAN [12] with 3D convolution and used as a milestone for video generation. Our work is motivated from this work in which two stream networks were used, one is to generate foreground and other is to generate background, which was then combined to produce video. It is then fed to the discriminator to discriminate probably fake output from real videos. Arjovsky et al. [13, 14] created a variant of GAN known as WGAN (Wasserstein GAN) which uses Wasserstein distance instead of Jensen Shannon distance, resulted in a more stable system. Mathieu et al. [15] created the deep multiscale system to predict future frames, but the accuracy was measured for only for few frames. Similarly, Zhu et al. [16] developed SeqGAN, and Yu et al. [17] developed CycleGAN, with the focus to resolve generator differentiation issue. Xue et al. [18] used single image instead of sequence of images, to generate future frames. Walker et al. [19] used optical flow to generate future frames. There are other GAN-related works which has been done for different applications [20–27].

The proposed paper addresses the following agendas:

(1) The generation of RGB-D data by using two-stream conditional deep 3D-convolutional generative adversarial networks

(2) Exploitation of spatio-temporal convolutional architecture for generating both depth as well as RGB videos

(3) Generation of 2 second RGB-D video with the rate of 30 frames per second where each RGB-D frame is having 512×512 spatial resolution

(4) Use of SR process to increase resolution of generated videos with good perception quality

(5) RGB prediction architecture for the future

3. Proposed Methodology

Figure 1 shows the diagrammatical flow of the proposed model. There are two streams serving for color video generation and depth video generation, respectively. At the end of each stream, super-resolution network is used to improve the quality of output video of each path. After improving the

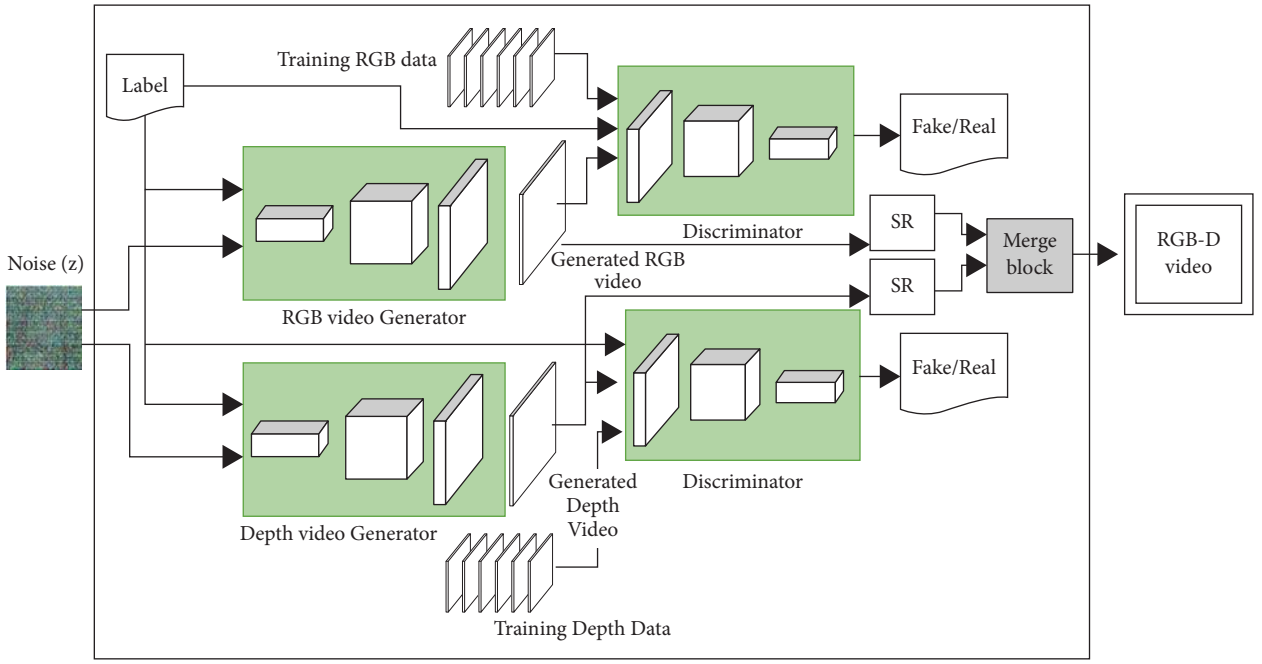

Figure 2: Detailed block diagram of the proposed framework.

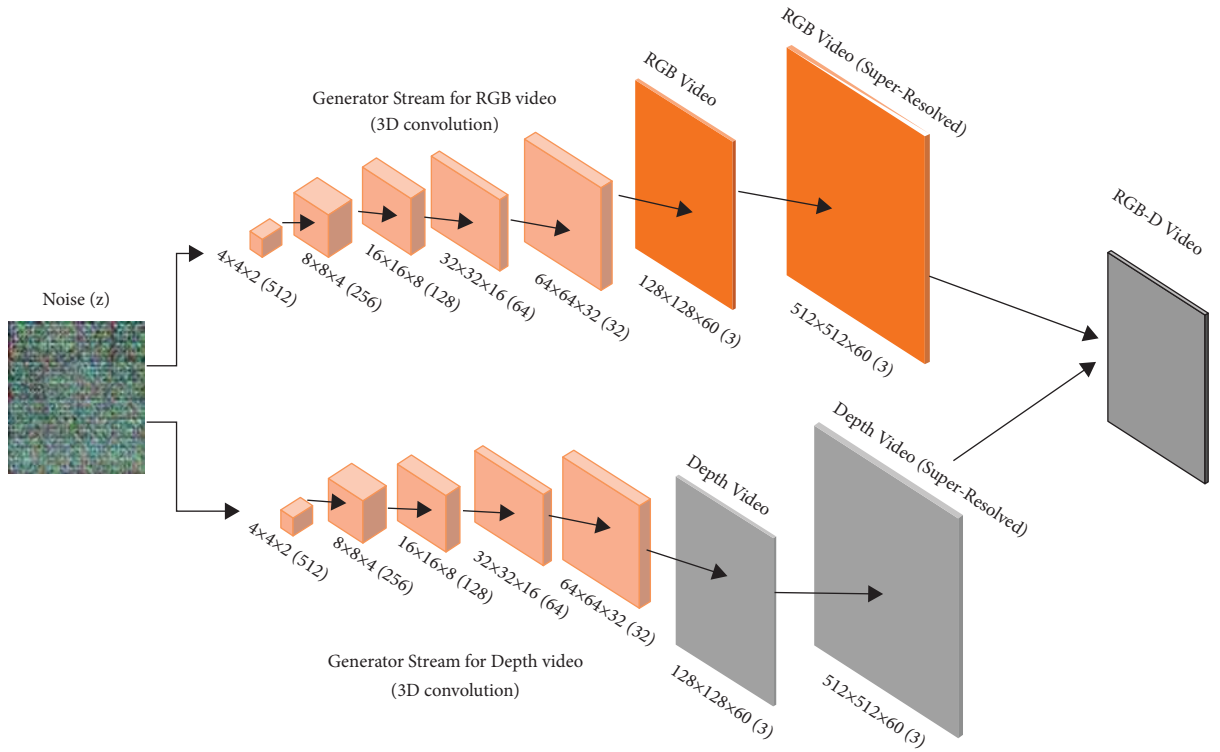

Figure 3: Detailed architecture of the generator part.

quality of both types of videos, both videos are concatenated using merge block to obtain RGB-D video.

Each stream has been implemented by a conditional GAN. In conditional GAN, the label of each type of video is provided along with noise sample to generator and same label with training data to the discriminator. The block diagram of the proposed framework is shown in Figure 2. Same noise sample with same label is used for both the generators. Same label is assigned to discriminator associated with real videos to obtain the expected video. The structure of generator and discriminator is discussed in the remaining sections.

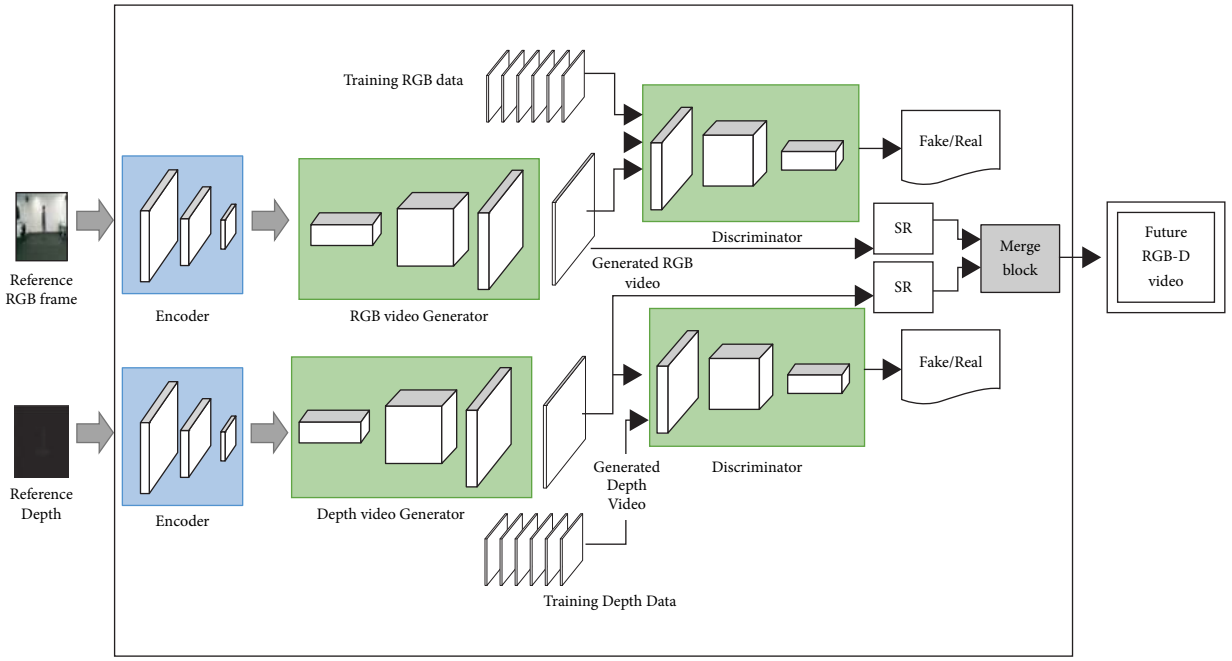

Figure 4: Block diagram for future prediction of video frames.

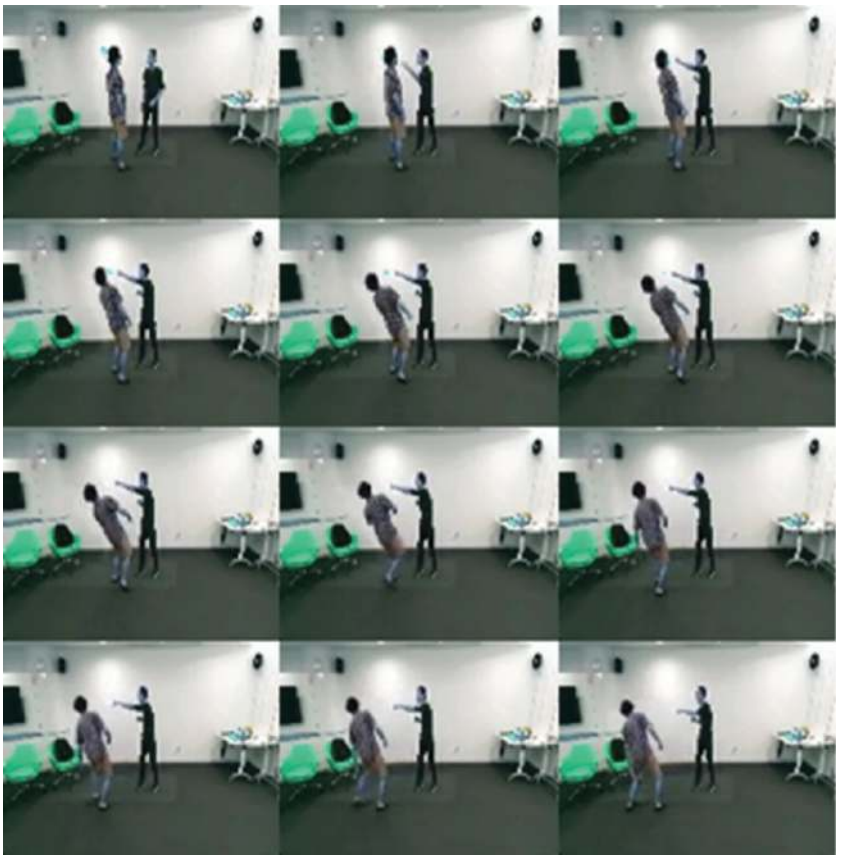

Figure 5: Ground truth or training video frames of class "kicking" with a frame gap of 5 frames.

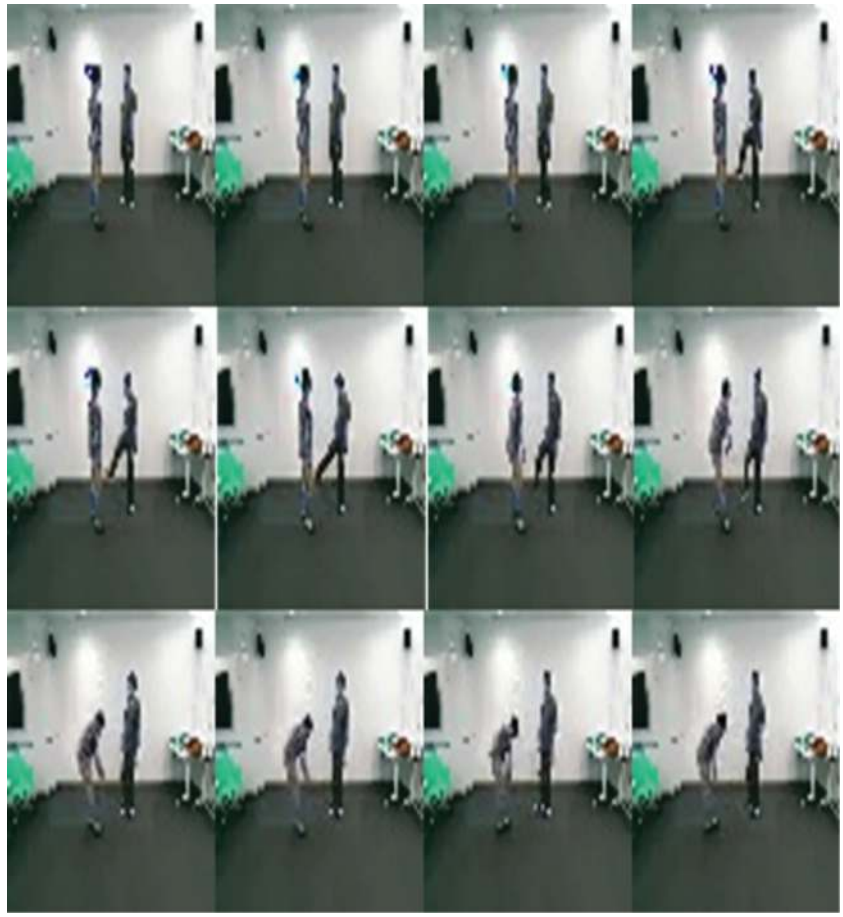

Figure 6: Generated video frames of class "kicking" with a frame gap of 5 frames.

3.1. Generator. As shown in Figures 2 and 3, we are using a noise vector of dimension 100 obtained from normal distribution. The noise vector is then concatenated with the label, and the output vector is later reshaped into a [1, 1, 1, 106].

Since we have only 6 classes of activity, we use one-hot encoding technique to get the label vector. To utilize both spatial and temporal information, a 3D convolution-transpose operation is performed over this reshaped vector with the kernel size of [2, 4, 4] where 4 and 4 are height and width, respectively, and 2 is the depth of the filter. 512 filters are used in the first convolution-transpose layer with the stride of [1, 1, 1]. The next five layers use the same 3D convolution-transpose operation having the kernel size of [4, 4, 4] with 256, 128, 64, 32, and 3 filters, respectively. To increase the size of image, we use stride of 2 in each dimension. The output shape of the last layer is 64, 128, and 128 which represents height and width of output video as 128×128 and 64 frames in depth. Rectified Linear Unit (ReLU) has been used as an activation function in all layers. We use only 60 frames of the last layer out of 64. The purpose of doing is to create 2 second video at 30 fps requiring only 60 frames. The final output of generator is 60 frames of dimension 128×128.

This generator network is replicated into two streams to generate RGB and depth video. The same label is fed into the depth video generating stream. After the generating both videos, we use super-resolution network to improve the perceptual quality of generated videos. Later, both videos have been merged to obtain RGB-D video. The output is the RGB-D video of 2 second length saved at 30 frames per second. The dimension of the output video is the same as the final output layer of each generator network which is 128×128.

3.2. Discriminator. The job of discriminator is to act as a classifier. It must distinguish between the real video and fake video. As we are using conditional GAN, the classification is also based on label. The discriminator comprises five 3D convolution layers having a kernel dimension of [4, 4, 4] at each layer except for the last layer which is [2, 4, 4]. The spatio-temporal information of both the videos is studied with the help of 3D convolution operation. The filter size of each layer is 64, 128, 256, 512, and 1, respectively. Before feeding the real video into the first layer of network, the label which is of one-hot encode in nature is reshaped into the same size of input video frame which is 128×128. After reshaping, it concatenates with input video frame and goes to the first layer of discriminator network. The leaky-ReLU activation function is basically employed in the first three layers, and the ultimate layer makes use of the sigmoid activation function.

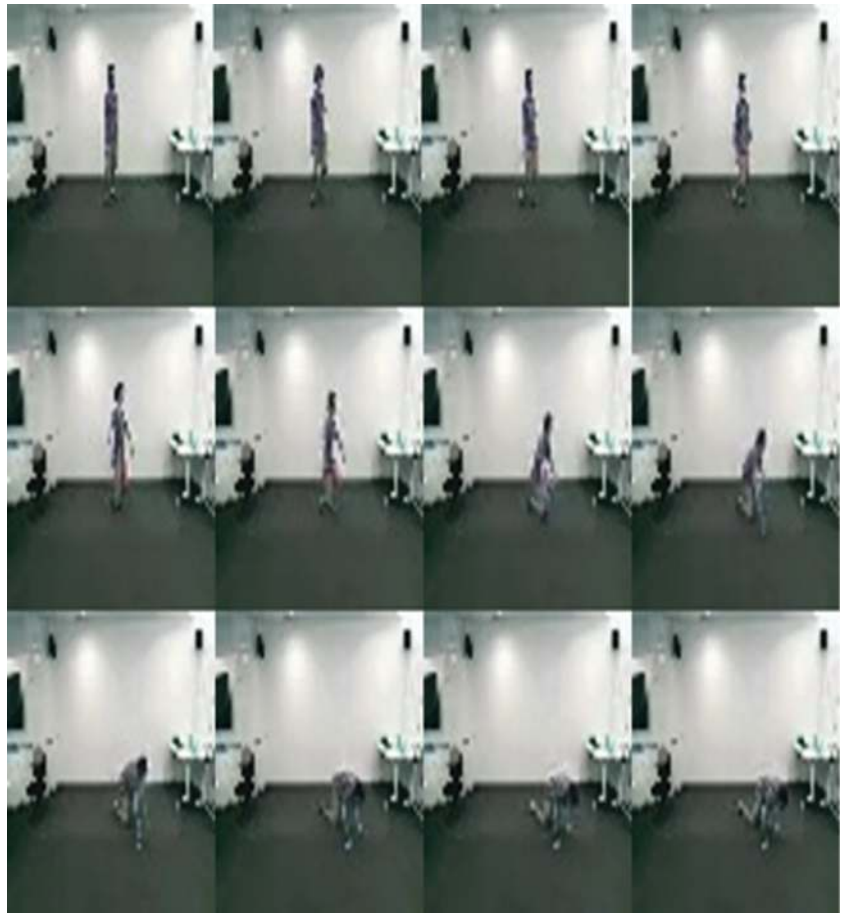

Figure 7: Generated video frames of class "sitting on knee" with a frame gap of 5 frames.

3.3. Training. The training data are augmented using various kinds of data augmentation techniques to prevent overfitting issue. The used data augmentation techniques are image rotation, image cropping, and image filters. Thereafter, all images are resized to the actual resolution. To build model, we use two steams one for generation of RGB video and other for depth video. The above-described generator and discriminator model is replicated in two streams. Each generator and discriminator are given proper training with the aid of cross entropy loss methodology. The optimizer is optimized at a learning rate of 0.002. The model is to be trained for 1000 epochs. In every epoch, we train our model for all the video in training folder, and after each complete iteration, we are generating sample video as well as saving the model file.

3.3.1. Super-Resolution (SR). Here, we are using 3D-CDCA [28] to 4x super-resolve the RGB video frames. This 4X super-resolution increases the spatial resolution of RGB frames from 128×128 to 512×512 and temporal 4X-SR increases the number of frames 4 times than earlier. For depth SR, we are simply interpolating space-time video frames by tri-cubic interpolation. This SR block helps in improving spatio-temporal resolution of the generated RGB and depth videos.

3.4. Future Generation: Prediction of Next Frames. We have adapted our previous proposed framework for future frames prediction as shown in Figure 4. Here, our input is static frames or reference frames, and we are predicting next frames, so the output is future frames. The working methodology is as follows. The reference frame is fed into a convolutional encoder (CE). The CE learns the features of reference frame and reduces the size of the reference frame equal to the first layer of generator network. Then, our proposed framework generates the video by creating the next future frames.

4. Experimental Results

4.1. Dataset. NTU RGB+D action recognition dataset containing 56,880 RGB videos each with a resolution of 1920×1080 along with the depth map has been implemented for our experiments. Videos of this dataset have 40 videos of classes of daily action, 9 classes of videos of medical conditions, and 11 classes of mutual conditions for each subject. For conditional video generation, we use only 6 classes. The NTU RBG+D till date is the largest dataset assembled by taking 40 subjects performing 60 different action classes which are subdivided as daily actions, mutual actions, and medical conditions. Daily actions include 40 days to day actions classes such as drinking water, jumping up, taking selfie, and

Reference Frame Predicted Frames

Figure 8: Future prediction: here first frame is static or reference frame and next six are future predicted frames with the gap of 10 frames.

Reference Frame Predicted Frames

Figure 9: Future prediction: here first frame is static or reference frame and next six are future predicted frames with the gap of 10 frames.

pointing to something. A total of 11 action classes are categorized as mutual conditions such as kicking, hugging, and punching, and 9 action classes are considered as medical conditions such as sneezing, vomiting, and falling. We have shown original video frames in Figure 5.

5. Result Analysis

Using this dataset, we generated videos as shown in Figures 6 and 7. Figure 7 shows videos of subjects hugging each other, and Figure 8 shows videos of fight. Figures 6 and 7 are of classes kicking and sitting on chair, respectively. All of the videos are 2 seconds long and have 60 frames per second. We are showing only 12 frames per video.

It is very difficult to measure the quality of generated video, and since there is no one to correspond between real data and generated data, calculated mean squared error is not suitable for this kind of quality measurement. Additionally, mean squared error does not reflect the human perception of reality. A commonly adapted tool is Amazon Mechanical Turk, commonly known as MTurk. We conducted a survey on Amazon Mechanical Turk, motivated from [29], to get the subjective MOS based on perceptual quality of generated videos. Our aim is to measure the perception quality of the motion in the generated video. We categorize the MOS (mean opinion score) in five categories such as bad, poor, fair, good, and excellent, score of 1 for bad (least perceivable) and 5 for excellent (most perceivable and realistic). For each HIT (Human Intelligent Test), subjects are asked to rate the videos of two seconds. A total of 1200 ratings were collected from more than 60 unique subjects, and the survey is still ongoing. Some subjects were rejected on the account of reliability, and no subject was allowed to take the survey more than once. By averaging the rating for individual video, we obtain MOS for our generated videos. 76 percent of the rating lies in the range of 3 to 5. By analyzing the scores, we can say that our generated videos look realistic and perceivable with its motion. Figures 8 and 9 show the output of future prediction of frames using a reference frame of any particular action. In these figures, output predicted frames are shown with the gap of 10 frames.

The quantitative analysis on testing dataset indicates that the proposed model obtains false-positive and false-negative values as 1.37% and 1.87%, respectively. These values also indicate that the proposed model does not suffer from the overfitting issue.

6. Conclusions

The proposed conditional deep 3D-convolutional generative adversarial network can generate more realistic videos of 2 second length as shown in the experimental results. This framework has proven to be more promising for predicting future frames. In the proposed framework, the super-resolution framework enables us to produce the video of high spatio-temporal resolution. In addition, we have generated RGB-D video for each action class. This RGB-D data generation will help in many computer-vision applications as well as help us in understanding of features responsible for action recognition of different classes. In future, we will explore CNN-LSTM-based generative architecture on the lines of the proposed architecture for RGB-D generation with more realistic quality and will develop end-to-end pipeline that contains data generation with application in improving action recognition accuracy for many classes.

Data Availability

The data that support the findings of this study are available upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] P. Vitoria, J. Sintes, and C. Ballester, "Semantic image inpainting through improved wasserstein generative adversarial networks," 2018, https://arxiv.org/abs/1812.01071.

[2] H. Huang, "Introvae: introspective variational autoencoders for photographic image synthesis," 2018, https://arxiv.org/abs/1807.06358.

[3] R. Theagarajan and B. Bhanu, "DeephESC 2.0: deep generative multi adversarial networks for improving the classification of hESC," *PLoS One*, vol. 14, no. 3, p. e0212849, 2019.

[4] M. Firman, "RGBD datasets: past, present and future," 2016, http://arxiv.org/abs/1604.00999.

[5] J. Heaton, "Ian goodfellow, yoshua bengio, and aaron courville: deep learning," *Genetic Programming and Evolvable Machines*, vol. 19, no. 1-2, pp. 305–307, 2018.

[6] I. J. Goodfellow, "NIPS 2016 tutorial: generative adversarial networks," 2017, http://arxiv.org/abs/1701.00160.

[7] H. Huang, P. S. Yu, and C. Wang, "An introduction to image synthesis with generative adversarial nets," 2018, https://arxiv.org/abs/1803.04469.

[8] X. Wei, "Improving the improved training of wasserstein gans: a consistency term and its dual effect," 2018, https://arxiv.org/abs/1803.01541.

[9] Y. Enokiya, "Automatic liver segmentation using U-Net with Wasserstein GANs," *Journal of Image and Graphics*, vol. 7, pp. 94–101, 2018.

[10] Y. Ji, H. Zhang, and Q. M. Jonathan Wu, "Saliency detection via conditional adversarial image-to-image network," *Neurocomputing*, vol. 316, pp. 357–368, 2018.

[11] J. I. Daniel, M. He, C. D. Kim, and W. Graham, "Generative adversarial parallelization," 2016, http://arxiv.org/abs/1612.04021.

[12] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, http://arxiv.org/abs/1710.10196.

[13] Z. Luo, B. Peng, D.-A. Huang, A. Alexandre, and L. Fei-Fei, "Unsupervised learning of long-term motion dynamics for videos," 2017, http://arxiv.org/abs/1701.01821.

[14] Z. Chai, "CMS-LSTM: context-embedding and multi-scale spatiotemporal-expression LSTM for video prediction," 2021, https://arxiv.org/abs/2102.03586.

[15] M. Mirza and O. Simon, "Conditional generative adversarial nets," 2014, http://arxiv.org/abs/1411.1784.

[16] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. P. Singh, "Action-conditional video prediction using deep networks in atari games," 2015, http://arxiv.org/abs/1507.08750.

[17] Y. Wei, X. Luo, L. Hu, Y. Peng, and J. Feng, "An improved unsupervised representation learning generative adversarial network for remote sensing image scene classification," *Remote Sensing Letters*, vol. 11, no. 6, pp. 598–607, 2020.

[18] Y. Guo, K. Kshatri, E. D. Matsumoto, and A. Kapoor, "Expert and crowdsourced evaluation of image quality from a novel endoscopy phone light adapter," *Urology*, vol. 146, pp. 54–58, 2020.

[19] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," 2016, http://arxiv.org/abs/1606.03498.

[20] K. Papadopoulos, "Vertex feature encoding and hierarchical temporal modeling in a spatial-temporal graph convolutional network for action recognition," 2019, https://arxiv.org/abs/1912.09745.

[21] D. Singh, V. Kumar, V. Yadav, and M. Kaur, "Deep neural network-based screening model for COVID-19-infected patients using chest X-ray images," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 35, no. 3, p. 2151004, 2021.

[22] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July 2017.

[23] N. Gianchandani, A. Jaiswal, and D. Singh, "Rapid COVID-19 diagnosis using ensemble deep transfer learning models from chest radiographic images," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, 2020.

[24] J. Wallach, H. M. Berger, and P. D. Greene, "Affective overdrive, scene dynamics, and identity in the global metal scene," *Metal Rules the Globe*, Duke University Press, Durham, NC, USA, 2011.

[25] D. Singh, V. Kumar, and M. Kaur, "Densely connected convolutional networks-based COVID-19 screening model," *Applied Intelligence*, vol. 51, no. 5, pp. 3044–3051, 2021.

[26] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and C. Bryan, "High-resolution image synthesis and semantic manipulation with conditional GANs," 2017, http://arxiv.org/abs/1711.11585.

[27] T. Xue, J. Wu, K. Bouman, and B. Freeman, "Visual dynamics: probabilistic future frame synthesis via cross convolutional networks," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds., Curran Associates, Inc., New York, NY, USA, 2016, http://papers.nips.cc/paper/6552-visual-dynamics-probabilistic-future-frame-synthesis-via-cross-convolutional-networks.pdf.

[28] Y. Yan, "Sequence generative adversarial nets with a conditional discriminator," *Neurocomputing*, vol. 429, pp. 69–76, 2021.

[29] J.-Y. Zhu, T. Park, P. Isola, and A. Alexei, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 2017, http://arxiv.org/abs/1703.10593.