Soft Computing (2022) 26:13153–13175
https://doi.org/10.1007/s00500-022-07014-x

FOCUS

Dynamic GAN for high-quality sign language video generation from skeletal poses using generative adversarial networks

B. Natarajan[1] · R. Elakkiya[1]

Accepted: 16 March 2022 / Published online: 28 June 2022

Abstract

The recent advancements of unsupervised deep generative models have produced incredible results in image and video generation tasks. However, existing approaches still pose huge challenges in the high-quality video generation process. The generated videos consist of blurred effects and poor video quality. In this paper, we introduce a novel generative framework named dynamic generative adversarial networks (dynamic GAN) model for regulating the adversarial training and generating photo-realistic high-quality sign language videos. The proposed model uses skeletal poses information and person images as input and produces high-quality videos. In generator phase, the proposed model uses U-Net-like network to generate target frames from skeletal poses. Further, the generated samples are classified using the VGG-19 framework to identify its word class. The discriminator network classifies the real and fake samples as well as concatenates the resultant frames and generates the high-quality video output. Unlike, existing approaches the proposed novel framework produces photo-realistic video quality results without employing any animation or avatar approaches. To evaluate the model performance qualitatively and quantitatively, the proposed model has been evaluated using three benchmark datasets that yield plausible results. The datasets are RWTH-PHOENIX-Weather 2014T dataset, and our self-created dataset for Indian Sign Language (ISL-CSLTR), and the UCF-101 Action Recognition dataset. The proposed model achieves average 28.7167 PSNR score, 0.921 average SSIM score, 14 average FID score and 8.73 $\pm$ 0.23 average inception score which are relatively higher than existing approaches.

1 Introduction

Sign language greatly improves the communication skills of the deaf-mute community as well as explores the needs and emotions of such people. Sign languages are highly structured, visual conveying, and multi-channel based one, expressed via gestures and utilizes human upper body parts such as hands, face, eyes and gaze movements (Elakkiya and Selvamani 2017, Elakkiya 2021). These components are usually termed as manual components for hands actions and non-manual components for facial and mouth expressions. In most countries, sign languages are developed based on their own culture, traditions, and surroundings. These variations are referred to as multimodal sign gestures for multilingual sentences. Recognition and translation of such different variations in sign gestures create numerous challenges to the researchers and require expert skills in computer vision and artificial intelligence domains. Research studies on sign language recognition and translation attained wider attention around the globe. The development of such systems assists normal people to easily communicate with deaf-mute people to provide training and education services. These objectives highly motivated us to develop such systems. Automation of these translation processes with the help of high-power computing devices will raise the digital technology advancements to the next level. The growth of hardware technology handles such high-level computing tasks using GPU devices. The proposed work is aimed to develop the software framework for translating of skeletal pose into sign language videos using deep generative networks. The proposed model creates a single unified framework to process multimodal skeletal poses and translate them into human-based sign gesture images and combines the gesture image sequences for video generation. The earlier

Communicated by Joy Iong-Zong Chen.

✉ R. Elakkiya
elakkiyaceg@gmail.com

[1] School of Computing, SASTRA Deemed to be University, Thanjavur, Tamil Nadu, India

generative models using deep learning approaches have reached various milestones by producing impressive results in image and video generation. The existing approaches like auto-encoder and its variants VAE and CVAE (Larsen et al. 2016; Xu et al. 2019; Pu et al. 2016) generate images with blurred effects. The quality of generated results does not comply with the expectations. The recent approaches in generative adversarial networks have been attained wider attention among researchers for developing various applications like synthesizing medical images (Nie et al. 2017; Elakkiya et al. 2021), text-to-image translation (Reed et al. 2016; Smys and Haoxiang 2021), video analytics (Wu et al. 2019), sentiment analysis (Pandian 2021) and creating human images that do not exist in the world (Beschizza 2019). This powerfulness of the GAN models directs the researchers to develop efficient models to generate high-quality images or videos.

However, the processing of a large number of images or videos and producing new images or video potentially requires high expert skills. Research and development of such models explore the capability of generative networks to the next level. Predicting the future frames (Cai et al. 2018), video completion (Cai et al. 2018) and video generation (Cui et al. 2019; Clark et al. 2019; Gao et al. 2019) showcases the current improvements in GAN model development. These advancements in the GAN techniques can be applied to generate high-quality photo-realistic sign videos from skeletal poses for the betterment of the deaf-mute community. In this paper, the development of the dynamic GAN model is divided into various folds. In the first fold, the mapping of skeletal poses and ground truth images takes place, and then the generator network generates human-based sign gesture images. In the next fold, we apply the image classification using VGG-19 and image alignment techniques. Further, we apply deblurring techniques (Shan et al. 2008) for improving the image quality. The generation of intermediate frames for connecting the sequences of gestures has been carried out in the proceeding steps. Finally, the discriminator network produces a photo-realistic high-quality sign video generation process by checking the reality of images. In the case of fake images, the model redirects the results to the generator to undergone fine-tuned training for generating high-quality results.

The primary invention of the generative adversarial network (GAN) framework (Goodfellow et al. 2014) greatly scaled up the growth of the deep generative model to generate high-quality images or videos. These models have attained greater attention among researchers to develop powerful models for high-quality image or video generation. The two networks of the GAN model play the minimax adversarial game competently to produce high-quality videos. The generator networks aim to produce images similar to the real ones from random noise vectors. The discriminator network classifies the real and fake images intelligently shown in Fig. 1. Based on such classification, the generator network fine-tunes its training performance to produce good quality videos which mimic real videos. In a first version of GAN (Goodfellow et al. 2014), the multi-layer perceptron-based fully connected layers and the activation function ReLU is applied in the generator network side and maxout activations are applied in the discriminator network. The model has been evaluated using the benchmark datasets such as the MNIST handwritten digits dataset and multi-class images-based CIFAR-10 dataset. The basic model has been upgraded to various levels to achieve greater emoluments in multiple domain datasets. The extended model known as DCGAN (Radford et al. 2015) was implemented for stabilizing the training process in the generator end using deep CNN approaches. The extended version of the basic GAN model is called as conditional GAN (Mirza and Osindero 2014) model which applies conditioning on class labels to produce high sharpened results in the generation of new digits images using the MNIST handwritten digits dataset depicted in Fig. 2.

The InfoGAN models (Chen et al. 2016) utilize the latent space for encompassing semantic label information with real images for generating improved quality images. Although it produces impressive results, the quality of images needs to be improved. Auxiliary classifier GAN (Odena et al. 2017) model employs the conditional GAN for conditioning the class labels and adds the auxiliary models for reconstructing the class labels. This model adds more complexity to produce the plausible results. The development of Stack GAN models (Zhang et al. 2017, 2019) uses hierarchical stacked approaches combined with conditional GAN networks for generating images from text. This model follows two stages of development. In the first level, it produces images based on the text by applying conditioning on text data, which results in the low-resolution images. In the second stage, this model improves the results by conditioning low-resolution images and text. Finally, it produces high-resolution images. Due to stepwise improvements and capability to handle only low resolution images, it fails to perform well for high resolution image scenarios. The context encoders (Pathak et al. 2016) use adversarial approaches to generate conditioned images by applying conditions on its surrounding parts. This model uses reconstruction loss and adversarial loss approaches to yield sharpened results. This model investigates mainly on missing portions of image to predict its actual pixels. The pix2pix generative model (Isola et al. 2017) extends the basic framework of GAN models to uplift its performance in the image to image translation tasks. This model incorporates the U-Net

framework in the generator phase and applies the PatchGAN framework in discriminator phases for supporting different domain applications like generating photos from semantic labels, black and white images to color image translation, edges to real image conversion, day and night scene translation, photo editing and creating new effects. We incorporate these techniques to translate the pose skeletal sequences into realistic human image.

The recent advancements in generative adversarial networks have been greatly improved the GAN performance to the next level by generating photo realistic images. The variant of the GAN network referred to as Wasserstein generative adversarial network (WGAN) (Arjovsky et al. 2017) introduces critic which alters the training steps for updating the discriminator network. The Wasserstein loss functions are introduced in this model for improving the output image quality. The cycle-consistent generative adversarial network (CycleGAN) (Zhu et al. 2017b) has been developed for performing the image-to-image translation tasks without using conditioned target images. This model follows reversible approaches to produce one form to another by utilizing the cycle consistent approaches. The progressive GAN (Karras et al. 2017) models emerged with new approaches for training the generative networks. This model adds extra layers to stabilize and progressively improve the performance of the model and yields low quality images. The BigGAN model (Brock et al. 2018) improves the image quality by scaling up the existing conditioning models and changes the training parameters. The use of the truncation trick in latent space highly boosts the model performance. The StyleGAN (Karras et al. 2019) models use different latent space embedding techniques to synthesis the images. It controls the features of output images by inputting the latent code with different points. Although these models produce significant improvements in frame quality, the training process requires higher stability and fails in case of diverse inputs.

Although a lot of advancements were proposed in various papers, there is a great demand for the development of a single unified generative framework to produce high-quality images or videos for diverse domains. The proposed dynamic GAN models introduce novel techniques to effectively train the generator models and applying novel selection of intelligent techniques to improve the frame quality in terms of variation, texture, edge sharpening, and diversity which lead to the production of photo-realistic sign videos from skeletal poses. The intermediate frame generation and video completion approaches lead the discriminator network to classify the generated videos as the real ones. From a development point of view, the translation of human skeletal pose images into sign videos incurs huge challenges in model development and also needs to address the bottlenecks of conventional sign language recognition tasks. The execution order of sign gestures is highly differing from the word order of spoken language sentences. To address this issue, we introduce novel image alignment techniques for arranging the sign gesture images. The selection of relevant sign images and generating the in-between frames requires much attention for the video completion process. In the output videos, we consider the video quality by avoiding the collision of sign gestures. In addition to this, preserving the naturalness and identification of epenthesis movements, resolving the gesture ambiguities, co-articulation issues, and ill-posedness (Elakkiya and Selvamani 2019, 2018) are also considered for generating good quality results. The continuous recognition of the dynamic changes of sign gestures related to spoken sentences poses huge challenges. On the other hand, processing large-scale datasets with multimodal features. We need to consider all these challenges for developing the powerful framework. Finally, we conclude that, the proposed model aimed to address the common challenges of generative models such as stabilizing the training process, handling the diverse inputs, generating high-quality videos (Wang et al. 2021) and issues persist with video sequence alignment and intermediary frame generation.

Our major contributions towards the development of proposed model are listed as follows.

- We develop a novel GAN framework for generating photo-realistic high-quality sign language videos by processing the skeletal pose images and ground truth images.
- We evaluate the proposed model performance qualitatively and quantitatively using different benchmark datasets such as RWTH-PHOENIX-Weather 2014T dataset, the ISL-CSLTR dataset, and the UCF-101 Action recognition dataset.
- We build a single unified architecture for generating videos in diverse domains such as action recognition, analysis of human behavior in public, and monitoring the activities of people in a crowded environment.

Further discussions about this work are planned as follows. The existing developments present in generative models were discussed in Sect. 2, the proposed system and implementation details were discussed in Sect. 3. In Sect. 4, the experimental results on benchmark datasets are discussed and finally the conclusion and future work part summarize the entire work.

2 Related works

Research studies on high-quality video generation using latent space data points have been identified as a challenging task since the last centuries. Due to the mode

collapse, robustness, instability, scalability, and inconsistent results, the earlier approaches on video generation produces low-resolution videos. The generation of high-quality photo-realistic videos requires a lot of supervision on the distribution of data points present in the latent space (Gulrajani et al. 2016). Each data points placed in a latent space contribute some portion to the video generation and maps the relationship exists between sign gestures sequences. Further, it provides support for automate the high-quality video generation process. The earlier approaches (Goodfellow et al. 2014; Mirza and Osindero 2014; Salimans et al. 2016; Liu and Tuzel 2016; Isola et al. 2017; Ma et al. 2017; Siarohin et al. 2018; Elakkiya 2021; Saito et al. 2017; Efros and Leung 1999) discusses the generation of images or videos from noise vector by randomly selecting some data points. Due to a lack of efficient training process and various factors, these models mostly produce blurry and inconsistent results. Although, the latent space provides necessary information about existing data points, still needs some efficient mechanism to enhance the selection of data points to produce high-quality photo realistic videos. The emergence of GAN models handles such image or video generation tasks efficiently using generator and discriminator networks. The production of sign videos needs much attention in selecting latent space data points due to the variants of input sentences and dynamic changes in selecting sign gesture images and incorporation of spatial and temporal features to produce videos. In order the preserve the consistency in output quality, we need to investigate various techniques for producing fine-grained human perceptual results.

In general, machine learning models can be classified as discriminative and generative models. Usually, the discriminative models work well for classification-based tasks like spam detection in email. On the other hand, generative models are powerful in creating samples based on underlying data distribution. The variants of generative models are parametric and nonparametric approaches. Parametric approaches are highly used for image reconstruction purposes, whilst nonparametric approaches are highly used in text synthesis (Efros and Leung 1999) and speech synthesis (Saito et al. 2017) processes. Learning the low-dimensional details of data distribution supports the image reconstruction process (Xie et al. 2020). Mostly the popular models such as deep auto-encoders (Xu et al. 2019; Pu et al. 2016) or the generative model known as restricted Boltzmann machines (RBM) (Zhang et al. 2017; Denton et al. 2015; Choi et al. 2018, 2020; Zhu et al. 2017) were primarily used for generating images. Due to higher complexity in generating images, these techniques are found as less effective one. The advent of variational auto-encoders (VAE) models (Xu et al. 2019) resolves this issue by adopting a variational sampling-based approaches, but which are limited to small scale datasets such as MNIST. The generation of human recognizable images with rich content the VAE model was stepped in new advancements (Xu et al. 2019; Pu et al. 2016).

GAN models are introduced by the author (Goodfellow et al. 2014) discuss the adversarial training process by placing two players (generator and discriminator) in a game of competing with each other using minimax approaches. The maximization of the first player score will minimize the second player score vice versa. This discriminative process aims to produce handwritten images, faces, and objects. The primary model was targeted to achieve global optimum by matching the produced results with original data. This model produces blurred results that need to be improved using conditional-based approaches and inference mechanisms. This model considers only a specific portion of the data distribution, divergence, and oscillation nature tends to training difficulties. The conditional-based GAN models (Mirza and Osindero 2014) apply conditional probabilistic rules on both the generator and discriminator sides to generate improved results. These models apply conditions on some portion of data. Applying conditioning on class labels over handwritten digit datasets and highly capable to learn the multimodal features. The condition-based predictive distribution produces good results over the learned distribution of data and results in the deterministic relationship between input and output. The conditional GAN models lead the development of image to image translation models (Salimans et al. 2016), face generation (Liu and Tuzel 2016), face aging (Isola et al. 2017), domain adaption models for alignment of multimodal features (Huang et al. 2018; Zhu et al. 2017a), image captioning(Pu et al. 2016), machine translation (Yang et al. 2017), text to image synthesis (Saito et al. 2017; Efros and Leung 1999).

The least-square GAN model (Mao et al. 2017) was developed to address the vanishing gradient issues persist with discriminator classifier by employing the decision boundary-based penalization strategies. Although, this model produces comparable results over the regular GAN model, requires much improvement for generating real images by automating the penalization steps. The 3CATN (Li et al. 2019) model is proposed to address the challenges in adversarial training that still need to be improved for unsupervised learning tasks. The StackGAN (Zhang et al. 2017) models follow two-stage approaches for translating the text scripts into real images. In the first stage, it creates outline-based low-resolution images, after applying condition-based augmentation techniques it produces photo-realistic results. However, these models fail to perform well on real-time image generation tasks due to model collapse and instability issues. The author (Denton et al. 2015) was developed Laplacian pyramid-based GAN

Table 1 Comparison of various generative models

Author	Model	Dataset	Metric	Image/ video
Goodfellow et al. (2014)	GAN	MNIST, TFD, CIFAR-10	log-likelihood estimate	Image
Salimans et al. (Salimans et al. 2016)	Improved GAN	MNIST, CIFAR-10, SVHN	Test error rate	Image
Ma et al. (2017)	PG^2 (pose guided person generation network)	Deep Fashion, Market-1501	SSIM, IS	Image, video
Siarohit et al. (2018)	Deformable GAN	Deep Fashion, Market-1501	SSIM, IS	Image, video
Mirza and Osindero (2014)	Conditional GAN	MNIST, MIRFlickr 25,000 dataset	Log-likelihood estimate	Image
Isola et al. (2017)	Pixel GAN	Cityscape dataset	FCN Score	Image
Shishir et al. (2020)	EsharaGAN	IsharaLipi dataset	IS	Image, video
Stoll et al. (2020)	Pix2pixHD + VAE-GAN	SMILE sign language dataset	SSIM, PSNR, MSE	Image, video
Zhao et al. (2018)	VariGANs	MVC, deep fashion	SSIM, IS	Image, video
Tulyakov et al. (2018)	MoCoGAN	Taichi video clips, MUG facial expression dataset	Motion control score (MCS), IS, average content distance (ACD), user preference score	Image, video
Vondrick et al. (2016)	VGAN	Flickr	User preference score	Image, video
Saito et al. (2017)	TGAN	Moving MNIST, UCF-101, Golf scene dataset	IS, GAM (generative adversarial metric)	Image, video
Arjovsky et al. (2017)	WGAN	LSUN Bedrooms dataset	Earth mover (EM) distance	Image, video
Radford et al. (2015)	Deep convolutional GAN	CIFAR-10, STL, LSUN	Classification accuracy	Image, video
Wang and Gupta (2016)	S^2GAN	$NYUV^2$ dataset	Classification accuracy	Image, video
He et al. (2018)	VideoVAE	Chair CAD, Weizmann human action dataset, YFCC, MIT Flickr	IS	Image, video
Wang et al. (Wang et al. 2019a)	P2P video generation	Moving MNIST dataset, Human 3.6 M, BAIR Robot pushing dataset, Weizmann action dataset	PSNR, SSIM, MSE	Image, video
Aigner and Körner (2018)	FutureGAN	Moving MNIST, Cityscape, KTH action	PSNR, SSIM, MSE	Image, video
Karras et al. (Karras et al. 2017)	PGGAN	CIFAR-10, CelebA, LSUN	IS	Image, Video

PSNR peak signal-to-noise ratio, *SSIM* structural similarity index measure, *IS* inception score, *MSE* mean squared error, *FCN* fully-convolutional semantic segmentation network

model (LPGAN). This model uses down sampling and up sampling methods to construct high-quality images places high complexity in the simple image generation process and bottlenecks with sequence images in video generation tasks. The large scale GAN models (Brock et al. 2018) are developed to handle huge classes of Imagenet database natural images using orthogonal regularization techniques. Although this model suits large-scale training, model collapse may occur and results in low-resolution images or videos. The author (Kim et al. 2017) proposed DiscoGAN models for identifying and classifying different domain objects and applied style transfer mechanisms to change

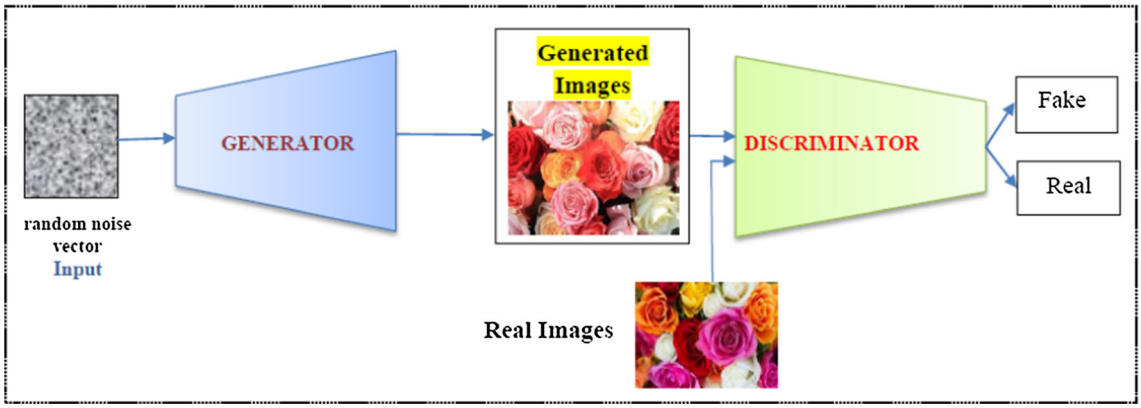

Fig. 1 Basic GAN model

the objects of one domain to another domain without losing identity key points. In Dosovitskiy et al. (2015), author introduced the chair image generation using CNN Techniques. The Laplacian pyramid GAN models (Denton et al. 2015) are developed to produce fine-grained images from coarse-grained inputs but fail to generate stable results. However, the DCGAN model provides promising results, requires the conditioning of class labels at each time step. The author (Mathieu et al. 2015) implemented the new approach for predicting the future frames by applying conditions on the previous frames. The author (Tulyakov et al. 2018) introduced the MoCoGAN model for dealing with the motions and content features in the video generation process. The random vector sequences are trained properly for generating sequence of frames in order, produces high-quality videos.

The FUNIT model (Liu et al. 2019) generates multiple domain images few-shot-based image translation approaches. This model requires fully class labeled image datasets and fails to handle the dynamic generation of frames for image or video creation. In parallel work, the author (Karras et al. 2019) combines multimodal and multi-domain translation by employing the manipulation schemes in latent spaces. These models limit their performance for learning various styles of multi-mapping translation tasks. In Choi et al. (2018) author developed the StarGAN model using one generator and one discriminator networks for the translation of image to image tasks and supports scalable services. This model aims to provide high-quality images owing to the generalization and multitasking capabilities. Besides, the simple mask vector concepts enrich the model performance to operate well in multiple domains. Since this model handles multiple domain inputs, fails to incorporate the distinguishable features. The revised version StarGAN v2 (Choi et al. 2020), produces improved results in multiple domains and handles diversity and scalability issues very well. Since all these models are implemented for working with inter-domain and intra-domain-based applications, need to be enhanced for video generation in multiple domains by adopting different features. The GAN model and the advancements have been found significantly important for learning the structure of deep generative models to generate images or videos similar to real-time data. However, the

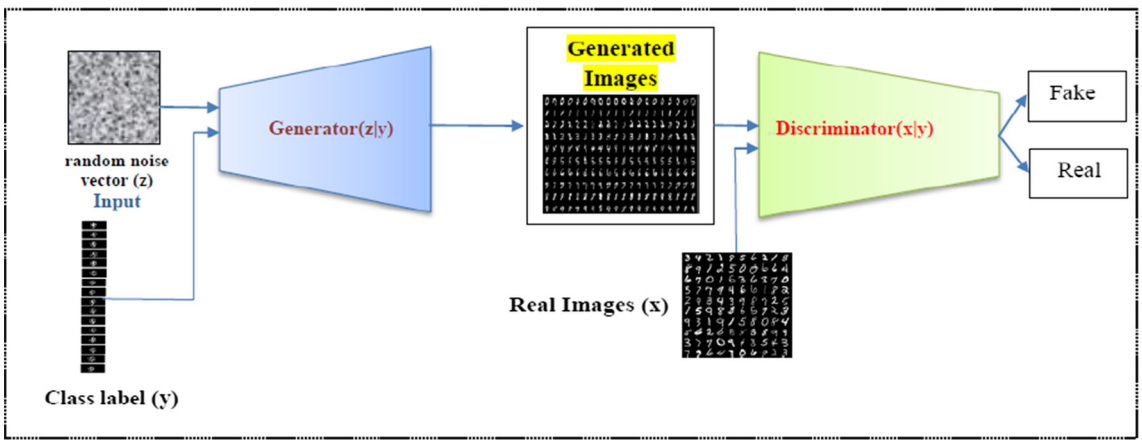

Fig. 2 Conditional GAN Model

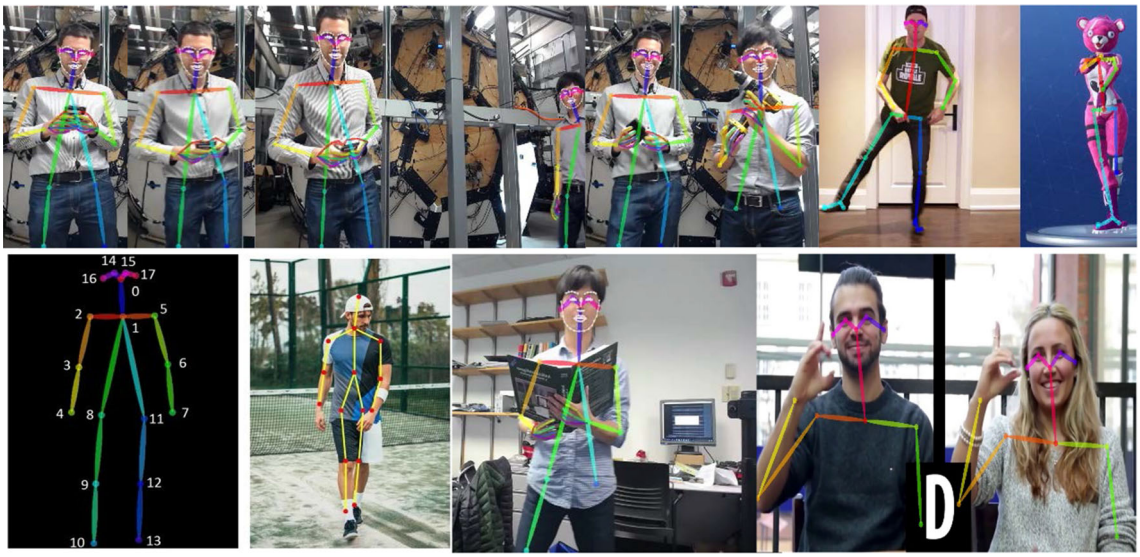

Fig. 3 Sample pose estimation results of open pose library

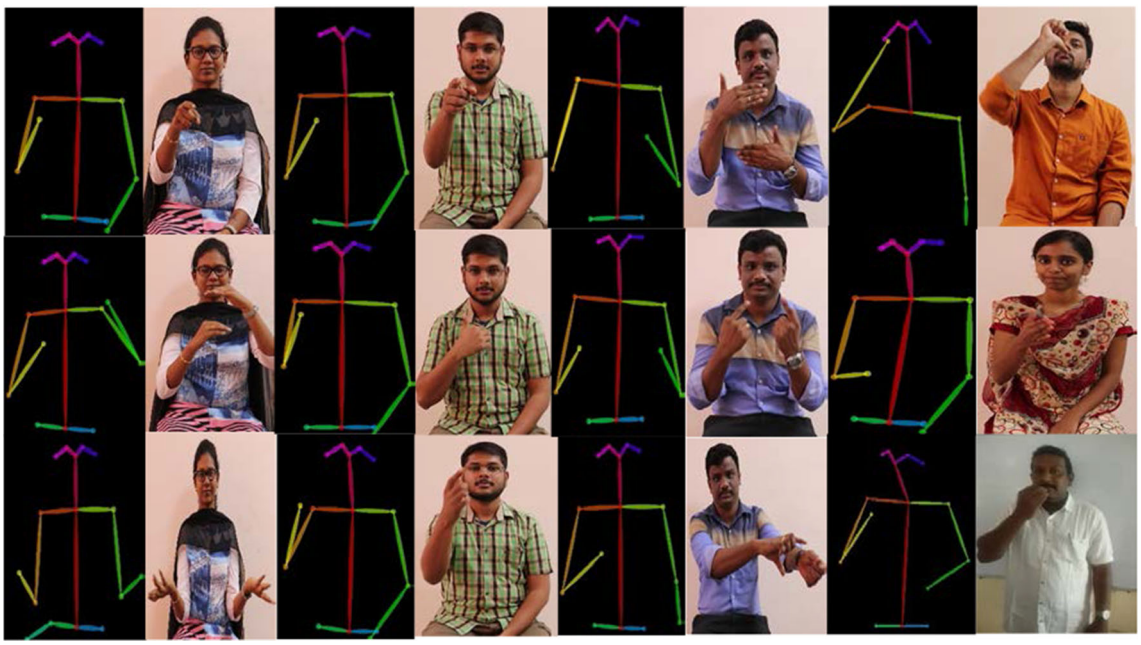

Fig. 4 Sample pose estimation results of open pose library for ISL-CSLTR dataset

persist of instability and mode collapse issues with the generated results, the evolutionary GAN model (Wang et al. 2019b) addresses these issues by employing different adversarial training methods and mutation operations in both generator network and discriminator network.

In Radford et al. (2015), the author investigated the CNN-based generative network for learning unsupervised feature representations, but this model needs to improve the learning of latent spaces to avoid mode collapse issues in generated results. We used an open pose library (Cao et al. 2019) in our model for extracting 18 different body joints form human poses. From which, we can predict the target poses and generate human images similar to real ones. The existing approaches (Pishchulin et al. 2013; Dantone et al. 2013) for estimating the human poses consider various parameters present in the input images. To find various gesture actions of human, local features identification and extraction techniques are used. Many of the researchers are proposed different techniques such as graphical models (Pishchulin et al. 2013), non-tree models (Dantone et al.

2013). In many cases, CNN (Pfister et al. 2015) found a highly very useful approach in pose estimation. Detecting multiple people poses in a single image creates higher complexity to the existing model. The open pose model (Yan et al. 2017; Wang et al. 2019a) gives a solution to this problem using part affinity fields. Table 1 describes the detailed information of existing generative frameworks implemented for image and video generation tasks.

3 The proposed system

3.1 GAN

The first GAN model was introduced by the author (Goodfellow et al. 2014) in the year 2014. The basic architecture of this model comprises the generator (G) and discriminator (D) networks. The generator network creates fake images similar to input images by manipulating the data distribution. The discriminator classifies the fake and real images using probability functions. The GAN architecture has been modeled like playing two players-based minimax games. Each network aims to increase its score by decreasing other network scores. Finally, it results in the production of high-quality images.

$$\min_{\text{Gen}} \max_{\text{Dis}} V(\text{Dis},\text{Gen}) = \mathbb{E}_x \sim p_{\text{data}}[\log .\text{Dis}(x)] + \mathbb{E}_{z \sim p_{z^{(z)}}} [\log(1 - \text{Dis}(\text{Gen}(z)))]. \quad (1)$$

In Eq. (1), p_{data} represents the real images and p_z denotes the noise vector values. We use the basic GAN network models in our work for generating videos. The generator and discriminator networks are fine-tuned to produce photo-realistic high-quality videos.

The Fig. 1 explains the basic components and its functions of GAN network. The generator network creates new samples using random noise vector values and the discriminator network classifies the plausible and implausible results.

3.2 Conditional GAN

The conditional GAN models (Mirza and Osindero 2014) have been achieved tremendous success in image or video generation. These models are evidenced as powerful ones in many GAN variant models (Isola et al. 2017; Li et al. 2019; Odena et al. 2017) to produce high-quality images. It applies conditioning on the class labels, assists the generator network to produce sharpened results by considering angles and orientation. Equation (2) describes the cGAN model.

$$\min_{Gen} \max_{Dis} V(D,G) = \mathbb{E}_{x,y} \sim p_{data}[\log \text{Dis}(x|y)] + \mathbb{E}_{z \sim p_{z^{(z)}}} [\log (1 - \text{Dis}(Gen(z|y)))] \quad (2)$$

In Fig. 2, the CGAN model uses class labels and random noise to generate the handwritten digits samples which are classified using discriminator as real or fake one. We have incorporated the conditional GAN techniques with our proposed model development for applying condition on sign glosses (class labels) to generate sharpened images with plausible sign gesticulations in the generated results.

3.3 OpenPose

The OpenCV-based OpenPose library (Cao et al. 2019) is mainly developed for human pose estimation in different environments like playground, meeting, dancing, street walking and interactions with others. The OpenPose techniques are developed by Carnegie Mellon University (CMU) researchers for developing the applications to track human actions, movements, behavior, and interactions in real-life environments. It detects the human body parts like the head, hands, limbs, and footpoints. It helps to identify the activity and pose orientation of a human in an image or videos by plotting color lines over the human images. This can be extensively used to track human activity in public or highly secured environments. This model learns the two-dimensional poses estimation from human body parts by

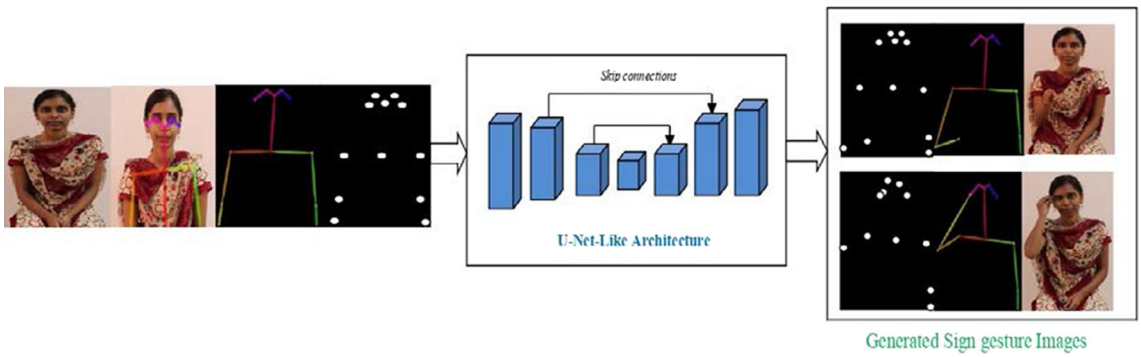

Fig. 5 Human-based Sign gesture realistic image Generation using key points

adopting Part Affinity Fields approaches described in Eq. (3). This technique follows bottom-up approaches that tend to produce improved results over earlier methods. It also detects key points in vehicle images and predicts the poses of hidden components in the human body. The overall pipeline of the open pose model comprises various folds. In the first fold, the input RGB color image gets processed for producing estimation of key points in 2D anatomical positions. For this estimation, the first ten layers of the popular CNN model known as VGG-19 are used. In the second fold, using part affinity fields and confidence maps the relationship association of body parts was

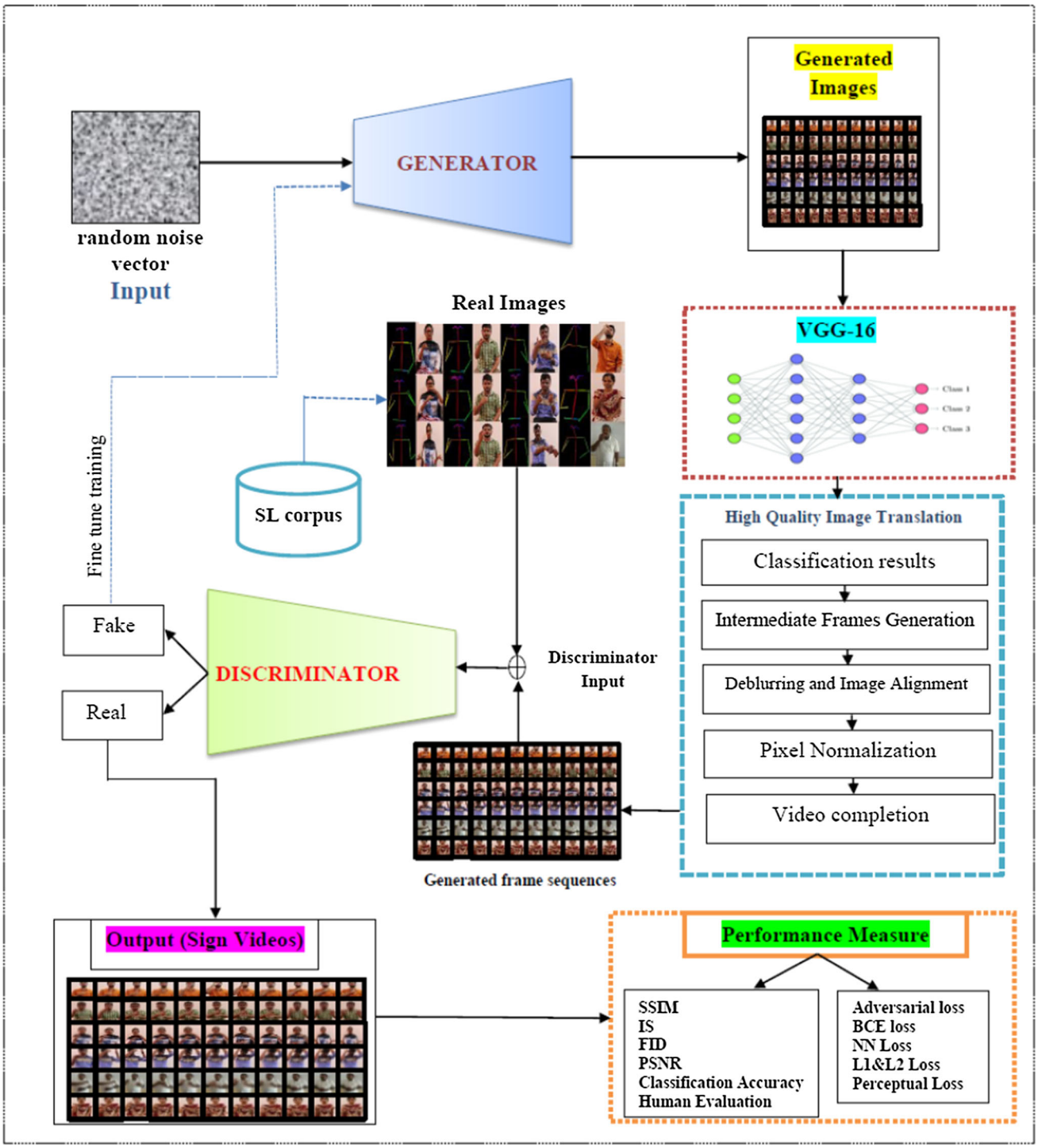

Fig. 6 The proposed dynamic GAN network architecture

identified. With the help of greedy inference techniques, the final poses are extracted (Fig. 3). We used double blending command to extract the skeletal poses alone from the sign videos for further processing.

$$\boldsymbol{L}_v^* = \frac{1}{n_v(\mathrm{Pt})} \sum_u \boldsymbol{L}_{v,u}^*(\mathrm{Pt}). \tag{3}$$

In Eq. 3, $n_v(\boldsymbol{P})$ denotes at point Pt the values of vector for u number of humans.

We use the open pose library in our work for extracting skeletal poses and key point information from sign gesture images. These extracted poses and key point helps to detect the movements of signs from one form to another in continuous sign cases. From which, we can modify the key points to create multiple views of sign gestures using the methods discussed in Zhao et al. (2018). The computation of affine transformation helps to identify the target pose feature maps. During the forward pass, the set of local affine transformations takes place to achieve the desired global pose-based deformation results. The coarse image generation using variants of Bayesian approaches provides a good approximation of conditional log-likelihood estimation and produces low-resolution images.

The fine image generation approach uses generative networks to improve the quality of images. We use affine transformation, coarse image generation, and fine image generation techniques to produce high-quality images by referring to the skeletal poses and ground truth (Fig. 4).

The improvements in open pose versions highly focus the possible failure cases like detecting poses in overlapping parts, presence of statues, various objects, and animals, misclassification of humans in highly crowded environments. These models train the machine to understand the interactions of humans in environments and estimates accurate poses. The model has been evaluated using three datasets (i) MPII human multi-person dataset, (2) COCO key point challenge dataset (3) foot dataset, MPII human multi-person dataset consists of a total of 5602 images categorized as 3844 training images and 1758 testing images. The open pose library efficiently detects 18 human body joints as key points in such multi-human-based images or videos. The COCO dataset results showcase the improvements of the open pose models by estimating 18 features. The foot dataset has experimented with open pose models for handling the failure cases which perhaps due to the variations in human images, occlusions, and hidden cues.

3.4 Pose to human-based sign gesture realistic image generation

This section discusses the translation process of skeletal key point information into human-based realistic sign gesture image generation procedure. We employ the PG^2 framework (Ma et al. 2017) and image inversion approaches stated in Eq. (4) to generate the desired target poses. We concatenate the source gesture image with skeletal pose information to provide input to the U-Net-like framework and optimized for producing the target gesture images.

$$z^* = \arg\min_z(\mathrm{dist}(G(z), x)). \tag{4}$$

Figure 5 explains the sign gesture image generation process using key points. The U-Net-like model used skip connections to understand the source and target representations and produces the high-quality results.

3.5 The proposed system

The generator network uses random noise vector values which are conditioned on sign glosses-based class labels to generate sign images. The sign gesture images and skeletal key point information are concatenated and passed as an input to the generator network. The generated results are classified according to class of sign glosses group using the VGG-16 framework. Further, we apply intermediate frame generation techniques to create intermediary frames between sign gestures. The creation of intermediary frames correlates the sequence of actions between signs to explore the real actions and changes. The noise present in the images is cleaned using deblurring approaches. The pixel normalization techniques and video completion techniques are used for smoothening the final results. We employ video completion techniques proposed in Cai et al. (2018) for generating intermediate frames between the sign images to synthesis the sequences of sign gestures as real one.

The perceptual loss and contextual L1 losses are combined to predict the intermediary frames between two sign gestures. Equation (5) describes the computations.

$$\hat{z} = \arg\min_z \left\{ \mathrm{Loss}_{\mathrm{context}}(z|I) + \alpha \times \mathrm{Loss}_{\mathrm{percept}}(z) \right\}. \tag{5}$$

This optimization strategy produces fine grained results for effective video generation. These results are fed into the discriminator network to analyze the realism of generated results. The discriminator network classifies the real and fake samples. In case of fake, it iterates the model training to improve the learning performance of the model. Figure 6 explores the detailed architecture of the proposed dynamic GAN network.

The proposed system functions are elaborated detail in algorithm (1) named as high-quality video generation using dynamic GAN model, in which random noise vectors (Z_i) are given as input to the generator network G. The training process will undergo several times based on input length and the final results fed into the discriminator network. The discriminator network (D) classifies the real and fake samples. Based on the training process, the quality of generator is improved to generate high-quality images similar to real images.

Algorithm 1: High Quality video generation using Dynamic GAN model

Begin

Input: Random noise vector (z_M) and RGB color Input images (I_K) Dataset

Output: Generation of photo realistic High Quality Videos

Procedure

1. Let Random vector input noise variable be z_i and Input Images I = $\{I_1,I_2,I_3,....I_N\}$where $I_1,I_2,I_3,....I_N$ denotes sequence of input images of count N
2. Initialize the buffer_size, batch_size, height and width of images
3. Load the training dataset to the generator network (G)
4. Apply resize, random cropping and normalization of pixel values
5. Feed the processed input (256 x 256 x 3) into GANgenerator network
6. **for** training the networks iteratively **do**

for t steps **do**

- apply minibatch of m noise vectors $\{z^{(1)}, z^{(2)},\ldots z^{(m)}\}$ using $p_{g^{(z)}}$
- applyminibatch of n image samples $\{I^{(1)}, I^{(2)},\ldots I^{(n)}\}$ using $p_{data^{(x)}}$
- update the discriminator (Dis) network by applying ascending SGD[6]

$$\nabla_{\theta_d}\frac{1}{m}\sum_{k=1}^{m}\left[\log \mathrm{Dis}(I_{input}^{(k)})\right] + \left[\log\left(1-\mathrm{Dis}\left(Gen\left(z_{\mathrm{noise}}^{(k)}\right)\right)\right)\right] \quad (6)$$

Calculate discriminator loss $\mathrm{Dis}_{\mathrm{Loss}}$

end for

- applyminibatch of m noise vectors $\{RN^{(1)}, RN^{(2)},\ldots RN^{(m)}\}$ using $p_{g^{(z)}}$
- update the generator (Gen) network by applying descending SGD[7]

$$\nabla_{\theta_d}\frac{1}{m}\sum_{k=1}^{m}\log\left(1-\mathrm{Dis}\left(Gen\left(z_{\mathrm{noise}}^{(k)}\right)\right)\right) \quad (7)$$

Calculate generator loss $\mathrm{Gen}_{\mathrm{Loss}}$

end for

7. Classify the generated samples using VGG-19 model
8. Apply Intermediate frame generation techniques, deblurring techniques and video blurring methods.
9. go to step 6

output(video)

End

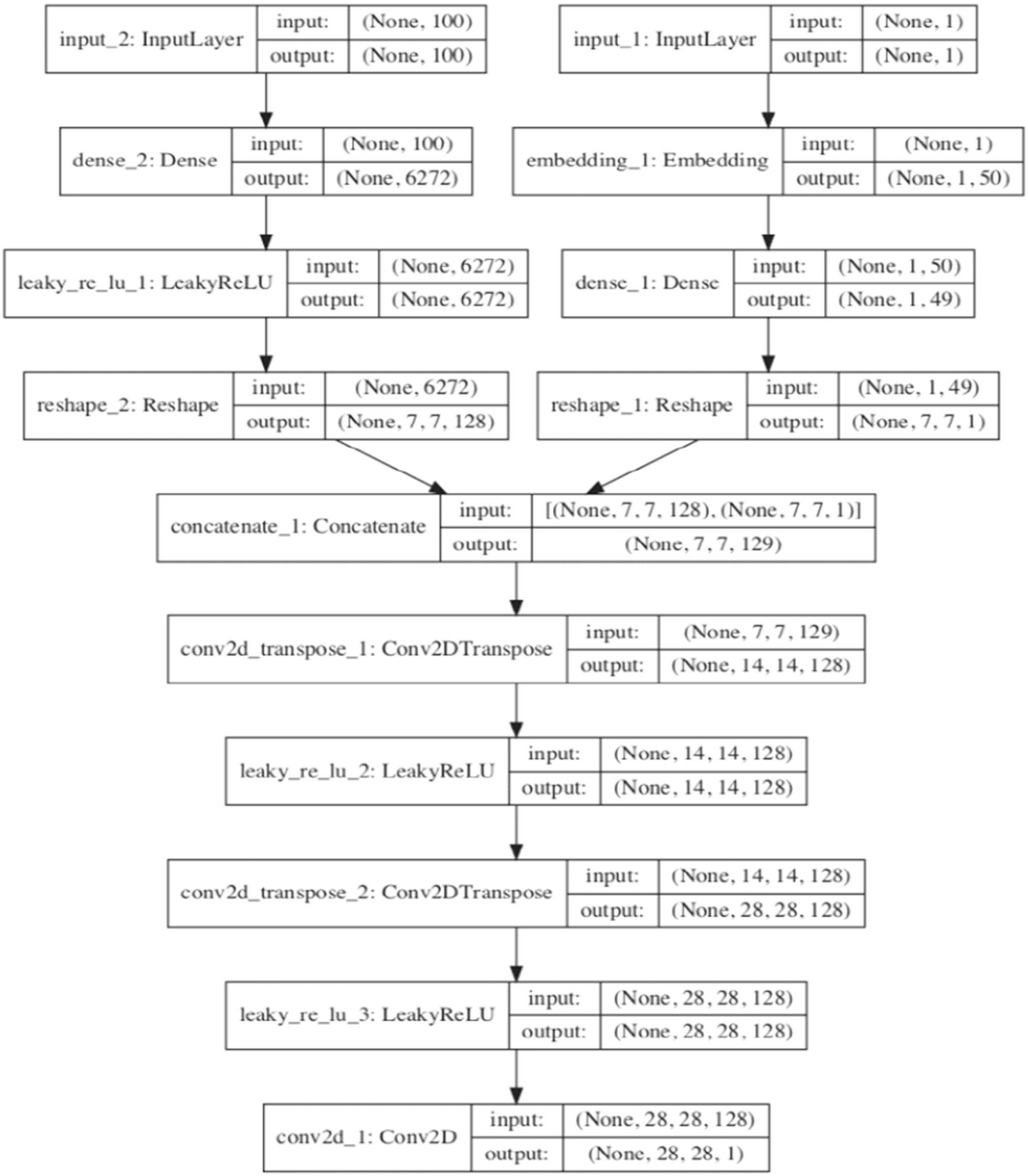

Fig. 7 Layer details of generator network

3.6 The generator network

The generator network generates sign gesture images using conditional GAN (Mirza and Osindero 2014) methods which are conditioned in class labels of sign glosses. The layer wise details are plotted in Fig. 7. Figure 8 explores the generation of sign gesture images from the skeletal pose information of the generator network. The generator network uses encoder–decoder-based approaches to generate the plausible results. The encoder unit uses

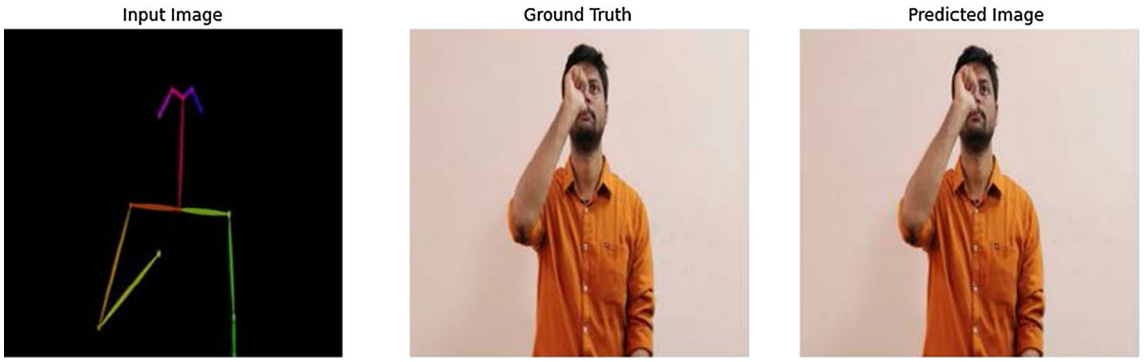

Fig. 8 The sample image generation from skeletal poses using generator network

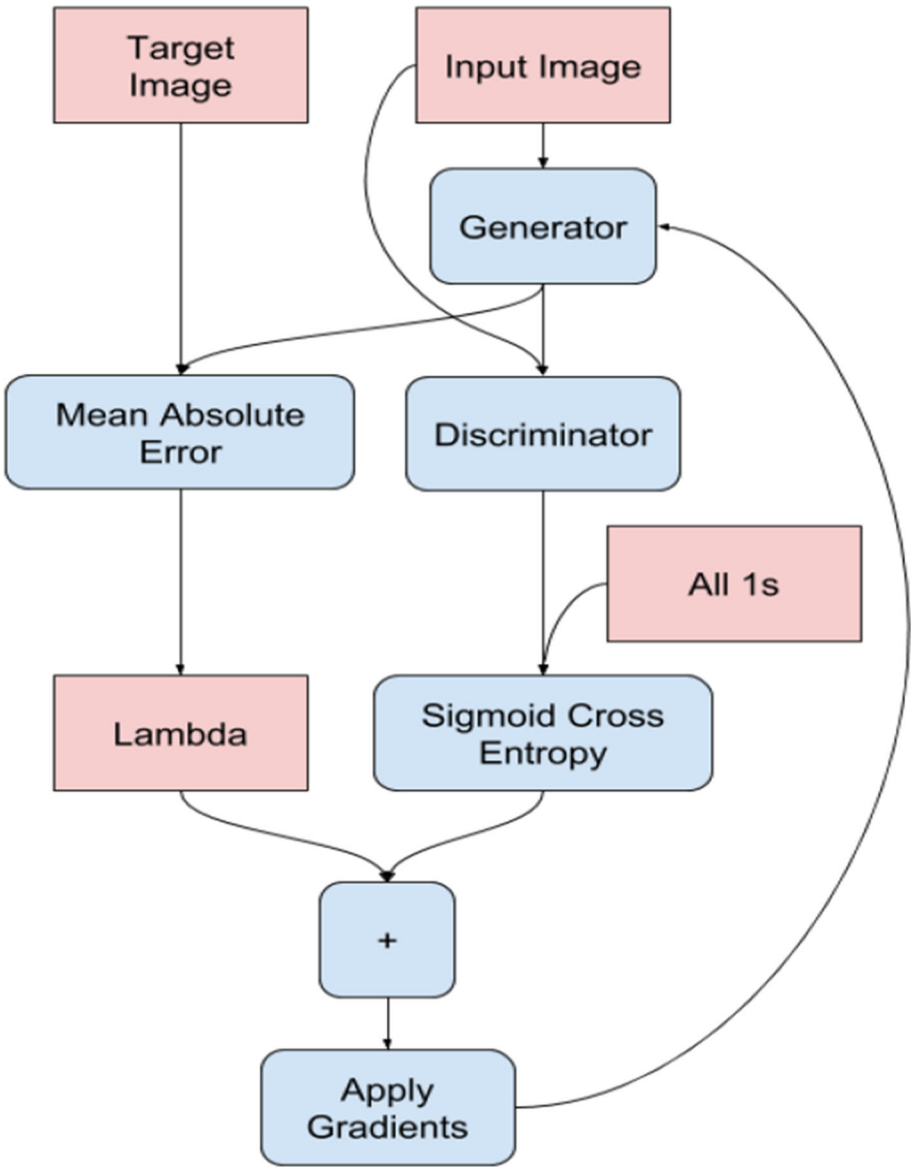

Fig. 9 The training procedure of generator network

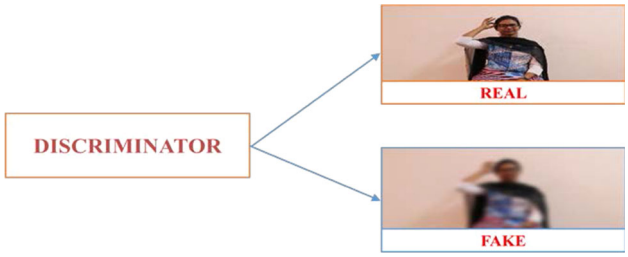

Fig. 10 The discriminator classification of real and fake samples

convolution, batch normalization and Leaky ReLU functions and decoder performs transposed convolution, batch normalization, dropout and ReLU activation functions. The skipped connection in U-Net-like framework emphasizes the better translation performance.

The generator loss is a sigmoid cross-entropy loss of the generated images. We also include L1 loss which refers to the mean absolute error between the generated image and the target image. This allows the generated image to become structurally similar to the target image. The generator network training procedures are illustrated in Fig. 9.

3.7 The discriminator network

The discriminator network uses the PatchGAN classifier approaches. Each block in the discriminator utilize the convolution operation, batch normalization, and the activation function Leaky ReLU to provide the accurate classification results. The discriminator evaluates the realness of the generated samples as shown in Fig. 10. The layer details of discriminator network are shown in Fig. 11. Each 30 × 30 patch of the output classifies a 70 × 70 portion of the input image.

The discriminator network classifies the real and fake samples by comparing the two data distributions. Based on such classification results, the generator improves its performance to generate samples matches with the real samples. The discriminator loss function compares the real samples and the generated samples. We use sigmoid cross-entropy loss function to estimate the loss of generated samples compared with real data distribution. Figure 12 depicts the training procedure of the discriminator network.

3.8 Training details

The proposed dynamic GAN model is developed using python programming language and high power GPU devices. We trained the proposed dynamic GAN model using the Dell Precision 7820 Tower work station. It has two Intel Xeon Silver 4210 2.2. GHz processors and 10 cores. It used Nvidia Quadro as RTX4000 support for providing GPU performance. The proposed model uses batch normalization techniques and Adam optimizer with α = 2e-4 β = 0.5 and β = 0.999 for optimization purpose. Batch size 128, dropout value 0.01 and initial learning rate is set as 0.01. We set Leaky ReLu value 0.1and ReLu activation functions. The mini batch size is 100 and momentum value is 0.05. We tested our model for five different benchmark datasets—RWTH-PHOENIX-Weather 2014T dataset, ISL-CSLTR dataset, UCF101Action Recognition dataset, MNIST Handwritten Digits dataset and CIFAR-10 Dataset.

3.9 Loss functions

The generator network generates the new samples using the random noise vector, to analyze the generated image quality. We measure the loss in generated results using mean squared error (MSE) metric is defined in Eq. (8).

$$\mathcal{L}_{\mathrm{MSE}}(\mathrm{gt},\mathrm{gen}) = \ell_{\mathrm{MSE}}(G(\mathrm{Xgt}),\mathrm{gen}) = \|G(\mathrm{gt}) - \mathrm{gen}\|^2. \tag{8}$$

The sigmoid cross-entropy loss comprises sigmoid activation plus a cross-entropy loss. This loss functions are independent for each vector component (class), meaning that the loss computed for every CNN output vector component is not affected by other component values. That's why it is used for multi-label classification, were the insight of an element belonging to a certain class should not influence the decision for another class. It is called binary cross-entropy loss because it sets up a binary classification problems. The loss functions are stated using Eqs. 9 and 10 as follows.

$$\mathrm{CE} = -t_1 \log(f(s_1)) - (1 - t_1) \log\,(1 - f(s_1))\,, \tag{9}$$

$$f(s_i) = \frac{1}{1 + \mathrm{e}^{-s_i}}. \tag{10}$$

4 Experimental results and discussion

4.1 The RWTH-PHOENIX-Weather 2014T dataset

The RWTH-PHOENIX-Weather 2014T dataset (Koller et al. 2015) was collected from the phoenix Television channel for the years 2009 to 2011. The 386 editions based on weather forecast information have been prepared as a dataset. The dataset contains video clips, frames, and annotation details clearly in the corpus repository and available for free access. This dataset highly supports the development of a German Sign Language-based assistive system for speech loss and hearing loss people. It is a first created corpus for handling continuous sign language process at the sentence level. The videos are available with

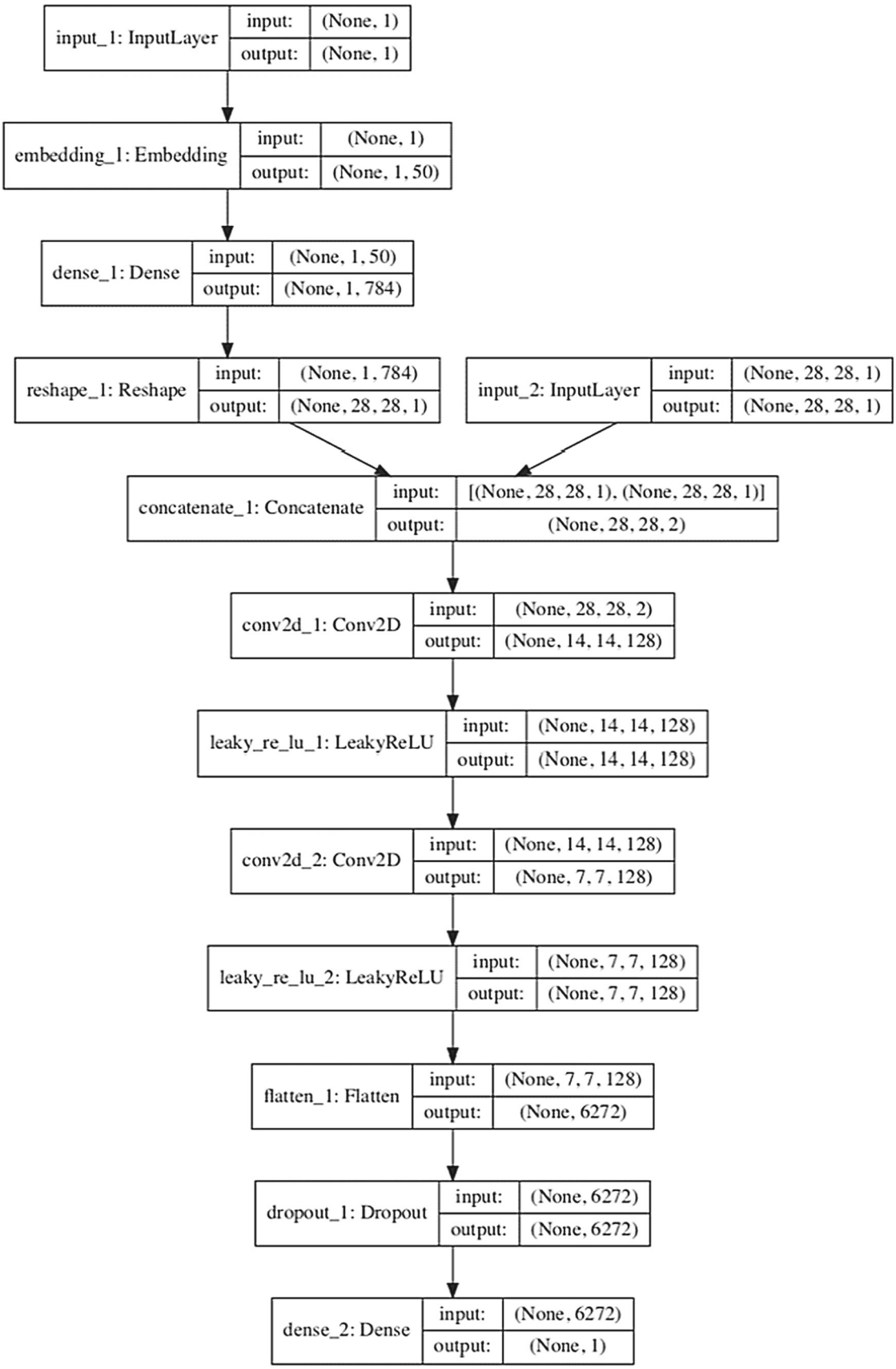

Fig. 11 Layer details of discriminator network

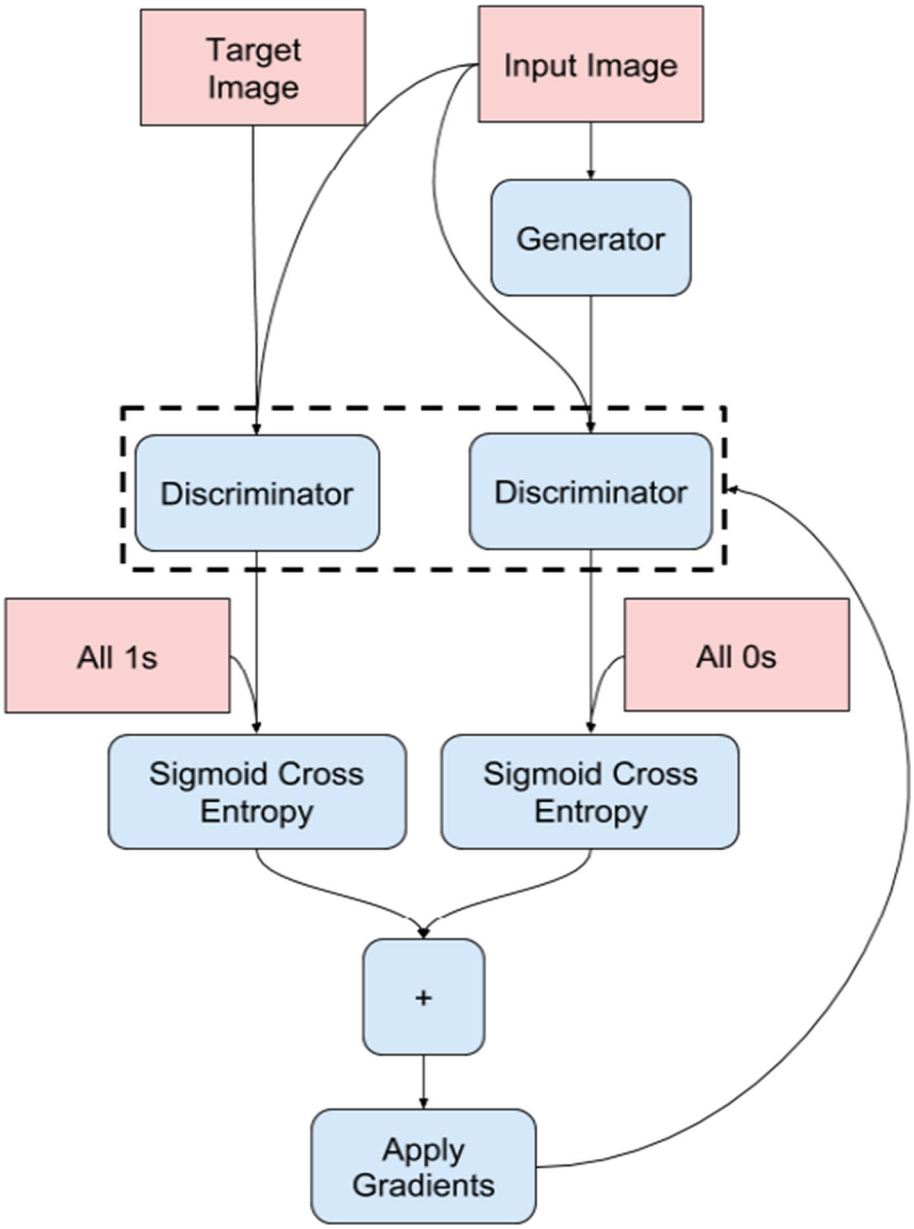

Fig. 12 The training procedure of discriminator network

210 × 260-pixel resolution and 25 frame rates per second. This dataset has been developed using 9 different signers and it consists of 7 k sentences. Figure 13 shows the generated videos frame sequences for the RWTH-PHOENIX-Weather 2014T dataset.

4.2 ISL-CSLTR: Indian sign language dataset for continuous sign language translation and recognition

We created the ISL-CSLTR dataset (Elakkiya and Natarajan 2021) for Indian sign language. This corpus has been created to support the deaf-mute community. This dataset is freely accessible and research works on sign languages can utilize it. This novel corpus consists of 700 videos collected from 7 different signers with different background environments and luminance conditions. This corpus was primarily developed for handling hundred English sentences that are used frequently in daily life. Figure 14 shows the generated videos frame sequences for the ISL-CSLTR dataset.

4.3 UCF101—action recognition data set

The UCF101—action recognition data set (Soomro et al. 1212) was collected from YouTube based on activities. This dataset has 101 different activity-based videos that project human daily life activities like applying makeup,

Figure. 13 The generated images of the RWTH-PHOENIX-Weather 2014 T dataset using the proposed dynamic GAN model

playing the game, swimming, brushing, vegetable cutting, and typing. Figure 15 shows the image samples for the UCF101—Action Recognition Data Set.

4.4 Structural similarity index measure (SSIM)

The structural similarity index measure (SSIM) metric (Wang et al. 2004) used for assessing the image quality. We use the SSIM metric (Eq. 11) for comparing the model performance with existing approaches. This metric assesses the structural information degradation of generated video frames and the results are tabulated in Table 2.

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}. \tag{11}$$

4.5 Inception score (IS)

The proposed dynamic GAN model performance has experimented with inception score metrics. The high score denotes the model performance over multiple domains and the generation capability of the generator. The computation of IS is performed using the following Eq. 12.

$$\mathrm{IS}(\mathcal{G}) = \exp\left(\mathbb{E}_{x\sim p_g}\mathcal{D}_{\mathrm{KL}}(p(y|x)\|p(y))\right). \tag{12}$$

Let x denotes the generated images of the generator network G, $p(y|x)$ denotes the class distribution of generated samples and the marginal probability function denoted as $p(y)$. The Inception score results are depicted in Table 3.

4.6 Peak signal-to-noise ratio (PSNR)

The generated video quality is evaluated using the PSNR quality metric (Eqs. 13 and 14). It compares the quality of generated results using ground truth images and provides the score. The higher PSNR value indicates improved quality in generated results. We compared our model performance with baseline models for the aforementioned three benchmark datasets and results are tabulated in Table 4 where gt denotes the ground truth samples and gen denotes the generated results.

$$\mathrm{PSNR}(\mathrm{gt},\ \mathrm{gen}) = 10\log_{10}\left(255^2/\mathrm{MSE}(\mathrm{gt},\ \mathrm{gen})\right), \tag{13}$$

$$\mathrm{MSE}(\mathrm{gt},\ \mathrm{gen}) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(\mathrm{gt}_{ij} - \mathrm{gen}_{ij}\right)^2. \tag{14}$$

4.7 Fréchet inception distance (FID)

The Fréchet inception distance (FID) metric (Heusel et al. 2017) evaluates the generated video quality by considering

Fig. 14 The generated images of the ISL-CSLTR dataset using the proposed dynamic GAN model

the visual features and temporal details. Equation (15) is used for computing the FID of generated videos. D denotes the CNN model used to extract the features in the video. We use the VGG-19 model for feature extraction. m_r, m_f denotes the mean values of features extracted from real and fake or generated samples, Σ_r, Σ_f denotes covariance matrix of features from real and generated samples. The lowest score of FID is always better for video quality estimation (Table 5).

$$d^2\big((m_r,\Sigma_r),(m_f,\Sigma_f)\big) = \|mr - mf\|_2^2 + \mathrm{Tr}\Big(\Sigma_r\Sigma_f - \Big(2\big(\Sigma_r\Sigma_f\big)^{1/2}\Big)\Big). \tag{15}$$

The results are shown in Fig. 16 compare the proposed dynamic GAN model performance with existing developments using different domain datasets. We compared the proposed model with Pixel GAN, MoCoGAN, PG^2 GAN Models. The results show the improved performance (Fig. 17).

In order to compare the video generation accuracy of the proposed model, we compared our model with existing developments and Fig. 18 shows the estimated results.

5 Conclusion and future work

In this paper, the proposed dynamic GAN network introduces a novel method for unsupervised learning-based sign video generation from skeletal poses. The proposed model incorporates the numerous techniques for generating photo-realistic high-quality video generation. The U-Net-Like generator model generates the target frames using skeletal pose key point information. The incorporation of VGG-19 model for classification and intermediate frame generation, de-blurring and image alignment, pixel normalization and video completion techniques significantly support the video generating process and yields high-quality photo-realistic sign gesture videos. The model has experimented with benchmark datasets (i) RWTH-PHOENIX-Weather

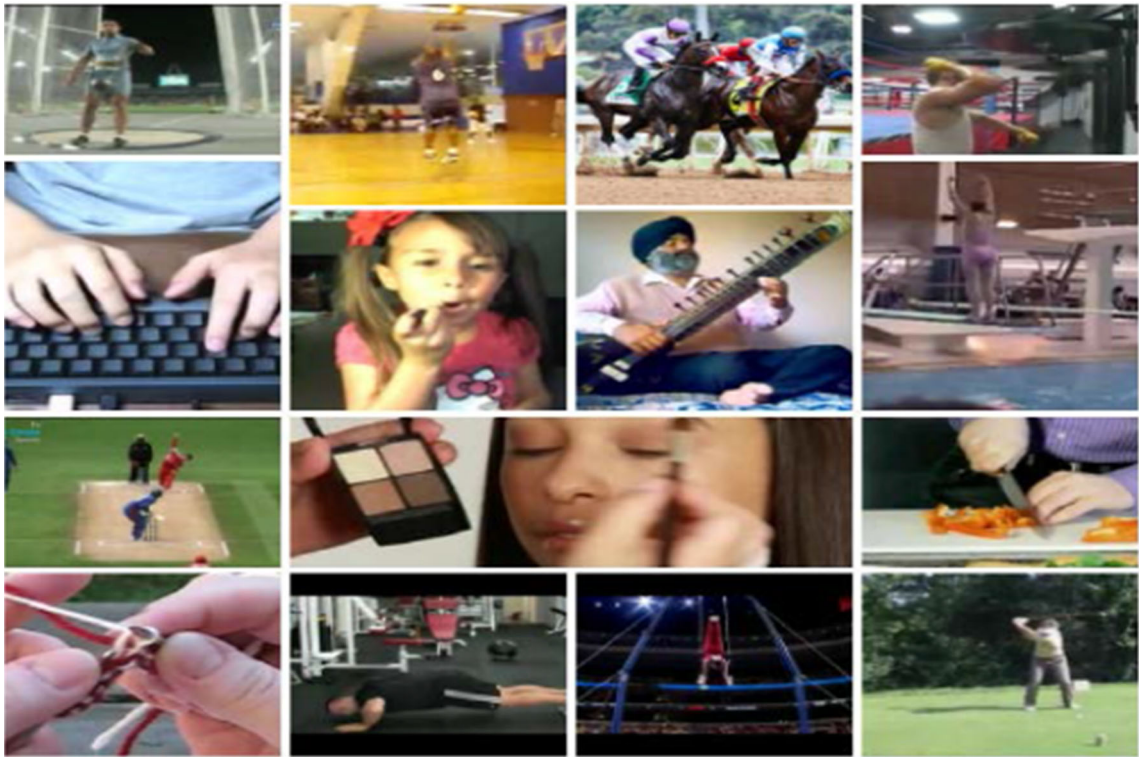

Fig. 15 The generated frames of the UCF101—action recognition data set using the proposed dynamic GAN model

Table 2 The comparison of structural similarity index measure (SSIM) metric with proposed dynamic GAN model

Framework	SSIM				
	RWTH-PHOENIX-weather 2014T dataset	ISL-CSLTR dataset	UCF101Action recognition dataset	MNIST handwritten digits dataset	CIFAR-10 dataset
MoCoGAN (Tulyakov et al. 2018)	0.702	0.802	0.856	0.841	0.872
PG^{2}GAN (Ma et al. 2017)	0.785	0.810	0.863	0.763	0.782
FutureGAN (Aigner and Körner 2018)	0.852	0.826	0.796	0.712	0.747
VGAN (2016)	0.891	0.901	0.892	0.791	0.772
Deformable GAN (Siarohin et al. 2018)	0.863	0.865	0.892	0.852	0.862
InfoGAN (Salimans et al. 2016)	0.836	0.796	0.783	0.723	0.743
Pixel GAN (Isola et al. 2017)	0.785	0.693	0.782	0.732	0.711
Ours (DynamicGAN)	0.901	0.937	0.925	0.911	0.935

2014T dataset (ii) ISL-CSLTR and (iii) UCF101-Action Recognition dataset. The proposed model addresses the existing challenges of generative networks such as instability training, high-quality image or video generation and generation of diverse images in an intelligent way. We also evaluated the model performance with various quality metrics shows the improved performance over multiple domain-based datasets. The proposed model will assist the development of applications for automating the translation of spoken text to sign videos to serve the deaf-mute

Table 3 The comparison of inception score (IS) metric with proposed dynamic GAN model

Framework	Inception score		
	RWTH-PHOENIX-Weather 2014T dataset	ISL-CSLTR dataset	UCF101Action recognition dataset
MoCoGAN (Tulyakov et al. 2018)	5.34 ± 0.05	5.74 ± 0.05	5.14 ± 0.05
PG^{2}GAN (Ma et al. 2017)	6.46 ± 0.11	6.45 ± 0.23	6.16 ± 0.26
FutureGAN (Aigner and Körner 2018)	7.07 ± 0.07	6.07 ± 0.07	8.01 ± 0.07
VGAN (2016)	7.56 ± 0.17	7.47 ± 0.87	7.28 ± 0.07
Deformable GAN (Siarohin et al. 2018)	7.86 ± 0.47	7.72 ± 0.81	7.63 ± 0.37
InfoGAN (Salimans et al. 2016)	7.96 ± 0.37	7.55 ± 0.85	7.64 ± 0.27
Pixel GAN (Isola et al. 2017)	8.19 ± 0.13	8.24 ± 0.19	8.16 ± 0.32
Ours (DynamicGAN)	8.69 ± 0.12	8.84 ± 0.09	8.66 ± 0.14

Table 4 The comparison of PSRN Score metric with proposed dynamic GAN model

Framework	PSNR value		
	RWTH-PHOENIX-Weather 2014T dataset	ISL-CSLTR dataset	UCF101Action recognition dataset
MoCoGAN (Tulyakov et al. 2018)	22.4201	21.3201	21.4277
PG^{2}GAN (Ma et al. 2017)	20.2306	22.2307	22.2356
FutureGAN (Aigner and Körner 2018)	22.6561	23.5571	23.6555
VGAN (2016)	23.2001	24.4061	23.2451
Deformable GAN (Siarohin et al. 2018)	22.6861	24.6741	22.6711
InfoGAN (Salimans et al. 2016)	24.1321	25.1361	24.1421
Pixel GAN (Isola et al. 2017)	25.3645	24.7645	25.3525
Ours (DynamicGAN)	28.2161	29.2201	28.7141

Table 5 The comparison of the Fréchet inception distance (FID) metric with the proposed dynamic GAN model, shows the generated video quality of our model compared with baseline models

Framework	Fréchet inception distance (FID)		
	RWTH-PHOENIX-Weather 2014T dataset	ISL-CSLTR dataset	UCF101Action recognition dataset
MoCoGAN (Tulyakov et al. 2018)	36.42	13.5	12.62
PG^{2}GAN (Ma et al. 2017)	33.23	14.3	13.23
FutureGAN (Aigner and Körner 2018)	36.65	35.9	39.25
VGAN (2016)	36.2	34.1	41.0
Deformable GAN (Siarohin et al. 2018)	28.68	37.36	39.6
InfoGAN (Salimans et al. 2016)	23.12	25.12	21.23
Pixel GAN (Isola et al. 2017)	17.36	19.23	13.32
Ours (DynamicGAN)	14.2	15.5	12.3

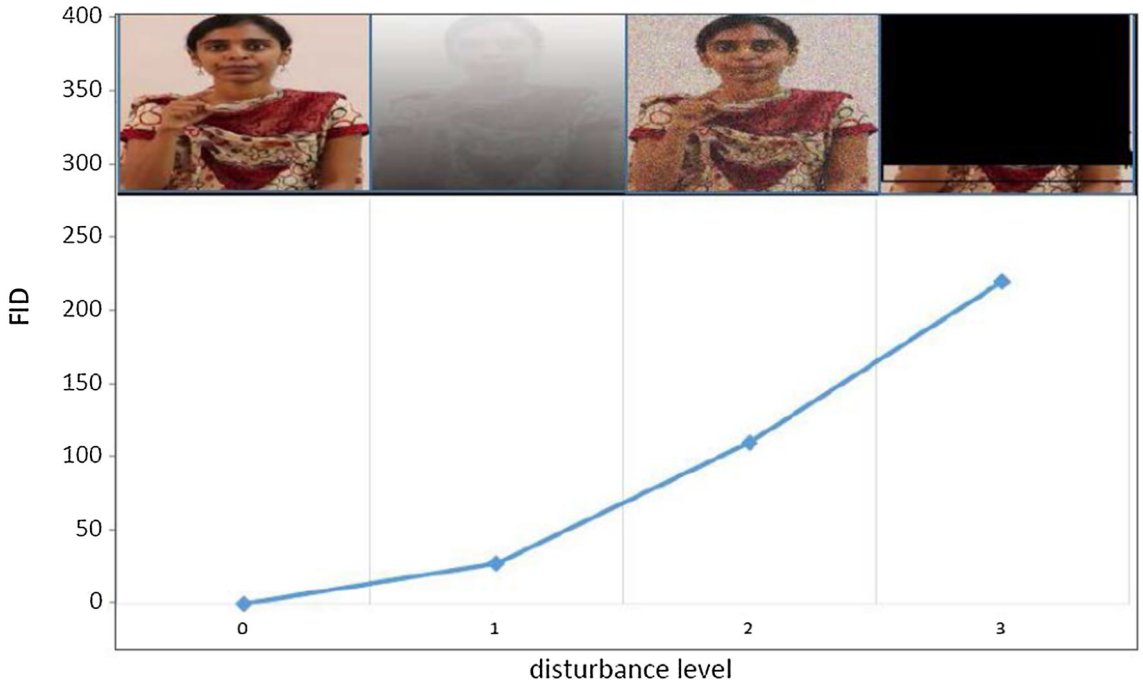

Fig. 16 The Fréchet inception distance (FID) metric evaluates the generated samples at different noise and blurred levels. The monotonic increase in this evaluation captures the different disturbance levels

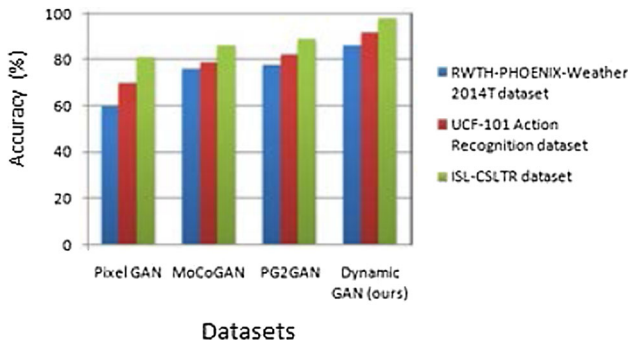

Fig. 17 Comparison of the proposed dynamic GAN model with diverse domain datasets

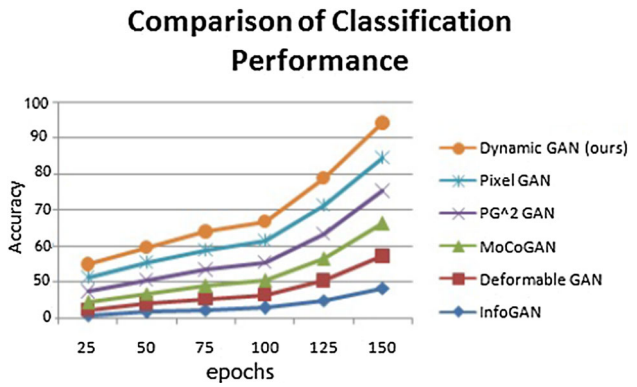

Fig. 18 Comparison of the classification performance of proposed dynamic GAN model

society. As a future work, we have planned to improve the limitations of the proposed framework for handling multiple domain datasets to develop numerous applications.

Acknowledgements The research was funded by the Science and Engineering Research Board (SERB), India under Start-up Research Grant (SRG)/2019–2021 (Grant No. SRG/2019/001338). We thank SASTRA Deemed University for providing infrastructural support to conduct the research. And also, we thank all the students for their contribution in collecting the sign videos and the successful completion of the ISL-CSLTR corpus. We would like to thank Navajeevan, Residential School for the Deaf, College of Spl. D.Ed & B.Ed, Vocational Centre, and Child Care & Learning Centre, Ayyalurimetta, Nandyal, Andhra Pradesh, India, for their support and contribution.

Funding The authors have not disclosed any funding.

Data availability Enquiries about data availability should be directed to the authors.

Declaration

Conflict of interest The authors have not disclosed any competing interests.

References

Aigner S, Körner M (2018) Futuregan: anticipating the future frames of video sequences using spatio-temporal 3d convolutions in progressively growing gans. http://arxiv.org/abs/1810.01325

Arjovsky M, Chintala S, Bottou L (2017) Wasserstein generative adversarial networks. In: International conference on machine learning. PMLR, pp 214–223

Beschizza R (2019) This person does not exist. Boing-Boing

Brock A, Donahue J, Simonyan K (2018) Large scale GAN training for high fidelity natural image synthesis. http://arxiv.org/abs/1809.11096

Cai H, Bai C, Tai YW, Tang CK (2018) Deep video generation, prediction and completion of human action sequences. In Proceedings of the European conference on computer vision (ECCV), pp 366–382

Cao Z, Hidalgo G, Simon T, Wei S-E, Sheikh Y (2019) OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans Pattern Anal Mach Intell 43(1):172–186

Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P (2016) Infogan: interpretable representation learning by information maximizing generative adversarial nets. http://arxiv.org/abs/1606.03657

Choi Y, Choi M, Kim M, Ha JW, Kim S, Choo J (2018) Stargan: unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8789–8797

Choi Y, Uh Y, Yoo J, Ha JW (2020) Stargan v2: diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8188–8197

Clark A, Donahue J, Simonyan K (2019) Adversarial video generation on complex datasets. http://arxiv.org/abs/1907.06571

Cui R, Cao Z, Pan W, Zhang C, Wang J (2019) Deep gesture video generation with learning on regions of interest. IEEE Trans Multimed 22(10):2551–2563

Dantone M, Gall J, Leistner C, Van Gool L (2013) Human pose estimation using body parts dependent joint regressors. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3041–3048

Denton E, Chintala S, Szlam A, Fergus R (2015) Deep generative image models using a Laplacian pyramid of adversarial networks. http://arxiv.org/abs/1506.05751

Dosovitskiy A, Springenberg JT, Brox T (2015) Learning to generate chairs with convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1538–1546

Efros AA, Leung TK (1999) Texture synthesis by non-parametric sampling. In: Proceedings of the seventh IEEE international conference on computer vision, vol 2. IEEE, pp 1033–1038

Elakkiya R (2021) Machine learning based sign language recognition: a review and its research frontier. J Ambient Intell Humaniz Comput 12(7):7205–7224

Elakkiya R, Natarajan B (2021) ISL-CSLTR: Indian sign language dataset for continuous sign language translation and recognition. Mendeley Data. https://doi.org/10.17632/kcmpdxky7p.1

Elakkiya R, Selvamani K (2017) Extricating manual and non-manual features for subunit level medical sign modelling in automatic sign language classification and recognition. J Med Syst 41(11):1–13

Elakkiya R, Selvamani K (2018) Enhanced dynamic programming approach for subunit modelling to handle segmentation and recognition ambiguities in sign language. J Parallel Distrib Comput 117:246–255

Elakkiya R, Selvamani K (2019) Subunit sign modeling framework for continuous sign language recognition. Comput Electr Eng 74:379–390

Elakkiya R, Sri Teja KS, Jegatha Deborah L, Bisogni C, Medaglia C (2021) Imaging based cervical cancer diagnostics using small object detection-generative adversarial networks. Multimed Tools Appl 1–17

Gao H, Xu H, Cai QZ, Wang R, Yu F, Darrell T (2019) Disentangling propagation and generation for video prediction. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9006–9015

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014). Generative adversarial nets. In: Advances in neural information processing systems, p 27

Gulrajani I, Kumar K, Ahmed F, Taiga AA, Visin F, Vazquez D, Courville A (2016) Pixelvae: a latent variable model for natural images. http://arxiv.org/abs/1611.05013

He, J., Lehrmann, A., Marino, J., Mori, G., & Sigal, L. (2018). Probabilistic video generation using holistic attribute control. In: Proceedings of the European conference on computer vision (ECCV) (pp. 452–467).

Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) Gans trained by a two time-scale update rule converge to a local Nash equilibrium. Adv Neural Inf Process Syst 30

Huang X, Liu M-Y, Belongie S, Kautz J (2018) Multimodal unsupervised image-to-image translation. In: Proceedings of the European conference on computer vision (ECCV), pp 172–189

Isola P, Zhu JY, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134

Karras T, Aila T, Laine S, Lehtinen J (2017) Progressive growing of gans for improved quality, stability, and variation. http://arxiv.org/abs/1710.10196

Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4401–4410

Kim T, Cha M, Kim H, Lee JK, Kim J (2017) Learning to discover cross-domain relations with generative adversarial networks. In: International conference on machine learning. PMLR, pp 1857–1865

Koller O, Forster J, Ney H (2015) Continuous sign language recognition: towards large vocabulary statistical recognition systems handling multiple signers. Comput vis Image Underst 141:108–125

Larsen ABL, Sønderby SK, Larochelle H, Winther O (2016) Autoencoding beyond pixels using a learned similarity metric. In: International conference on machine learning. PMLR, pp 1558–1566

Li J, Chen E, Ding Z, Zhu L, Lu K, Huang Z (2019) Cycle-consistent conditional adversarial transfer networks. In: Proceedings of the 27th ACM international conference on multimedia, pp 747–755

Liu M-Y, Huang X, Mallya A, Karras T, Aila T, Lehtinen J, Kautz J (2019) Few-shot unsupervised image-to-image translation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10551–10560

Liu MY, Tuzel O (2016) Coupled generative adversarial networks. Adv Neural Inf Process Syst 29:469–477

Ma L, Jia X, Sun Q, Schiele B, Tuytelaars T, Van Gool L (2017) Pose guided person image generation. http://arxiv.org/abs/1705.09368

Mao X, Li Q, Xie H, Lau RY, Wang Z, Paul Smolley S (2017) Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2794–2802

Mathieu M, Couprie C, Le Cun Y (2015) Deep multi-scale video prediction beyond mean square error. http://arxiv.org/abs/1511.05440

Mirza M, Osindero S (2014) Conditional generative adversarial nets. http://arxiv.org/abs/1411.1784

Nie D, Trullo R, Lian J, Petitjean C, Ruan S, Wang Q, Shen D (2017) Medical image synthesis with context-aware generative adversarial networks. In: International conference on medical image computing and computer-assisted intervention. Springer, Cham, pp 417–425

Odena A, Olah C, Shlens J (2017) Conditional image synthesis with auxiliary classifier gans. In: International conference on machine learning. PMLR, pp 2642–2651

Pandian AP (2021) Performance evaluation and comparison using deep learning techniques in sentiment analysis. J Soft Comput Paradig (JSCP) 3(02):123–134

Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros AA (2016) Context encoders: feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2536–2544

Pfister T, Charles J, Zisserman A (2015) Flowing convnets for human pose estimation in videos. In: Proceedings of the IEEE international conference on computer vision, pp 1913–1921

Pishchulin L, Andriluka M, Gehler P, Schiele B (2013) Poselet conditioned pictorial structures. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 588–595

Pu Y, Gan Z, Henao R, Yuan X, Li C, Stevens A, Carin L (2016) Variational autoencoder for deep learning of images, labels and captions. Adv Neural Inf Process Syst 29:2352–2360

Radford, Alec, Luke Metz, and SoumithChintala (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. http://arxiv.org/abs/1511.06434

Reed S, Akata Z, Yan X, Logeswaran L, Schiele B, Lee H (2016) Generative adversarial text to image synthesis. In: International conference on machine learning, PMLR, pp 1060–1069

Saito Y, Takamichi S, Saruwatari H (2017) Statistical parametric speech synthesis incorporating generative adversarial networks. IEEE/ACM Trans Audio, Speech Lang Process 26(1):84–96

Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X (2016) Improved techniques for training gans. Adv Neural Inf Process Syst 29:2234–2242

Shan Q, Jia J, Agarwala A (2008) High-quality motion deblurring from a single image. Acm Trans Graph (tog) 27(3):1–10

Shishir FS, Hossain T, Shah FM (2020) EsharaGAN: an approach to generate disentangle representation of sign language using InfoGAN. In: 2020 IEEE region 10 symposium (TENSYMP). IEEE, pp 1383–1386

Siarohin A, Sangineto E, Lathuiliere S, Sebe N (2018) Deformable gans for pose-based human image generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3408–3416

Smys S, Haoxiang W (2021) Naïve Bayes and entropy based analysis and classification of humans and chat bots. J ISMAC 3(01):40–49

Soomro K, Zamir AR, Shah M (2012) UCF101: a dataset of 101 human actions classes from videos in the wild. http://arxiv.org/abs/1212.0402

Stoll S, Hadfield S, Bowden R (2020) SignSynth: data-driven sign language video generation. In: European conference on computer vision. Springer, Cham, pp 353–370

Tulyakov S, Liu M-Y, Yang X, Kautz J (2018) Mocogan: decomposing motion and content for video generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1526–1535

Vondrick C, Pirsiavash H, Torralba A (2016) Generating videos with scene dynamics. http://arxiv.org/abs/1609.02612

Wang X, Gupta A (2016) Generative image modeling using style and structure adversarial networks. In: European conference on computer vision. Springer, Cham, pp 318–335

Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612

Wang TH, Cheng YC, Lin CH, Chen HT, Sun M (2019a) Point-to-point video generation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10491–10500

Wang C, Chang Xu, Yao X, Tao D (2019b) Evolutionary generative adversarial networks. IEEE Trans Evolut Comput 23(6):921–934

Wang Z, She Q, Ward TE (2021) Generative adversarial networks in computer vision: a survey and taxonomy. ACM Comput Surv (CSUR) 54(2):1–38

Wu H, Feng J, Tian X, Xu F, Liu Y, Wang X, Zhong S (2019) secgan: a cycle-consistent gan for securely-recoverable video transformation. In: Proceedings of the 2019 workshop on hot topics in video analytics and intelligent edges, pp 33–38

Xie Z, Baikejiang R, Li T, Zhang X, Gong K, Zhang M, Qi J (2020) Generative adversarial network based regularized image reconstruction for PET. Phys Med Biol 65(12):125016

Xu W, Keshmiri S, Wang G (2019) Adversarially approximated autoencoder for image generation and manipulation. IEEE Trans Multimed 21(9):2387–2396

Yan Y, Xu J, Ni B, Zhang W, Yang X (2017) Skeleton-aided articulated motion generation. In: Proceedings of the 25th ACM international conference on multimedia, pp 199–207

Yang Z, Chen W, Wang F, Xu B (2017) Improving neural machine translation with conditional sequence generative adversarial nets. http://arxiv.org/abs/1703.04887

Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, Metaxas DN (2017) Stackgan: text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 5907–5915

Zhang H, Goodfellow I, Metaxas D, Odena A (2019) Self-attention generative adversarial networks. In: International conference on machine learning. PMLR. CVA, pp 7354–7363

Zhao B, Wu X, Cheng Z-Q, Liu H, Jie Z, Feng J (2018) Multi-view image generation from a single-view. In: Proceedings of the 26th ACM international conference on multimedia, pp 383–391

Zhu J-Y, Zhang R, Pathak D, Darrell T, Efros AA, Wang O, Shechtman E (2017a) Toward multimodal image-to-image translation. http://arxiv.org/abs/1711.11586

Zhu JY, Park T, Isola P, Efros AA (2017b) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232