KSII Transactions on Internet and Information Systems

A Novel HAT-cGAN-DIFF Model for Generating a High-Quality Video from Facial Image

Journal:	*KSII Transactions on Internet and Information Systems*
Manuscript ID	TIIS-AB-2023-Dec-1119
Manuscript Type:	Artificial Intelligence & Big Data
Date Submitted by the Author:	03-Dec-2023
Complete List of Authors:	Bui, Hung; Industrial University of Ho Chi Minh City, Data Science, Faculty of Information Technology Ho, Duy; Industrial University of Ho Chi Minh City, Data Science Department, Faculty of Information Technology Vo, Huy; Industrial University of Ho Chi Minh City, Data Science Department, Faculty of Information Technology
Keywords of your Paper:	cGAN, Diffusion Model, High quality video creating, Hidden affine transform, HAT-cGAN-DIFF

KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS VOL. X, NO. X, December 20XX 134

A novel HAT-cGAN-DIFF model for generating a high-quality video from facial image

Bui Thanh Hung*, Ho Vo Hoang Duy, and Vo Quoc Huy
[1] Data Science Laboratory, Faculty of Information Technology, Industrial University of Ho Chi Minh city
Ho Chi Minh city, Vietnam
[e-mail: buithanhhung@iuh.edu.vn, h.hoangduy2002@gmail.com, huyvo8500@gmail.com]
*Corresponding author: Bui Thanh Hung

Abstract

The development of high-quality videos from facial images is a pivotal pursuit with multifaceted applications in our daily lives, spanning from the creation of portraits to the analysis of facial attitudes and expressions, as well as advancements in recognition technology, notably face shape recognition. Despite the numerous methods proposed in previous studies to tackle this challenge, these approaches contend with persistent limitations. These constraints include shortcomings like inadequate clarity in the generated videos and the presence of disruptive noise, which hinders the fluidity of facial movements, ultimately leading to a diminished sense of naturalness. In this study, we introduce a novel method designed to enhance the quality of videos generated from facial images. We propose a novel HAT-cGAN-DIFF model to create high-quality videos from images with two modules. The first module combines Conditional Generative Adversarial Neural Network (cGAN) with Hidden Affine Transform to create a quality video from images. The second module denoises the video generated from cGAN to ensure the generated video is clear and natural by using Diffusion model. We have conducted experiments and evaluated our method on two datasets: the color image dataset MUG datasets (The MUG Facial Expression Database) and the CK+ gray image dataset (CK-Mixed datasets). Our experimental results, when contrasted with previous research methods, reveal a notable enhancement in the quality of generated videos.

Keywords: cGAN, Diffusion model, High quality video creating, Hidden affine transform, HAT-cGAN-DIFF

A preliminary version of this paper appeared in IEEE ICC 2020, June 15-19, Dresden, Germany. This version includes a concrete analysis and supporting implementation results on MICAz sensor nodes. This research was supported by a research grant from the IT R&D program of MKE/IITA, the Korean government [2020-Y-001-04, Development of Next Generation Security Technology]. We express our thanks to Dr. Donald Lincoln who checked our manuscript.

http://doi.org/10.3837/tiis.2020.0X.00X **ISSN : 1976-7277**

1. Introduction

Recent strides in computer vision have ushered in a transformative era, fundamentally reshaping the landscape of creative art and user experience. Historically, the conversion of still images into high-quality videos posed a considerable challenge due to constraints in computer processing speed and performance. However, the landscape has evolved significantly, and with the progress in artificial intelligence and neural networks, this predicament has become more attainable than ever before.

The capacity to transform images into high-quality videos has garnered considerable interest and is experiencing robust growth, carrying with it numerous potential and practical applications. This challenge not only signifies a significant advancement in the realm of computer vision but also acts as a compelling catalyst for the ingenuity of scientists and researchers.

Transforming static images into high-quality videos holds significant importance across various domains. In the realm of portrait creation, the ability to generate video portraits from photographs of relatives or deceased individuals imbues profound spiritual significance. Within the media sector, this conversion process contributes to the creation of visually captivating advertisements and content, thereby elevating viewer engagement and facilitating effective message delivery. In entertainment, the production of high-quality videos amplifies audience experiences. And in education, the transformation of images into high-quality videos proves instrumental in rendering teaching more engaging and efficacious, particularly in instructing complex subjects such as science or history. Nonetheless, converting still images into videos presents formidable challenges for scientists. This necessitates researchers and developers possessing extensive expertise in computer vision and artificial intelligence. Working with intricate algorithms becomes imperative to ensure that the resulting videos are not only aesthetically pleasing but also authentic and impactful. The preservation of security and privacy during the image conversion process is a crucial consideration, especially in sensitive sectors like healthcare or security.

As technology continues to advance, the ongoing research and development of converting still images into high-quality videos persist. This intersection of art and computer science promises to foster more distinctive and enjoyable user experiences. Challenges pertaining to honesty, performance, and security will remain pivotal concerns for the research community in the foreseeable future.

In this study we propose an innovative and effective method for generating high-quality videos from facial images that integrate eye-lip movements and relevant facial parts into the generated video. Our contributions in this research include the following:

• We propose a novel HAT-cGAN-DIFF model based on Hidden Affine Transformation combined with image processing techniques and deep learning techniques to be able to synthesize dynamic video sequences from an input image.

• We use cGAN deep learning model to learn Affine transformations and to create facial images with desired features. Then, frames will be created and synthesized into videos with movements of lips, eyes and related parts of the face. By this way, the lips, eyes and related parts of the face will move in the smoothest and most realistic way.

• We use Diffusion Models to denoise the videos generated and synthesized from cGAN so that we can produce videos with scratch quality and realistic movements of facial parts. We initially add noise to the input video and then use U-Net 3D in the denoising part of the diffusion model to perform better denoising in the video, from there, high quality videos can be created.

• We conduct experiments on two datasets of color images and gray images, evaluate the results using various metrics to demonstrate the effectiveness of the proposed method.

• We compare with existing methods and evaluate the results based on the quality, realism and sharpness of the videos synthesized from the model.

Beyond the introductory section, the subsequent portions of the paper will unfold as follows: Section 2 delves into the presentation of related work, while Section 3 introduces our proposed model. Part 4 encompasses experiments and evaluation, and the conclusive insights derived from our research, along with directions for future development, are encapsulated in Part 5.

2. Related Work

In recent years, significant strides have been made in the realm of generating images and videos through deep learning models. Notably, the creation of videos featuring facial expressions stands out as a perennially captivating topic within this domain. Numerous research efforts have been dedicated to developing methods for generating images and videos using cGANs [1] that authentically capture facial movements.

This section delves into the research direction concerning the topic at hand. Initially, we explore recent advancements in the generation of static images depicting facial expressions. Subsequently, we delve into related works aimed at addressing the challenge of generating lifelike facial movements. Finally, in our pursuit of elevating the quality and sharpness of the generated images and videos, we also examine the application of the Diffusion model in the denoising process [2-5].

Image generation: Image generation has achieved great advances with various techniques. Among them, Generative Adversarial Network (GAN) [6] is commonly used in this problem. Image generation has been successfully applied in a number of problems such as: image synthesis [7], facial expression image creation [8], video to video translation [9], etc... Also in creating images, there have been many researches on improving image quality to create high-resolution images [10-14]. Some recent methods apply Conditional GANs [6] in generating facial expression images, based on an input image and generating a corresponding output image. For example, some studies have integrated the target expression into GANs as a one-hot vector that defines the target class encoding, thus generating faces based on discrete expression states [15-16]. Although this method is simple, it only produces discrete facial expressions, significantly reducing the variety of generated patterns. Hajar Emami et al. [17] proposed a model called SPA-GAN for image-to-image translation tasks. In SPA-GAN model, the attention mechanism is computed in the discriminator part and used to help the imager focus more on the most important regions between the source and target domains, as a result, the output images become more realistic. Yunjei Choi et al. [18] proposed an image-to-image translation model-StarGANv2 to learn the mapping between different image domains and at the same time ensure the diversity of the generated images. Their model is capable of generating images with many different styles across multiple domains. Philip Isola et al. [19] proposed a cGAN model for image-to-image translation. This model not only learns the mapping from the input image to the output image but also learns how to create a loss function to train this mapping. They demonstrated that their method is effective in imaging and many other tasks. From these studies, we proceed to discuss methods for video creation.

Video generation: Along with the previous success of GANs in image generation. There have been many studies on the problem of creating videos using GAN models such as: Text-to-Video (T2V) [20-24], Video-to-Video (V2V) [25-29], Image-to-Video (I2V) [30-34]. We go deeper into describing some models for creating videos from images such as:

R. Rombach et al. used the cINN model [35] and other methods such as Affine transform, nonlinear deep learning, and autoregressive models to synthesize videos from input images. However, the above method has some limitations such as limited capacity of the estimation model, and difficulty in handling large as well as in complex distributions. Haomiao Ni et al. proposed Late Flow Diffusion Models (LFDMs) [36] which is a class of generative models based on latent flow used to simulate complex data distribution such as images and videos. The proposed model can generate videos by transforming given images with flow sequences created in the hidden space based on layer conditions. The limitation of this model is that LFDM is limited to processing videos containing a single moving object, comparing to GAN models LFDM is much slower when sampling with 1000-step DDPM. Saito et al. proposed the TGANv2 model [37] to generate high-resolution time series in a memory-efficient manner. Yang Zhou et al. [38] introduced a method using deep learning to generate dynamic videos of speakers with smooth lip movements and natural portrait expressions. However, the above model still has some limitations that prevent it from generating high-resolution images of people talking. Although many significant achievements have been achieved, the quality of videos produced has not improved significantly. Therefore, we continue to learn about the Diffusion model so that the videos created are sharper and higher quality.

Diffusion Models: Diffusion models are a class of generative models based on the idea of gradually adding noise to an image until it is completely degraded. The reverse process is the imaging process. This process is modeled mathematically as a Markov process [39]. DDPM [40], Denoising probabilistic diffusion model, was the first successful diffusion model for high-quality imaging. Latent Diffusion Model [41] the underlying model of Stable Diffusion, this model allows diverse control via condition input, can be used to create custom images, is effectively in generating high resolution images, allows diverse control through condition input. Stable Diffusion is an improved diffusion model based on the Latent Diffusion Model. DALL-E2 [42]: OpenAI's diffusion model uses CLIP's spatial embedding. GLIDE [43]: Anthropic's diffusion model uses Bert [44] and CLIP [45] as text encoders. Imagen [46]: Google's diffusion model focuses on improving text encoders. Parti [47]: diffusion model uses transformer architecture instead of U-Net. V-DALL-E [48]: extends DALL-E2 for video, using a frame-by-frame simulation of the Markov diffusion process. Video Diffusion Model [49] uses a spatiotemporal self-attention mechanism to model dependencies between frames. DALL-E3 [50]: the latest version of DALL-E supports both photos and videos. G-Diffusion [51]: Google's next generation diffusion model for video.

Diffusion model methods offer several advantages, including the capability to craft high-quality, lifelike images at impressive resolutions. The image generation process is adaptable and controllable, allowing manipulation through condition inputs such as text, bounding boxes, and segmentation maps. The model's versatility is highlighted by its ease of extension and fine-tuning, achieved through processes like retraining or the addition of layers. Additionally, the model excels in visualizing novel or infrequently encountered concepts through an efficient retrieval system. There are many techniques to improve image quality such as cross-attention map optimization, hyper-parameter optimization. Diffusion models also have some disadvantages such as: Still having difficulty creating images with many objects accurately. It is not yet possible to image new concepts perfectly without additional training data. Photo quality is still not really consistent and may lack detail. The model is large in size, consuming a lot of computing resources and training time. In this research paper, we propose a diffusion model using U-Net 3D for denoising of the diffusion model architecture.

From previous studies, we propose a novel HAT-cGAN-DIFF model to synthesize high-quality videos from still images with eye-lip movements and relevant facial parts to be able to

create smooth, realistic and sharp facial movements based on image processing techniques and deep learning methods. We use the cGAN model combined with Hidden Affine Transformation to synthesize videos from facial images. Then, passing the video generated from the cGAN model through the Diffusion model, we will initially add noise to the video and then use the U-Net 3D architecture for denoising the diffusion model to remove noise and increase the quality of the video. The model will be presented in detail in the next section.

3. Methodology

3.1 Proposed HAT-cGAN-DIFF Model

A Starting with the input image, we initiate data preprocessing steps to realign and prepare the input. Subsequently, we synthesize videos from these facial images. Once the facial image video is synthesized, we introduce noise to the video and subsequently pass it through a noise reduction layer to enhance sharpness by denoising. This final step ensures the production of videos characterized by the utmost quality and sharpness in facial movements. We present the architecture of the proposed HAT-cGAN-DIFF model in **Fig. 1**. Each component of the proposed HAT-cGAN-DIFF model is presented as follows.

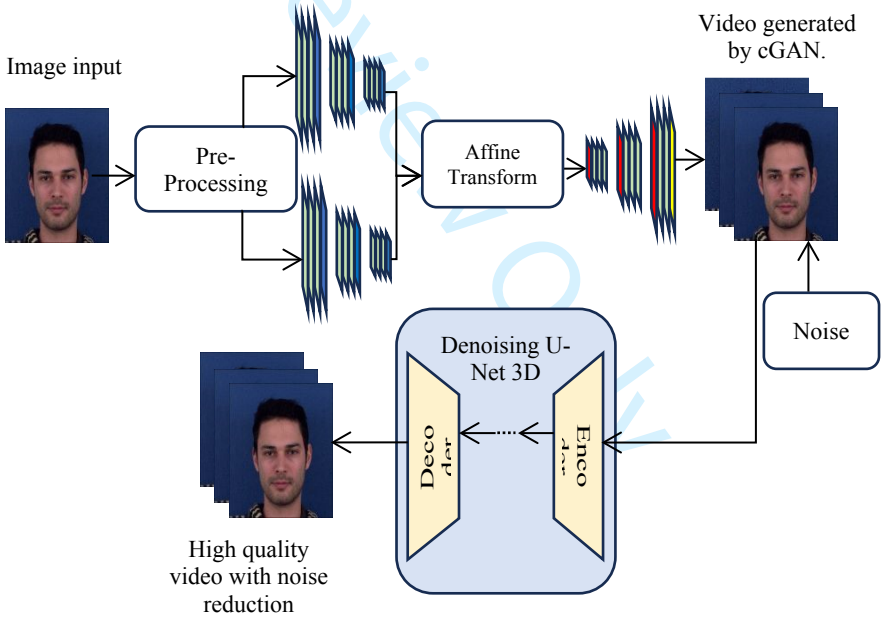

Fig. 1 The proposed HAT-cGAN-DIFF model

3.2 Conditional Generative Adversarial Network (cGAN)

cGAN stands for Conditional Generative Adversarial Network [1], which is an adversarial deep learning model introduced by Mirza and Osindero in the paper "Conditional Generative Adversarial Nets" in 2014. cGAN is an extension of GAN, complementing add a condition to both the generator and the differentiator. This condition can be any information the generator needs to generate the desired data samples. **Fig. 2** describes architecture of cGAN.

Generator in cGAN: is a neural network that can generate new data patterns. The generator

takes as input a condition and produces an output as a data sample. The structure of the Generator in cGAN is usually a neural network with a symmetric structure. This means that, if the network is divided into two halves, these two halves will mirror each other. This symmetrical structure helps the Generator create high-resolution and good quality data samples.

Discriminator in cGAN: is a neural network that can distinguish between real and fake data samples. The discriminator takes as input a data sample and returns a probability value indicating whether the data sample is real or fake.

The structure of the Discriminator in cGAN is usually a neural network with an asymmetric structure. This means, the two halves of the network do not mirror each other. This asymmetric structure helps Discriminator better distinguish between real and fake data samples.

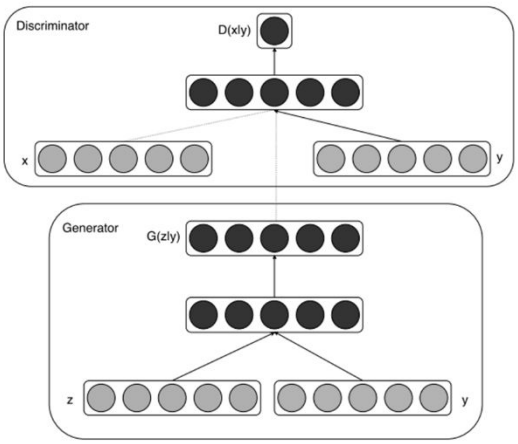

Fig. 2 cGAN architecture [1]

Formula formulation:

$$\min_{G}\max_{D} V(D,G) = \mathbb{E}_{a\sim p_{data}(a)}[\log D(a)] + \mathbb{E}_{c\sim p_c(c)}[\log(1 - D(G(c)))] \quad (1)$$

Generative Network:

The network receives as input a single image containing a human face. And try to create videos with smooth and realistic facial movements. In the generative network of cGAN we used U-Net [52] for the generator to be able to learn a high-level representation of the image. U-Net consists of two main components, Encoder and Decoder, which is a deep neural network architecture designed for the task of image segmentation. It can create clearly defined and precise image patches.

Learning these representations from U-Net helps cGAN's generative network create images of good quality, detail and high accuracy. U-Net is computationally efficient and easy to train.

In this research paper, we use two Encoder blocks for the U-Net model in the cGAN generating network including: Basic Encoder and Auxiliary Encoder. To perform the task of feature extraction from input images.

Basic Encoder is responsible for extracting basic, general features from the input face image to help the model learn preliminary features from the input image.

Auxiliary Encoder extracts more detailed features such as lips and eyes to create more realistic and sharper movements.

Then, we use Affine Transform to map features from image to video. Affine Transform is

a linear transformation on an image, performing mappings from face space to video space. This helps create videos with more realistic and smoother movements. The Decoder maps the features from the Encoder and creates a new image.

Discriminative Network:

The discriminative network is trained to be able to distinguish between images from the original data (real images) and images generated and synthesized from the generating network (fake images). The discriminative network uses a classification architecture to perform the task of distinguishing between real videos and images generated from the generative network. The task of the generating network is to generate images in the most realistic way so that the discriminating network cannot distinguish between real images and generated images.

Architecture of combining cGAN and Affine Transformation is presented in **Fig. 3**.

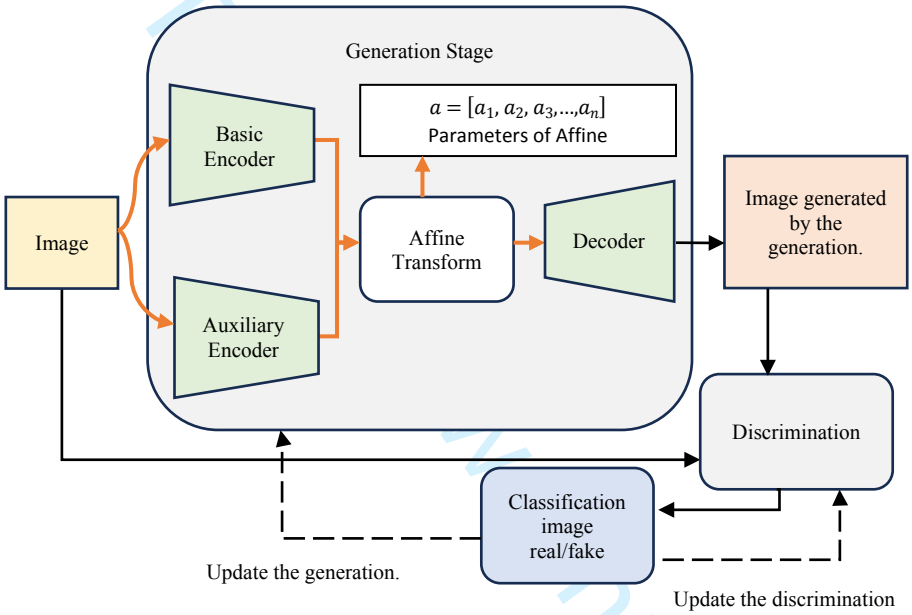

Fig. 3 Combining cGAN and Affine Transformation

3.3 Hidden Affine Transformation

In computer vision, an affine transform is used to perform linear transformations on an image. Affine transformations include a linear combination of scaling, rotation, translation, and shearing. Affine transformations can be represented by matrices and vectors. An affine transformation in 2D space (a,b) is represented as follows:

$$\begin{bmatrix} a' \\ b' \end{bmatrix} = \begin{bmatrix} a_0 & a_1 \\ b_0 & b_1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} a_2 \\ b_2 \end{bmatrix} \quad (2)$$

where (a,b) is the initial coordinates of the point on the original image, (a', b') is the coordinates of the point after performing the affine transformation, $(a_0, a_1, a_2, b_0, b_1, b_2)$ are the coefficients of the affine transformation and matrix 2x2 is the transformation matrix. After the image is converted the intensity level of the pixel is assigned using interpolation.

Hidden affine transformation is a method for applying affine transformations on deep learning data without explicitly defining affine parameters. Instead, the deep learning model learns to generate video frames from the original facial image using affine transformations based on the training data.

The way to implement implicit affine transform in a deep learning model usually involves using a neural network to learn and predict the affine parameters corresponding to each video frame. These affine parameters are calculated based on input information, such as facial images, and are used to generate new video frames.

In this research, we use the Hidden Affine transform to improve the process of transforming facial images into videos by applying hidden linear transforms to reproduce natural expressions and movements in the face.

3.4 Diffusion model

Diffusion models is a class of probabilistic models that allow the generation of high-quality video samples. They are based on the idea of using a diffusion process to gradually generate the final video from an initial random initialization vector.

Specifically, diffusion models start with an initial random vector f_0. The model then applies a sequence of upward diffusion steps f_0 to gradually turn it into the desired video.

This diffusion process gradually removes noise and increases detail in the video. It helps the generated frames become sharper and more realistic after each diffusion step.

Diffusion models are capable of generating high quality videos from an initial random vector. They are also very versatile, can be used for many different purposes such as video editing, restoring damaged videos, removing noise, compressing videos while maintaining the quality.

Diffusion model composes of two main components:

Diffusing process:

$$f_t = f_{t-1} + \epsilon_t \tag{3}$$

where f_t is the frame of video at time t, f_{t-1} is frame of the frame at time t-1, ϵ_t: is noise is added to the video frame at a time *t*

Reverse diffusion:

$$f_t = \sum_{j=0}^{t} \gamma_i \epsilon_i \tag{4}$$

where γ_i: is a small positive real number, often called as coefficient of diffusion. In this study Diffusion models are used as follows:

Video generated from cGAN also contains a lot of noise, making the video not realistic and clear. We recommend using a diffusion model to reduce noise, making the video sharp and high-resolution. The model consists of two parts: Diffusing Processing, which is responsible for adding noise to the input videos. And Reverse Diffusion will learn how to reduce noise to create sharper, higher quality video. **Fig. 4** describes architecture of Diffusion model.

We add Gaussian noise to the original video data (video generated from cGAN). Adding Gaussian noise to create latent vectors with a continuous distribution. Initially the data is discrete (pixels in the image). If we want to turn it into a continuous vector, we need to add noise. Noise Gaussian is the optimal choice because it produces a continuous distribution, this makes the training process easier.

We use U-net 3D architecture in denoising model. The basic U-net 3D architecture is the

same as U-net 2D for image processing, consisting of a downsampling block and an upsampling block with skip connections from the downsampling block. Instead of using 2D convolutional layers, the model uses 3D convolutional layers. These layers apply convolution over both space and time. During the downsampling process, the model gradually reduces the spatial size of the image to 1x1 but retains the time dimension. After each 3D convolutional layer block, the model inserts a temporal attention block. This block pays attention to the time dimension of the data. During the upsampling process, the model gradually restores the original spatial dimension of the video but retains the temporal dimension. This architecture allows the model to learn the spatial and temporal features of the video effectively. The model can be trained on both videos and images by enabling temporal attention to be turned off to process the image data. During the sample generation process, the model can predict each video frame based on the noise vector z and previously generated frames. After performing the above steps, the model will generate high quality videos from training data as input videos.

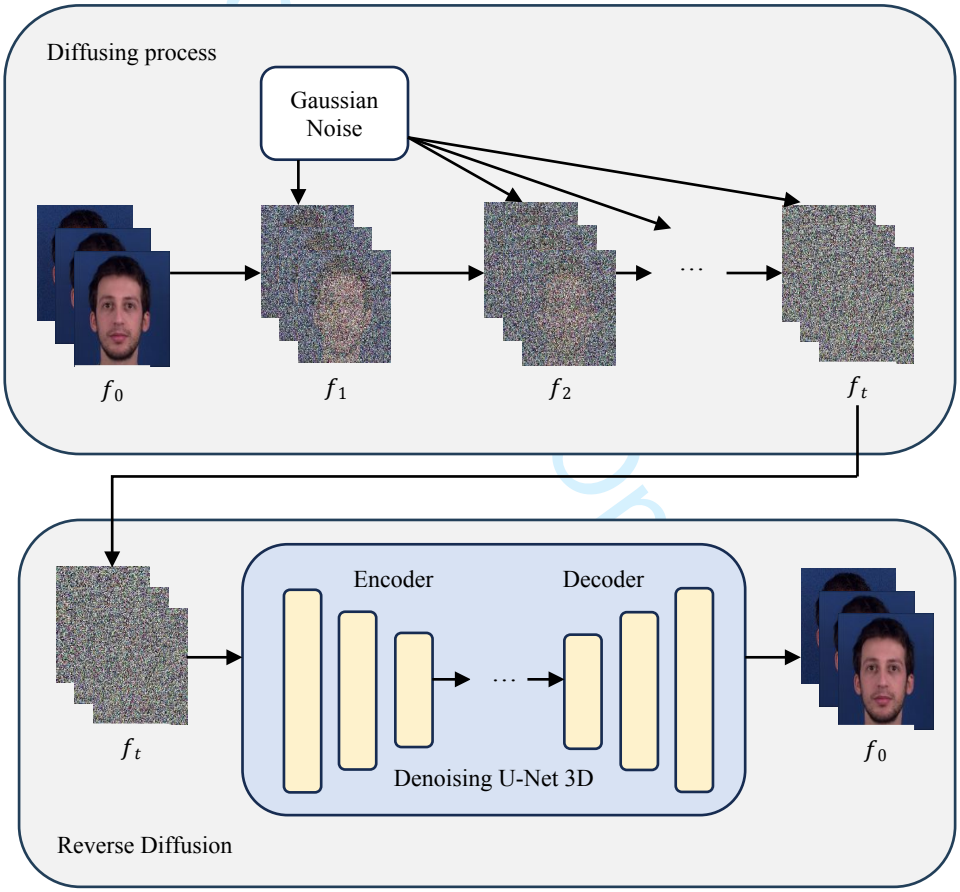

Fig. 4 Diffusion model

3.5 Loss Function

While the GANs model learns to map a random noise vector c to the output image b, G:c→b. In contrast, cGAN learns a mapping from an input image a and a random noise vector c to

generate b, G:{a,c}→b.

During training, Generator G's main goal is to produce output that is indistinguishable from "real" images through Discriminator D trained to detect as best it can "fake" images. " created by Generator. Below are some loss functions we use during training:

Adversarial loss:

Adversarial loss is the loss function used to train the discriminator. It calculates the error between the discriminator's predicted label (real or fake) and the actual label (real or fake). Adversarial loss minimizes the distance between the predicted probability distribution and the target probability distribution.

$$\mathcal{L}_{adv}(G,D) = E_{a,b}[\log D(a,b)] + E_{a,c}[\log (1 - D(a,G(a,c))] \quad (5)$$

The goal is to train the discriminator so that it can accurately distinguish between real data and fake data generated by the generator. This forces the generator to create increasingly realistic fakes to fool the discriminator. This competitive process helps both models become better.

Reconstruction loss:

Reconstruction loss is the reconstruction loss, which calculates the distance between the real image and the fake image generated by the generator. One of the popular choices for loss reconstruction is L1 loss [53]. L1 loss calculates the absolute value of the difference between the real image and the fake image. L1 loss helps the generator create fake images that are more similar to real images. Here we define Reconstruction loss according to L1 loss as follows:

$$\mathcal{L}_{rct}(G) = E_{a,b,c}[\|b - G(a,c)\|_1] \quad (6)$$

The goal is to minimize this loss, thereby helping the generator create fake images that are more similar to real images. Reconstruction loss measures the pixel-wise difference between real and fake images, forcing the generator to create more detailed and accurate fake images.

Full loss:

To generate the target image b, we construct the loss function L by linearly combining all previous losses:

$$\mathcal{L} = \mathcal{L}_{adv}(G,D) + \lambda_1 \mathcal{L}_{rct}(G) \quad (7)$$

Where λ_1 used to help stabilize the training and balance the optimization of our model. Finally, we define the optimization problem as follows:

$$G^* = arg\ min_G\ max_D\ \mathcal{L} \quad (8)$$

The procedure for generation phase of the proposed approach is given in pseudocode (Algorithm 1).

Algorithm 1: Steps for generating a video from an image

Proposed HAT-cGAN-DIFF Framework

1 : Input : Image I ∈ RWxH

2 : Output : a video

3 : Preprocess I: Resize image to W' x H'

Normalize pixel values

f= Image feature ∈ R
4: Hidden Affine Transformation a= Affine(f)
5 : for t = 1 to T do
for k = 1 to K_g do
Generator model using cGAN: g= cGAN(I, p)
Calculate Adversarial loss $\mathcal{L}_{adv}$ (G,D) in (5)
end for
for k = 1 to Kd do
Discriminator model using cGAN: d= cGAN(I)
Calculate Reconstruction loss $\mathcal{L}_{rct}$ (G) in (6)
end for
Calculate Full Loss L in (7)
Optimization on Full Loss L in (8)
6 : end for
7: Return generated video v = cGAN(I, a)
9 : Diffusion model from generated video v:
Diffusing Processing
Adding Gaussian noise
Reverse Diffusing
for t = 1 to T do
Denoising d = U-Net 3D(v)
10: Return high quality generated video v:

4. Experiments

4.1 Dataset

Fig. 5 Some images of the MUG dataset

MUG Datasets

In this study, we perform training and evaluation of the proposed HAT-cGAN-DIFF model on the Facial Expression Database-Multimedia Understanding Group (MUG) [54]. The dataset contains videos capturing natural facial expressions of 86 people (56% male, 44% female, mixed ethnicity). A total of 7 expressions were recorded: happy, sad, scared, disgusted, angry, surprised, neutral. We process and extract frames from the video and obtain 10000 images. **Fig. 5** shows some images in the dataset.

CK-Mixed datasets

We use the CK-Mixed (CK+) dataset [55] which is a grayscale image dataset to conduct experiments in the proposed method and compare the results with the color image dataset MUG. CK-Mixed is a popular dataset in facial emotion analysis, this dataset includes 593 videos with 6 different types of human facial emotions. All videos are in grayscale. After performing preprocessing steps and extracting image data from videos, our dataset will have 8000 images. Some images in the dataset are presented in **Fig. 6**.

Fig. 6 Some images of the CK-Mixed dataset

4.2 Hyperparameters

We used optimization by Adam [56] with learning rate 0.0002, β_1=0.5, β_2=0.999. Parameters in formula (7) are set 1 and 10. We applied instance normalization and set batch size is 1. Our encoders and decoders are designed as 8-layer neural networks with skip connection layers [57]. The Discriminator is a 3-layer convolutional neural network. For the tests on MUG and CK-Mixed we used 10 videos for the test set. All other videos are used for training. The data training process took approximately 50 hours for MUG and 30 hours for CK-Mixed on GPU Nvidia GeForce GTX 1650Ti (4GB of memory).

4.3 Evaluation metrics

To evaluate our HAT-cGAN-DIFF model in quantitatively, we use the following evaluation metrics: PSNR, SSIM, ACD, ACD-I, FID, IS.

Structural Similarity Index Measure (SSIM) [58] indicates the degree of similarity between the real image and the reconstructed image. Peak Signal-to-Noise Ratio (PSNR) [59] is used

to evaluate image or video quality. The higher the SSIM and PSNR scores, the better the image and video quality produced.

ACD (Average Content Distance) [60] used to evaluate content consistency in created videos. ACD-I (Average Content Distance- Identity) [61] is an extended version of the ACD evaluation criterion to measure the identity preservation of the input face in the generated video. Low ACD and ACD-I scores indicate similarity between faces in consecutive generated video frames and similarity between faces in the input image and generated video.

Frechet Inception Distance (FID) [62] is an index used to evaluate the video creation process. It measures both the image quality and temporal consistency of the generated video. Inception Score (IS) [63] is a common metric used to evaluate the quality of generated samples, and has a strong correlation with human assessment. A low FID score indicates high quality of the generated videos. Conversely, a high IS score shows that the model can produce diverse and highly identifiable images.

4.4 Result

Based on the method introduced above, we conduct experiments to create videos from a single input face image. The proposed HAT-cGAN-DIFF method is capable of generating videos with realistic and sharp movements. The results are evaluated based on criteria such as similarity to the original image and sharpness of the image created by the proposed model. We also conduct experiments on different methods on MUG dataset, we make comparisons with previous models such as VGAN [64], MoCoGAN [65], ImaGINator [66]. We made comparisons with Video GAN [67], GANimation [68], FGST [69] models on CK-Mixed dataset.

On both datasets, we compare with Resnet 6 blocks and Resnet 9 blocks in the generative network of the cGAN model.

Comparing with our proposed HAT-cGAN-DIFF method is to use U-Net in the cGAN generation network combined with the Diffusion denoising model with the methods presented above. **Fig. 7** presents the experimental results of creating color images with our model and with previous models. **Fig. 8** shows the results when performed on grayscale images.

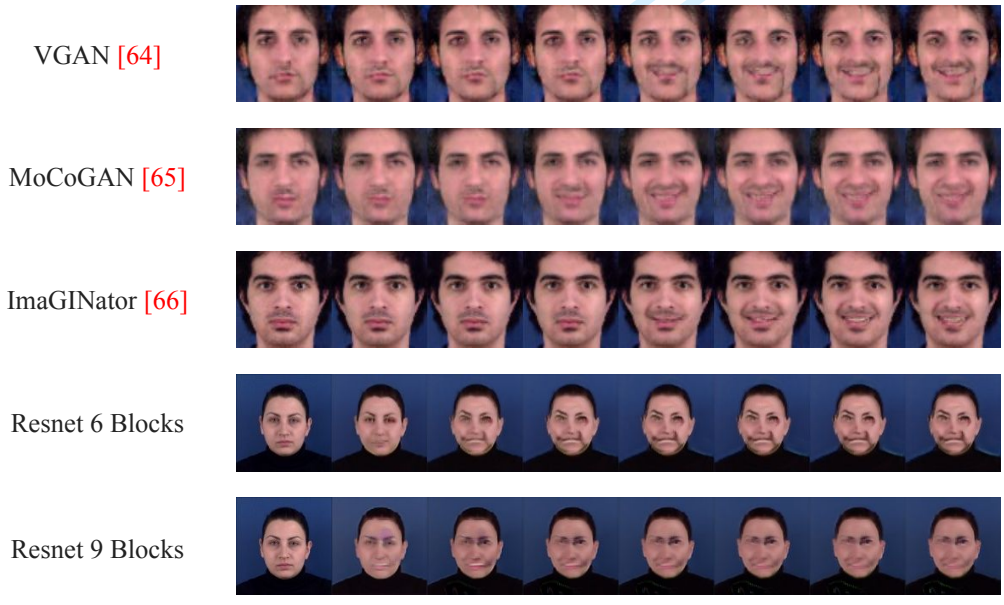

Fig. 7 Results of six models when creating color images

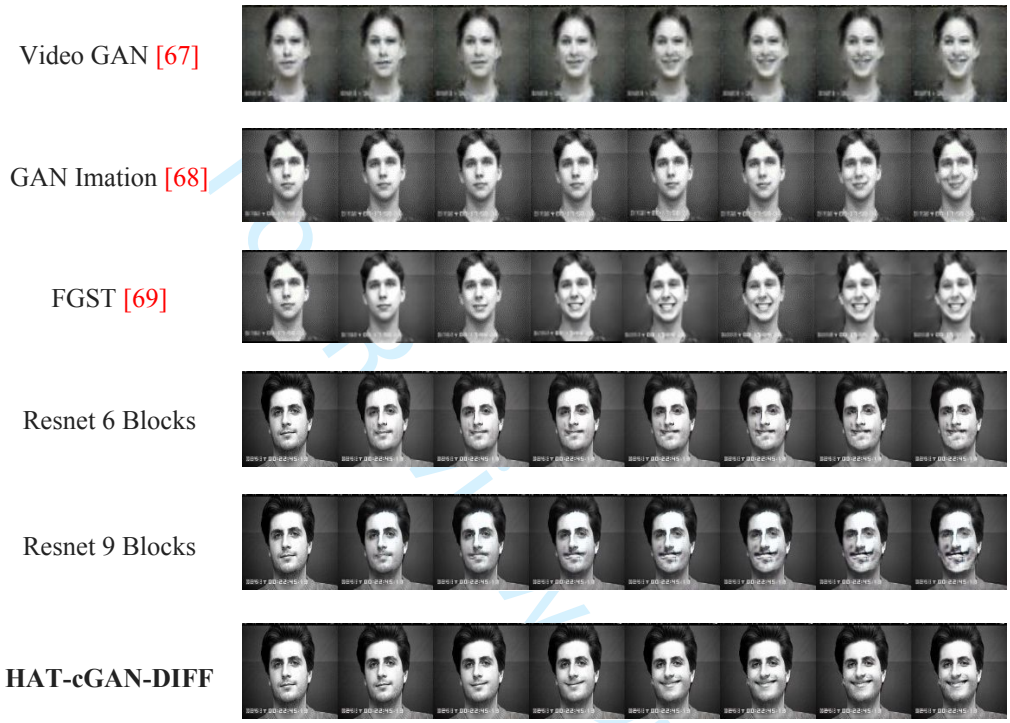

Fig. 8 Results of six models when creating gray images

Experimental results achieved on grayscale and color images show that our HAT-cGAN-DIFF model is more effective.

For color images, our proposed HAT-cGAN-DIFF model produces better, clearer and more realistic results than the VGAN, MoCoGAN and ImaGINator models. Previous models often produced images with high noise and unclear faces.

With grayscale images, the Video GAN, GAN Imation and FGST models gave poor results. The images are not realistic and many are blurry and have high noise. Our HAT-cGAN-DIFF method overcomes the above problems to produce high-resolution and sharp images.

Finally, our proposed HAT-cGAN-DIFF model can create more realistic and sharper images than using Resnet 6 Blocks and Resnet 9 Blocks in the cGAN generation network. These methods in both grayscale and color images produce poor quality, face images are distorted on color images and appear more numerous on grayscale images. Our HAT-cGAN-DIFF method produces images with higher contrast, lower noise, and higher resolution. This is due to the more complex U-Net network architecture combined with a denoised diffusion model of the video after being generated by cGAN, helping the network learn more complex features of the input data. Our HAT-cGAN-DIFF method can be used in applications that need to produce realistic and sharp videos.

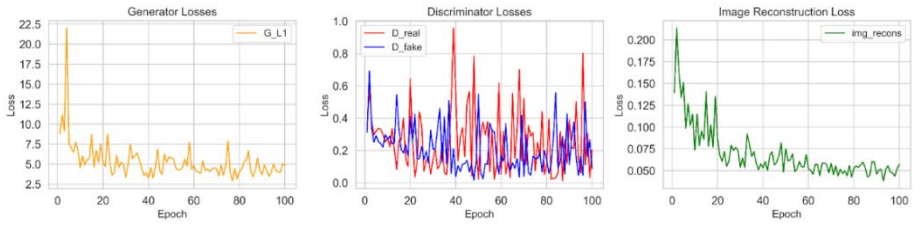

Fig. 9 Loss of the training processing on MUG datasets (color image)

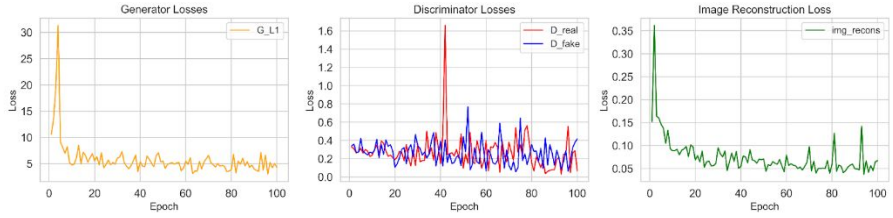

Fig. 10 Loss of the training processing on CK-Mixed datasets (gray image)

We evaluate the difference between the generated image (`fake_image`) and the real image (`real_image`) based on Generator loss. The goal is to minimize the deviation between the generated image and the real image. In **Fig. 9** and **Fig. 10**, we see that the loss gradually decreases over time as the Generator improves its image generation ability. The ultimate goal is to achieve the lowest G_L1 value, when the Generator creates images with high similarity to real images, which means the model creates high quality images that are difficult to distinguish from real images.

For Discriminator, *D_real* and *D_fake* are the output results of the discriminator when inputting real data and fake data respectively. D_real is the output of the discriminator when given real data. *D_fake* is the output of the discriminator when given fake data. The goal of the model is to create fake data that is difficult for the discriminator to distinguish from real data. From the results of Discriminator loss in **Fig. 9** and **Fig. 10**, we see that *D_real* is close to 1 and *D_fake* is close to 0, which helps the Generator create high-quality fake data that is difficult to distinguish from real data.

We also evaluate the difference between the reconstructed image (`*res_img*`) and the real image (`*real_image*`) based on Reconstruction loss. The goal is to minimize this loss to minimize the difference between the reconstructed image and the real image. Our results also show that the model is capable of accurately reproducing the original data as *img_recons* gradually decrease over time and without large fluctuations.

Quantitative Results:

To evaluate our HAT-cGAN-DIFF model against other models, we make a choice 5 progressive models that achieves good results for this problem is built based on the GAN model for comparison: VGAN [64], MoCoGAN [65], ImaGINator [66], FEV-GAN [70], CwGAN [71].

Regarding the models and metrics used for comparison, we use the results available in: ImaGINator, FEV-GAN, CwGAN. All these models are trained on the MUG dataset.

Table 1 Comparison results between our model and baselines using metrics SNR, SSIM, ACD and ACD-I

Model	PSNR	SSIM	ACD	ACD-I
VGAN [64]	14.54	0.28	0.14	1.55
MoCoGAN [65]	18.16	0.58	0.15	0.90
ImaGINator [66]	22.63	0.75	0.08	0.29
FEV-GAN [70]	27.10	0.91	0.09	0.23
CwGAN [71]	25.90	0.90	0.011	0.12
HAT-cGAN-DIFF	**31.56**	**0.95**	**0.06**	**0.10**

Looking at the results in **Table 1**, we see that our proposed method achieves the best results compared to other comparison methods in all 4 metrics:

First, we compare the reproducibility between our method and other methods using PSNR and SSIM scores. The results in **Table 1** show that our model outperforms other methods. This shows that the model we propose has good reconstruction ability and produces high quality videos and images.

Table 2 Inception Score and Frechet Inception Distance

	FID	IS
MocoGan [64]	45.46	3.24 ± 0.59
VGAN [65]	74.72	3.10 ± 0.38
ImaGINator [66]	29.02	-,
FEV-GAN [70]	-,	-,
CwGAN [71]	-,	3.52 ± 0.55
HAT-cGAN-DIFF	**11.27**	**4.05 ± 0.04**

Next, we evaluated the content consistency in generating facial expression videos using ACD, ACD-I scores. Our results in **Table 1** also show superiority over the remaining methods. This result shows that the ability of our proposed model to preserve appearance information in the generated videos is highly effective.

Finally, we evaluate the temporal consistency and quality of the generated video using FID and IS scores. The results in **Table 2** show low FID and high IS scores from our model, demonstrating that the videos produced with our method are highly consistent in time and of good quality. best images.

However, when analyzing in detail, we see that our model also has limitations such as: there are still some special cases where video synthesis is difficult, for example when the original photo is of poor quality. low or unclear, or when facial movements are complex and subtle. These restrictions can cause problems such as unnatural or inaccurate video.

5. Conclusion

This study introduces the novel HAT-cGAN-Diff model to generate high quality videos from facial images. We use Affine transforms and deep learning techniques to synthesize and process video sequences, then cGAN learn Affine transformations and synthesize images into videos. Ultimately, the video undergoes processing through the Diffusion model, which effectively filters out noise and enhances overall video quality. Rigorous experimentation on

various datasets, coupled with comparisons against diverse methods, substantiates the efficacy of our model. Looking ahead, we aim to augment the model's architecture to adeptly address specific and challenging scenarios, encompassing issues such as low-quality training data, variability in facial movements, and realistic video representation. Our ongoing efforts will delve into uncovering the model's potential and practical applications.

References

[1]. Mirza, Mehdi, and Simon Osindero. "Conditional generative adversarial nets." arXiv preprint arXiv:1411.1784, 2014. Article (CrossRef Link)

[2]. Nikhat Parveen, Prasun Chakrabarti, Bui Thanh Hung, Amjan Shaik. "Twitter Sentiment Analysis using Hybrid Gated Attention Recurrent Network". Journal of Big Data, 2023. Article (CrossRef Link)

[3]. B. Prasanalakshmi, Hung Bui Thanh, Chakrabarti Prasun, Chakrabarti Tulika, Elngar Ahmed A. and Rajanikanth Aluvalu. "A Novel Artificial Intelligence-Based Predictive Analytics Technique to Detect Skin Cancer". Peer J Computer Science Journal, 2023. Article (CrossRef Link)

[4]. Bui Thanh Hung, "Content based Image Retrieval using Multi-Deep Learning Models". Next Generation of Internet of Things. Lecture Notes in Networks and Systems, pp 347-357, vol 445. Springer, Singapore, 2022. Article (CrossRef Link)

[5]. Bui Thanh Hung, "Using Deep Unsupervised Method for Stock Prediction". Lecture Notes in Networks and Systems book, LNNS, volume 288, 2022. Article (CrossRef Link)

[6]. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y. "Generative adversarial nets". Advances in neural information processing systems, 2014. Article (CrossRef Link)

[7]. Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. "High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs". In CVPR, 2018. Article (CrossRef Link)

[8]. Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. "Style Aggregated Network for Facial Landmark Detection". In CVPR. 379–388, 2018. Article (CrossRef Link)

[9]. Mallya, A., Wang, T. C., Sapra, K., & Liu, M. Y. "World-consistent video-to-video synthesis". In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16 (pp. 359-378). Springer International Publishing, 2020. Article (CrossRef Link)

[10]. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. "Analyzing and improving the image quality of stylegan". In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8110-8119), 2020. Article (CrossRef Link)

[11]. Liu, B., Zhu, Y., Song, K., & Elgammal, A. "Towards faster and stabilized gan training for high-fidelity few-shot image synthesis". In International Conference on Learning Representations, 2020. Article (CrossRef Link)

[12]. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., & Aila, T. "Alias-free generative adversarial networks". Advances in Neural Information Processing Systems, 34, 852-863, 2021. Article (CrossRef Link)

[13]. Yang, S., Jiang, L., Liu, Z., & Loy, C. C. "Unsupervised image-to-image translation with generative prior". In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 18332-18341), 2022. Article (CrossRef Link)

[14]. Wang, T. C., Liu, M. Y., Zhu, J. Y., Tao, A., Kautz, J., & Catanzaro, B. "High-resolution image synthesis and semantic manipulation with conditional gans". In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8798-8807), 2018. Article (CrossRef Link)

[15]. Zhang, F., Zhang, T., Mao, Q., & Xu, C. "Joint pose and expression modeling for facial expression recognition". In Proceedings of the IEEE conference on computer vision and pattern recognition

(pp. 3359-3368), 2018. Article (CrossRef Link)

[16]. Lai, Y. H., & Lai, S. H. "Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition". In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) (pp. 263-270). IEEE, 2018. Article (CrossRef Link)

[17]. Emami, H., Aliabadi, M. M., Dong, M., & Chinnam, R. B. "Spa-gan: Spatial attention gan for image-to-image translation". IEEE Transactions on Multimedia, 23, 391-401, 2020. Article (CrossRef Link)

[18]. Choi, Y., Uh, Y., Yoo, J., & Ha, J. W. "Stargan v2: Diverse image synthesis for multiple domains". In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8188-8197), 2020. Article (CrossRef Link)

[19]. Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. "Image-to-image translation with conditional adversarial networks". In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1125-1134), 2017. Article (CrossRef Link)

[20]. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., ... & Taigman, Y. "Make-a-video: Text-to-video generation without text-video data". arXiv preprint arXiv:2209.14792, 2022. Article (CrossRef Link)

[21]. Hong, W., Ding, M., Zheng, W., Liu, X., & Tang, J. "Cogvideo: Large-scale pretraining for text-to-video generation via transformers". arXiv preprint arXiv:2205.15868, 2022. Article (CrossRef Link)

[22]. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., & Zhang, S. "Modelscope text-to-video technical report". arXiv preprint arXiv:2308.06571, 2023. Article (CrossRef Link)

[23]. Wu, J. Z., Ge, Y., Wang, X., Lei, S. W., Gu, Y., Shi, Y., ... & Shou, M. Z. "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation". In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 7623-7633), 2023. Article (CrossRef Link)

[24]. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., & Shi, H. "Text2video-zero: Text-to-image diffusion models are zero-shot video generators". arXiv preprint arXiv:2303.13439, 2023. Article (CrossRef Link)

[25]. Chan, C., Ginosar, S., Zhou, T., & Efros, A. A. "Everybody dance now". In Proceedings of the IEEE/CVF international conference on computer vision (pp. 5933-5942), 2019. Article (CrossRef Link)

[26]. Mallya, A., Wang, T. C., Sapra, K., & Liu, M. Y. "World-consistent video-to-video synthesis". In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16 (pp. 359-378). Springer International Publishing, 2020. Article (CrossRef Link)

[27]. Ni, H., Liu, Y., Huang, S. X., & Xue, Y. "Cross-identity video motion retargeting with joint transformation and synthesis". In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 412-422), 2023. Article (CrossRef Link)

[28]. Wang, T. C., Liu, M. Y., Tao, A., Liu, G., Kautz, J., & Catanzaro, B. "Few-shot video-to-video synthesis". arXiv preprint arXiv:1910.12713, 2019. Article (CrossRef Link)

[29]. Wang, Y., Yang, D., Bremond, F., & Dantcheva, A. "Latent image animator: Learning to animate images via latent space navigation". arXiv preprint arXiv:2203.09043, 2022. Article (CrossRef Link)

[30]. Sushko, V., Gall, J., & Khoreva, A. "One-shot gan: Learning to generate samples from single images and videos". In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2596-2600), 2021. Article (CrossRef Link)

[31]. Yao, X., Newson, A., Gousseau, Y., & Hellier, P. "A style-based gan encoder for high fidelity reconstruction of images and videos". In European conference on computer vision (pp. 581-597). Cham: Springer Nature Switzerland, 2022. Article (CrossRef Link)

[32]. Wu, Y., Singh, V., & Kapoor, A. "From image to video face inpainting: spatial-temporal nested

GAN (STN-GAN) for usability recovery". In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 2396-2405), 2020. Article (CrossRef Link)

[33]. Li, S., Han, B., Yu, Z., Liu, C. H., Chen, K., & Wang, S. "I2v-gan: Unpaired infrared-to-visible video translation". In Proceedings of the 29th ACM international conference on multimedia (pp. 3061-3069), 2021. Article (CrossRef Link)

[34]. Sushko, V., Gall, J., & Khoreva, A. "One-shot gan: Learning to generate samples from single images and videos". In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2596-2600), 2021. Article (CrossRef Link)

[35]. Dorkenwald, M., Milbich, T., Blattmann, A., Rombach, R., Derpanis, K. G., & Ommer, B. "Stochastic image-to-video synthesis using cinns". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3742-3753), 2021. Article (CrossRef Link)

[36]. Ni, H., Shi, C., Li, K., Huang, S. X., & Min, M. R. "Conditional Image-to-Video Generation with Latent Flow Diffusion Models". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18444-18455), 2023. Article (CrossRef Link)

[37]. Saito, M., Saito, S., Koyama, M., & Kobayashi, S. "Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan". International Journal of Computer Vision, 128(10-11), 2586-2606, 2020. Article (CrossRef Link)

[38]. Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., & Li, D. "Makelttalk: speaker-aware talking-head animation". ACM Transactions On Graphics (TOG), 39(6), 1-15, 2020. Article (CrossRef Link)

[39]. Ching, Wai-Ki, and Michael K. Ng. "Markov chains. Models, algorithms and applications", 2006. Article (CrossRef Link)

[40]. Ho, J., Jain, A., & Abbeel, P. "Denoising diffusion probabilistic models". Advances in neural information processing systems, 33, 6840-6851, 2020. Article (CrossRef Link)

[41]. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. "High-resolution image synthesis with latent diffusion models". In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695), 2022. Article (CrossRef Link)

[42]. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. "Hierarchical text-conditional image generation with clip latents". arXiv preprint arXiv:2204.06125, 1(2), 3, 2022. Article (CrossRef Link)

[43]. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., ... & Chen, M. "Glide: Towards photorealistic image generation and editing with text-guided diffusion models". arXiv preprint arXiv:2112.10741, 2021. Article (CrossRef Link)

[44]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. "Bert: Pre-training of deep bidirectional transformers for language understanding". arXiv preprint arXiv:1810.04805, 2018. Article (CrossRef Link)

[45]. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. "Learning transferable visual models from natural language supervision". In International conference on machine learning (pp. 8748-8763). PMLR, 2021. Article (CrossRef Link)

[46]. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., ... & Norouzi, M. "Photorealistic text-to-image diffusion models with deep language understanding". Advances in Neural Information Processing Systems, 35, 36479-36494, 2022. Article (CrossRef Link)

[47]. Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., ... & Wu, Y. "Scaling autoregressive models for content-rich text-to-image generation". arXiv preprint arXiv:2206.10789, 2(3), 5, 2022. Article (CrossRef Link)

[48]. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. "Zero-shot text-to-image generation". In International Conference on Machine Learning (pp. 8821-8831). PMLR, 2021. Article (CrossRef Link)

[49]. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., ... & Salimans, T. "Imagen video: High definition video generation with diffusion models". arXiv preprint arXiv:2210.02303, 2022. Article (CrossRef Link)

[50]. DALL-E 3. https://openai.com/dall-e-3
[51]. Dhariwal, P., & Nichol, A. "Diffusion models beat gans on image synthesis". Advances in neural information processing systems, 34, 8780-8794, 2021. Article (CrossRef Link)
[52]. Ronneberger, O., Fischer, P., & Brox, T. "U-net: Convolutional networks for biomedical image segmentation". In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18 (pp. 234-241). Springer International Publishing, 2015. Article (CrossRef Link)
[53]. Lee, B., Lei, F., Chen, H., & Baudron, A. "Bokeh-loss gan: Multi-stage adversarial training for realistic edge-aware bokeh". In European Conference on Computer Vision (pp. 619-634). Cham: Springer Nature Switzerland, 2022. Article (CrossRef Link)
[54]. Aifanti, N., Papachristou, C., & Delopoulos, A. "The MUG facial expression database". In 11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10 (pp. 1-4). IEEE, 2010.
[55]. Lucey, Patrick, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, 2010 IEEE computer society conference on computer vision and pattern recognition-workshops. IEEE, 2010.
[56]. Kingma, D. P., & Ba, J. "Adam: A method for stochastic optimization". arXiv preprint arXiv:1412.6980, 2014. Article (CrossRef Link)
[57]. Orhan, A. E., & Pitkow, X. "Skip connections eliminate singularities". arXiv preprint arXiv:1701.09175, 2017. Article (CrossRef Link)
[58]. Hore, A., & Ziou, D. "Image quality metrics: PSNR vs. SSIM". In 2010 20th international conference on pattern recognition (pp. 2366-2369). IEEE, 2010. Article (CrossRef Link)
[59]. Abu-Srhan, A., Abushariah, M. A., & Al-Kadi, O. S. "The effect of loss function on conditional generative adversarial networks". Journal of King Saud University-Computer and Information Sciences, 34(9), 6977-6988, 2022. Article (CrossRef Link)
[60]. Bao, S. "Review on Generative Adversarial Network in Computer Vision: Methods and Metrics". In 2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE) (pp. 535-542). IEEE, 2021. Article (CrossRef Link)
[61]. Zhao, L., Peng, X., Tian, Y., Kapadia, M., & Metaxas, D. "Learning to forecast and refine residual motion for image-to-video generation". In Proceedings of the European conference on computer vision (ECCV) (pp. 387-403), 2018. Article (CrossRef Link)
[62]. Soloveitchik, M., Diskin, T., Morin, E., & Wiesel, A. "Conditional frechet inception distance". arXiv preprint arXiv:2103.11521, 2021. Article (CrossRef Link)
[63]. Obukhov, A., & Krasnyanskiy, M. "Quality assessment method for GAN based on modified metrics inception score and Fréchet inception distance". In Software Engineering Perspectives in Intelligent Systems: Proceedings of 4th Computational Methods in Systems and Software 2020, Vol. 1 4 (pp. 102-114). Springer International Publishing, 2020. Article (CrossRef Link)
[64]. Chen, J., Konrad, J., & Ishwar, P. "Vgan-based image representation learning for privacy-preserving facial expression recognition". In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 1570-1579), 2018. Article (CrossRef Link)
[65]. Tulyakov, S., Liu, M. Y., Yang, X., & Kautz, J. "Mocogan: Decomposing motion and content for video generation". In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1526-1535), 2018. Article (CrossRef Link)
[66]. Wang, Y., Bilinski, P., Bremond, F., & Dantcheva, A. "Imaginator: Conditional spatio-temporal gan for video generation". In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1160-1169), 2020. Article (CrossRef Link)
[67]. Vondrick, C., Pirsiavash, H., & Torralba, A. "Generating videos with scene dynamics". Advances in neural information processing systems, 29, 2016. Article (CrossRef Link)
[68]. Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A., & Moreno-Noguer, F. "Ganimation: Anatomically-aware facial animation from a single image". In Proceedings of the European conference on computer vision (ECCV) (pp. 818-833), 2018. Article (CrossRef Link)
[69]. Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. "Flow-

Grounded Spatial-Temporal Video Prediction from Still Images". In The European Conference on Computer Vision (ECCV). 609–625, 2018. Article (CrossRef Link)

[70]. Bouzid, H., & Ballihi, L. "Facial expression video generation based-on spatio-temporal convolutional GAN: FEV-GAN". Intelligent Systems with Applications, 16, 200139, 2022. Article (CrossRef Link)

[71]. Naima Otberdout, Mohamed Daoudi, Anis Kacem, Lahoucine Ballihi, Stefano Berretti. "Dynamic facial expression generation on hilbert hypersphere with conditional wasserstein generative adversarial nets". IEEE Transactions on Pattern Analysis and Machine Intelligence 44.2: 848-863, 2020. Article (CrossRef Link)