Latent Flow Diffusion for Deepfake Video Generation

Aashish Chandra K[1*] Aashutosh A V[1*] Srijan Das[2]
Abhijit Das[1†]
[1]Birla Institute of Technology & Science, Pilani, Hyderabad campus, India,
[2] University of North Carolina Charlotte, USA
abhijit.das@hyderabad.bits-pilani.ac.in

Abstract

Image-to-video generation with conditional identity swap popularly known as deepfake, aims to synthesize a new video for the target identity guided by an image of the target and a video of the source identity. The biggest challenge of these tasks lies in the simultaneous generation of realistic spatial appearance and temporal dynamics corresponding to the given target image and source video. In this paper, we propose a deepfake generation technique using novel latent flow diffusion (LFD) that includes an optical flow sequence in the latent space based on a given source video to warp the given target image. Compared to previous works on video diffusion, our proposed LFD can swap the spatial details maintaining temporal information by utilizing the spatial content of the given target image and employing the latent flow of the source video. Our model consists of three stages: a Flow predictor model captures the optical flow of the source video, two-fold Transformer encoding layers predict the driving frame and a conditioned image-to-video generator guided by the driving frame generates the final deep fake video. We conducted multiple experiments and our proposed model has consistently outperformed prior video diffusion models for deepfake generation.

1. Introduction

Given an input reference image and a video, a deepfake model tries to generate a fake video by morphing the face in the reference image onto the input video [35]. Several generative techniques have been employed in the literature [45]. The main challenge is to preserve both the intricacies of the facial structure of the input image and the facial actions of each frame in the input video. Among the generative techniques, diffusion models have proven their superiority in recent years.

Diffusion models have proven to be potent tools in various content creation tasks, spanning from image-to-image generation to text-to-image and 3D object generation [8]. However, while these models have seen success in static image generation, generating high-quality videos is challenging due to the complex spatio-temporal information they encompass. Recent advancements have emphasized the importance of reimagining the backbones of diffusion models, highlighting the significance of innovative architectures in enhancing performance. Building onto pre-trained video diffusion models[18], image-to-video generation is a prominent topic in the avenue of video diffusion [29].

In image-to-video diffusion a single image and a condition or a video aim to generate a realistic video, replicating the given image and satisfying the condition. Alike to conditional image synthesis[19], the existing conditional image-to-video generation techniques[29, 37] generate frames in the video based on the given image and condition. The major challenge faced by such models is seamless spatial and temporal details in the generated videos.

Recently, latent flow diffusion models (LFDM) have been proposed [29], which tackles this by employing a latent optical flow sequence and condition, to warp the image along with the latent space for synthesising the new videos. Deviating from existing latent diffusion-based video generation which either uses direct-synthesis or warp-free methods, LFDM uses the spatial content of the given image. Hence, the spatial content can be used consistently to generate temporally coherent flow maintaining subject appearance and motion dynamics. Even then, generating image-to-video with conditional identity swap,*i.e.*, deep fake generation has not been well explored using diffusion.

A deep fake video generation model aims to synthesize a new video for the target identity guided by an image of the target identity and a video of the source identity. The challenge is the simultaneous generation of realistic spatial appearance and temporal dynamics corresponding to the given

*Equally contributing first authors.
†Corresponding authors.

target image and source video. Such identity swapping has a more intense challenge than the image-to-video generation scenario. To disentangle the generation of spatial content of the target image maintaining the temporal properties of the source video needs extra attention. We need to preserve the flow at the same time the spatial details.

Hence, to solve this challenge we proposed LFD deepfake generation (LFD2G) which works in three stages (see Fig 1). For capturing the flow of the source video we employed a latent Flow predictor. This finds the optical flow between every two frames of the source video. This is needed as an input to our conditional image-to-video generator. This has been trained in an unsupervised fashion on the MUG dataset [26]. Next, we have two transformer encoding layers, one being a self-attention block on the target image and the other being a cross-attention block between the target image and the first frame of the source video (see Fig 1). This dual Transformer block predicts the driving frame which is the first frame of our generated deepfake video. This stage can be replaced by any deepfake-image generation technique, but we used a vision Transformer [13] as our backbone. In the final stage, a diffusion model (DM) is trained using a driving frame and latent flow sequence extracted from source video produced from the trained flow predictor. From the driving frame and the generated frame, the DM aims to learn temporally coherent latent flow sequences by 3D convolutions and thereby produces a final deepfake video. The latent feature space treats the spatial and temporal information, in a simple and low-dimensional latent flow space which is only responsible for the motion dynamics. Hence, the diffusion generation process remains computationally efficient which is very important, as most of the Video-diffusion models are very computationally expensive. Our contributions are summarized as follows:

- We propose a novel latent flow diffusion for deepfake generation (LFD2G) by employing a temporally coherent flow sequence in the latent space, which is based on the given source video. To the best of our knowledge, this is the first work to apply latent flow diffusion models to generate deepfake generation.
- A novel three-stage learning strategy to disentangle the generation of the spatial context of the target image and temporal dynamics of the source video, by training a flow predictor, dual-transformer encoding block and a target image-guided 3D diffusion model.
- We conduct extensive experiments on multiple scenarios and ground truth references, where proposed LFD2G consistently outperforms previous state-of-the-art methods both in terms of qualitative and quantitative analysis.

2. Related Work

We proceed to list the recent works in the literature on diffusion and deep fake generation.

2.1. Diffusion Models for Image Generation

Diffusion Models(DMs)[17] have recently taken centre stage in Generative AI. These models have had immense success in generating images[11][33][49]. Diffusion models (DM) have a 2D UNet backbone[34], and have been experimentally proven to be better, in most cases, than GANs[15]. Ideas like latent diffusion[33] where DMs are applied within the latent space of pre-trained autoencoders have been proposed, which further demonstrates the versatility of DM in generating high-quality images.

2.2. Diffusion models for Video Generation

Diffusion models have also found a strong footing in the video generation[3][4], by using 3D UNet[6] backbone. Specifically, the model LFDM[29] is different from most other models, instead applies diffusion to generate latent flow sequences for conditional image-to-video generation.

Conditional video generation[29] represents a significant advancement in the field of computer vision, aiming to synthesize videos guided by user-provided signals. This approach encompasses various methodologies tailored to different input modalities, including text-to-video (T2V)[21], video-to-video (V2V)[47], and image-to-video (I2V) generation[48][4]. Within the realm of I2V generation, which closely aligns with video prediction from single images, there exists a distinction between stochastic methods utilizing only a given image as input and conditional generation techniques, which incorporate additional conditions alongside the base image.

Traditionally, conditional image-to-video (cI2V) generation[29] has relied on diverse strategies to achieve realistic and diverse video synthesis. Noteworthy among these approaches are methods such as pose-guided synthesis, interactive models enabling user-guided motion specification, and techniques leveraging optical flow estimation for motion synthesis. However, despite their efficacy, these methods often face challenges in generating videos with complex motions or fine-grained details.

2.3. Vision Transformers

Transformers[46] have recently made significant strides in revolutionizing the field of computer vision, and dominated traditional convolutional neural networks(CNNs)[30]. While CNNs have been remarkably successful in image classification, object detection[14], and segmentation[40] tasks, they struggle with capturing long-range dependencies and contextual information across images effectively. Transformers address this limitation by employing self-attention mechanisms, enabling them to capture global context information efficiently. Vision transformers (ViTs)[13], introduced by Dosovitskiy et al. in the landmark paper "An Image is Worth 16x16 Words," break down images into fixed-size patches, which are then flattened

and fed into a transformer encoder for processing. This approach allows ViTs[13] to achieve competitive performance on various vision tasks, such as image classification, object detection, and semantic segmentation, often surpassing the performance of traditional CNN architectures on challenging datasets. There are other state-of-the-art models like StableVITON[22], which incorporate the use of transformers for computer vision tasks. The success of transformers in computer vision underscores their versatility and potential for advancing the state-of-the-art in visual recognition tasks.

2.4. Deepfakes Detection

Deepfakes refer to multimedia content in which faces have been digitally altered or synthetically created using deep neural networks. Despite significant progress based on traditional and advanced computer vision, artificial intelligence, and physics, there is still a huge arms race surging up between attackers/offenders/adversaries (i.e., DeepFake generation[1, 32] methods) and defenders (i.e., DeepFake detection[1, 16, 44] methods)[28]. Some of the popular deepfake detection models which we have seen are Selim (DFDC Winner [12]), SSAT[10], Cross Efficient ViT (CFNet)[7] which can detect deepfakes with upto 60% accuracy. Therefore, there is a lot of opening in this field for vision scientists to participate.

2.5. Deepfake Generation

Some of the most widely used Deepfake datasets are Celeb-DF[27] and the FaceForensics++[35] dataset. FF++ is a forensics dataset consisting of 1000 original video sequences that have been manipulated with three automated face manipulation methods[2]: Deepfakes[5], Face2Face[43] and NeuralTextures[42]. The authors of this paper also provide 1000 Deepfakes models to generate and augment new data. With deepfake generation being a very pursued problem recently, we find a lot of active research in this field, one of which is Deepfake generation using GANS[36][39][41]. While GANs have been state-of-the-art (SoTA) for generating deepfakes for a long, there are other methods that have provided good, if not better results. The FaceShifter model[25][9] trains attention layers to adaptively integrate attributes of the face. The FaceSwap[24][9] model using Generative Autoencoders[31][50]. Since GANs often struggle to preserve subtle yet crucial identity details of source faces, there has been a shift towards exploring alternative architectures for deepfake generation. What has become the more interesting domain to explore is generating Deepfakes through Diffusion[5]. With diffusion having higher image fidelity than GANs and other AutoEncoder Models, it is natural that the preference towards Diffusion has increased for generating face images, and therefore deepfakes using them.

3. Proposed method

In this section, we aim to describe the preliminaries of our proposed model, an overview of the method and details of the proposed LFD2G.

3.1. Preliminaries

Let $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be a Gaussian noise volume with the shape of $K \times H \times W \times C$, where K H, W, and C are length, height, width, and channel number, respectively. Given one starting image x_0 (tatget image) and condition y (source video in our scenario), let $\mathbf{x}_0^K = \{x_0, x_1, \ldots, x_K\}$ be the real video of condition y, the goal of conditional image-to-video generation (cI2V) [29] is to learn a mapping that converts the noise volume $\mathbf{n}$ to a synthesized video, $\hat{\mathbf{x}}_1^K = \{\hat{x}_1, \ldots, \hat{x}_K\}$, so that the conditional distribution of $\hat{\mathbf{x}}_1^K$ given x_0 and y is identical to the conditional distribution of $\mathbf{x}_1^K$ given x_0 and y, *i.e.*, $p(\hat{\mathbf{x}}_1^K|x_0, y) = p(\mathbf{x}_1^K|x_0, y)$. This would also imply that conditional distribution of $\hat{\mathbf{x}}_0^K$ given y is identical to the conditional distribution of $\mathbf{x}_0^K$ given y ,*i.e.*, $p(\hat{\mathbf{x}}_0^K|y) = p(\mathbf{x}_0^K|y)$.

3.2. Method Overview

Given a source video of a person $\mathbf{x}_0^K = \{x_0, x_1, \ldots, x_K\}$, and a target image y_0 of another person we aim to generate a target person video imitating the source video, $\hat{\mathbf{y}}_0^K = \{\hat{y}_0, \hat{y}_1, \ldots, \hat{y}_K\}$. This is the basic formulation of deepfake video generation.

Here $\mathbf{x}_0^K \in \mathbb{R}^{K \times H \times W \times C}$ and $x_0 \in \mathbb{R}^{H \times W \times C}$, where K, H, W, C are the number of frames, height, width and number of channels respectively. Given $\hat{y}_0$ we would be able to produce $\hat{\mathbf{y}}_0$ using cI2V formulation [29] . Note that $\mathbf{y}_0 \neq \hat{\mathbf{y}}_0$, we will first need to find $\hat{y}_0$. For this, we can use any deep fake image generation model (vision transformer [13] for our scenario)

The LFD2G model is based on denoising diffusion probabilistic models (DDPM)[17]. In DDPM given a sample from the data distribution $S_0 \sim Q(S_0)$, in the forward process a Markov chain $S_1, ..., S_T$ is produced by adding Gaussian noise to S_0 maintaining a variance schedule $\beta_1, ..., \beta_T$, where variances of β_t are constant. If β_t value is small, the posterior $Q(S_{t+1}|S_t)$ can be estimated by diagonal Gaussian. Further, if T of the chain is large, S_T will well estimate the standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Hence the true posterior of $Q(S_{t+1}|S_t)$ can be estimated.

In DDPM reverse process, samples $S_0 \sim P_\theta(S_0)$ are produced by starting with Gaussian noise $\mathbf{S}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and gradually reducing noise in a Markov chain of $S_{T-1}, S_{T-2}, ..., S_0$ with learnt $P_\theta(S_{T-1}|S_T)$. To get $P_\theta(S_{T-1}|S_T)$, Gaussian noise ϵ is added to S_0 to generate samples $S_T \sim Q(S_T|S_0)$, then a model ϵ_θ is trained to predict ϵ using mean-squared error loss. A time step T is uniformly sampled from the time stamp $1, ..., T$. For denoising the model ϵ_θ is done by time-conditioned U-Net

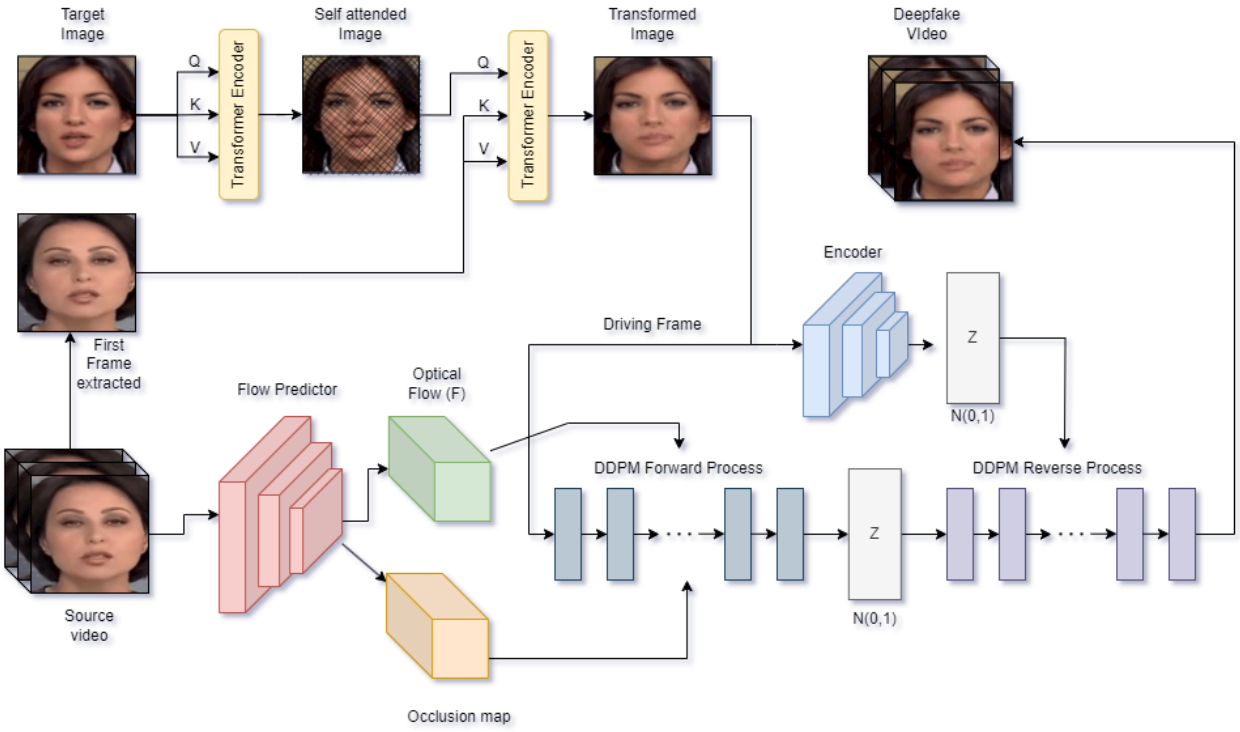


Figure 1. Architecture of the proposed LFD2G.

with residual blocks and self-attention layers. The time step T is specified to ϵ_θ via the sinusoidal position embedding. To conditionally generate, it can be learned simply via a y-conditioned model.

For our scenario, the condition is the source video and the input data is the target deepfake image to generate the target deepfake video.

3.3. The proposed LFD2G details

An overview of LFD2G is presented in Fig. 1. Our pipeline consists of three stages. In the first stage, a flow predictor F predicts the latent flow f for the final conditioned video ($\hat{\mathbf{y}}_0^K$) by taking the source video as input ($\mathbf{x}_0^K$). The flow considers both the horizontal and vertical movement between frames. The model also used *backward* flow by a differentiable bilinear sampling operation [20]. Similar to [29, 38, 47], flow predictor F also estimates a latent occlusion map m. This is needed as input to the conditional image-to-video generator.

The second stage uses a dual Vision transformer [13] encoding block to find $\hat{y}_0$, *i.e.*, the target deepfake image. We chose this architecture because of the transformer's ability to capture long-range dependencies with the input data. This is particularly crucial in facial manipulation tasks where subtle spatial relationships between facial features play a significant role. The query, key, and value to the first Multi-head attention block are the target image (y_0). The query to the second Multi-head Attention block is the output of the first Multi-head Attention block (y_0') and the key and value are the first frame of the source video (x_0). The output of the second Multi-head Attention block is $\hat{y}_0$ which is the Transformed image in Fig. 1. It is the starting frame of the deepfake target video which is the last stage of the pipeline.

The last/ third stage of the pipeline is the conditional image-to-video generation[29]. Here the condition is the source video ($\mathbf{x}_0^K$) (the latent flow captured) and the driving image is the output of our Attention layers ($\hat{y}_0$). The 3D-UNet-based DM is trained to achieve a temporally coherent latent flow sequence conditioned on the driving image. A 3D Gaussian noise by the DDPM forward process is employed. The encoder represents the starting frame as a latent map and the target frame encodes the condition as image embedding. The Denoising model is trained to predict the added noise based on a conditional 3D U-Net[6] with the diffusion loss.

As we can see from Fig. 1, both the optical flow and the occlusion map are passed as input to the DDPM forward process, along with the Driving frame as the transformed image. During the DDPM reverse process, an encoding of the transformed image is passed to the denoising model. Conditioned on the transformed image, the denoising model

(a) FaceSwap

(b) Face2Face

(c) NeuralTextures

Figure 2. The 1st row of each group represents the baseline results for LFDM in zero-shot and the 2nd row represents our trained model's results. The columns from left to right represent the Target Image (to be morphed on the video), Sample source video, Warped Output Video, Model Output Video, and corresponding Deepfake example from FF++.

is trained to predict the deepfake video.

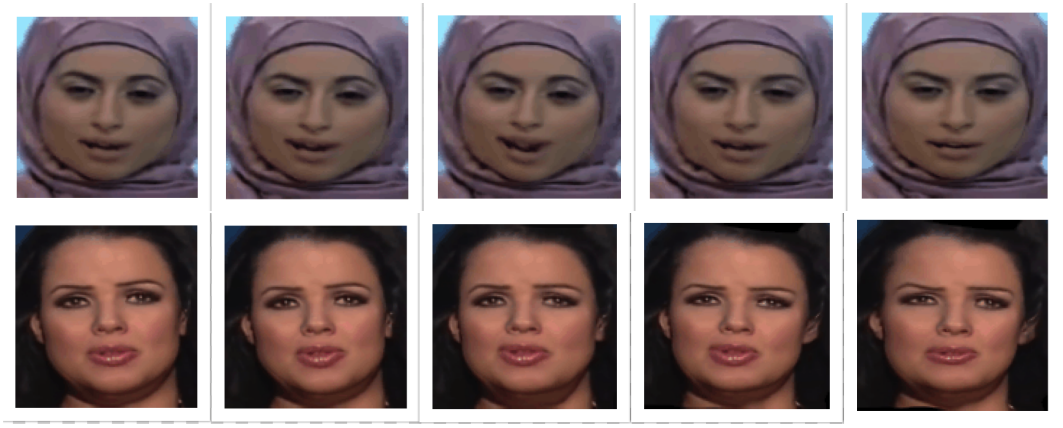
Each iteration of training involves the model processing both the real video and the target image, with the attention mechanism being trained every iteration undergoing optimisation using the Adam optimizer[23], with the goal of reducing the reconstruction loss between the generated deepfake video and the original deepfake video from the dataset.

4. Experimental results

In this section we will describe the experimental protocol and analysis we employed to validate our proposed model.

(a) FaceSwap

(b) Face2Face



(c) NeuralTextures

Figure 3. These figures from left to right show subsequent frames of our generated videos with the 3 techniques: FaceSwap, Face2Face and NeuralTextures.

The analysis includes both the quantitative analysis w.r.t deepfake detection algorithms and also w.r.t FVD and FID.

4.1. Datasets and Metrics

- **Pretraining Dataset:** The diffusion model being used has been first pre-trained on MUG dataset [26] which consists of 52 videos of shape $40 * 128 * 128 * 3$.
- **Dataset:** Our model is tested for generation capabilities on Deepfake datasets such as the FaceForensics++ (FF++) Dataset[35], containing manipulation techniques such FaceSwap[25], the Neural Textures[42] and the Face2Face[43]. Each of these consists of 1000. By us-

ing these three variety of manipulation techniques encompassing a wide range of facial manipulation scenarios, they provide a robust evaluation platform for our model. The fake videos from FF++ were used as ground truth for the generation of our target video, original videos from FF++ were used as the source video, and a frame from each corresponding fake video was treated as the target image.

- **Data Preprocessing**. All the videos from FF++ are resized to 128128 resolution and clipped to 40 frames.
- **Evaluation metrics:** We compute the Frechet Video Distance (FVD) to measure the dissimilarity between the distributions of feature representations extracted from real and synthesized videos. It assesses the visual quality, temporal coherence, and sample diversity of generated videos by comparing their feature distributions.. We also used Frechet Inception Distance (FID) to evaluate the quality of generated images by computing the distance between feature distributions of real and generated images. It leverages the Inception network to extract feature representations and provides a metric of image quality., FVD first employs a video classification network I3D pretrained on the Kinetics-400 dataset to obtain feature representation of real and synthesized videos. Next, it calculates the Frechet distance between the distributions of real and synthesized video features. To gauge how well a generated video corresponds to the condition i.e. the subject relevance, similar to the conditional FID, subject conditional FVD (sFVD) is employed. The sFVD compares the distance between real and synthesized video feature distributions under the same condition or the same subject.

  As our model aims to generate deepfakes, we consider several deep-fake detection techniques to check the generation quality. We did not fine-tune the model. Rather we just used it for testing. The deep fake detection models considered are Selim (DFDC Winner [12]), SSAT[10], Cross Efficient ViT[7]. All models are trained on DFDC[12] and FF++[35].

4.2. Implementation details

We build our model based on a comprehensive implementation of the Latent Flow Diffusion Model[29]. We used a latent diffusion[33] model because of its ability to manipulate the frames in a more controlled way by only modifying the latent code. We employed the pre-trained weight of the Latent Flow Diffusion Model to train our model.

Our model includes three trainable modules: a Flow Predictor, a Denoising model from DDPM and two Transformer Encoding layers. The flow predictor F is implemented with [38], which can estimate latent flow and occlusion map based on detected objects. The DDPM model is implemented using reference code from [17, 29] which has a backbone of 3D-Unet [6]. The Transformer Encoding layers have been implemented using a vision transformer [13]. Any deepfake-image generation technique can be used here, but we chose a transformer to show that the pipeline works.

Training a 3D diffusion model from scratch would require a lot of computing power, so we used a pre-trained diffusion model (weights from LFDM [29]) and just trained the transformer encoding layers. We used a learning rate of 10^{-4} with $\alpha = 0.9$, $\beta_1 = 0.9$ and $\beta_2 = 0.99$. We had to use a multi-GPU approach since the transformers use a lot of VRAM. This involved loading the first transformer into one GPU and the second transformer into another GPU. We trained our model for 5 epochs and noticed that we had almost reached convergence. This is because transformers are very good at remembering things. Trying to give attention to a full video would take too much VRAM and much more time for training.

Finally, the inference pipeline is the same as the training pipeline. The only difference is that all the weights are frozen. The final outputs have been saved as videos, on which our analysis has been done.

4.3. Result Analysis

The output of our model, the ground truth from FF++ dataset (Fake video generated using deepfake technique) and the base LFDM model with zero-shot are in Fig. 2. The output of our model, for two different scenarios with different groudtruth conditions, are in Fig. 3. It can be concluded from the visual analysis that our model has a very close output to the ground truth video and the best output was generated for Face2face. Additionally, the quality of the proposed model is much better than the output of the base model (LFDM [29]) with zero-shot.

Table 1. The quantitative analysis w.r.t deep fake detection, all results are in accuracy %

		Selim	CFNet	SSAT
Face to face	Ground truth	61.10	61.91	64.22
	LFDM	65.50	63.45	69.12
	Proposed	**61.23**	**61.74**	**64.90**
Neural texture	Ground truth	63.11	63.41	69.73
	LFDM	75.00	74.50	77.64
	Proposed	**63.30**	**63.71**	**69.90**
Face swap	Ground truth	65.01	66.45	72.81
	LFDM	75.00	74.50	77.64
	Proposed	**65.30**	**66.71**	**72.90**

Now we proceed for quantitative analysis w.r.t to deep fake detection techniques considering Selim (DFDC Winner [12]), SSAT[10], Cross Efficient ViT (CFNet)[7] (See Table 1). It can be concluded that our model has performed much better (lower score of the deep fake detection model

implies better performance of our generation model) than the baseline and is quite near to the ground truth.

Table 2. The quantitative analysis w.r.t FVD, sFVD and FID

		FVD	FID	sFVD
Face to face	LFDM	128.10	73.45	307.80
	Proposed	**75.23**	**25.11**	**101.00**
Neural texture	LFDM	131.20	74.20	319.10
	Proposed	**81.11**	**33.41**	**108.12**
Face swap	LFDM	172.00	75.50	311.00
	Proposed	**85.60**	**36.11**	**110.01**

Table 2 provides a detailed quantitative analysis of our proposed deepfake generation model in terms of Frechet Video Distance (FVD), subject conditional FVD (sFVD), and Frechet Inception Distance (FID). The FID, sFVD and FVD of the proposed model are quite low compared to the baseline (the lower the value of sFVD, FID and FVD, the better the performance of the generation model). For calculating sFVD, FID and FVD the corresponding fake image from FF++ was considered as reference video or ground truth. FVD proves proper visual quality was attended by our generative model. On the other side, low FID concludes with the overall image quality of the model. Moreover, low sFVD proves the subject condition of the generated image. Hence this analysis proves the superior performance and effectiveness of our model for deepfake generation. From all the qualitative and quantitative analysis it can also be concluded that while Faceforencics++ dataset our proposed model produces very good videos with on par performance as the Ground Truth.

4.4. Model Inference Analysis

Some analyses on model inference such as FLOPs, MACs and Params are in Table 3.

Table 3. Computational Complexity Comparison

Model	**FLOPs**	**MACs**	**Params**
Flow Diffusion	1.1992 T	598.548 G	42.7312 M
Attention	9.6652 G	4.8321 G	1.6107 B

Table 3 provides a comprehensive comparison of the computational complexity between the Flow Diffusion and Attention components of our proposed model. The Flow Diffusion component exhibits a significantly lower computational load compared to the Attention component. Specifically, the Flow Diffusion component requires 1.1992 TFLOPS (Floating Point Operations per Second) and 598.548 GMACs (Giga Multiply-Accumulates), with a parameter count of 42.7312 million. In contrast, the Attention component demands substantially higher computational resources, with 9.6652 GFLOPS and 4.8321 GMACs, along with a parameter count of 1.6107 billion. These findings highlight the disparity in computational requirements between the two components, providing insights into the resource allocation and optimization strategies for our model implementation.

5. Conclusion

In this paper, we propose a novel deepfake generation technique employing latent flow diffusion. The proposed model LFD2G, generates fake or target videos by warping given target images with flow sequences generated in the latent space based on source video. We conducted comprehensive experiments from which we can infer that the proposed model achieves state-of-the-art performance under diverse conditions for fake video generation. The model can generate fake videos of faces concentrating on the facial region.

There are still challenges with artifacts such as motions and temporal detailing that we will address in our future work. Further, the future scope will be to extend the generation for whole body structure.

Acknowledgements

This work was funded by the Institute of Data Engineering, Analytics, and Science (IDEAS) Technology Innovation Hub (TiH) Indian Statistical Institute, Kolkata, under the aegis of National Mission on Interdisciplinary Cyber-Physical Systems (NM-ICPS), Department of Science and Technology (DST), Government of India under the project titled "Generalized Tampering Detection in Media (GTDM)" and project number OO/ISI/IDEAS-TIH/2023-24/86.

References

[1] Zahid Akhtar. Deepfakes generation and detection: A short survey. *Journal of Imaging*, 9(1), 2023. 3

[2] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 3

[3] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation, 2024. 2

[4] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation, 2023. 2

[5] Yunzhuo Chen, Nur Al Hasan Haldar, Naveed Akhtar, and

Ajmal Mian. Text-image guided diffusion model for generating deepfake celebrity interactions, 2023. 3

[6] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016. 2, 4, 7

[7] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. Combining efficientnet and vision transformers for video deepfake detection. In *International conference on image analysis and processing*, pages 219–229. Springer, 2022. 3, 7

[8] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[9] Kaiwen Cui, Rongliang Wu, Fangneng Zhan, and Shijian Lu. Face transformer: Towards high fidelity and accurate face swapping, 2023. 3

[10] Srijan Das, Tanmay Jain, Dominick Reilly, Pranav Balaji, Soumyajit Karmakar, Shyam Marjit, Xiang Li, Abhijit Das, and Michael S Ryoo. Limited data, unlimited potential: A study on vits augmented by masked autoencoders. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6878–6888, 2024. 3, 7

[11] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *CoRR*, abs/2105.05233, 2021. 2

[12] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020. 3, 7

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2, 3, 4, 7

[14] Reagan L. Galvez, Argel A. Bandala, Elmer P. Dadios, Ryan Rhay P. Vicerra, and Jose Martin Z. Maningo. Object detection using convolutional neural networks. In *TENCON 2018 - 2018 IEEE Region 10 Conference*, pages 2023–2027, 2018. 2

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[16] David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2018. 3

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *CoRR*, abs/2006.11239, 2020. 2, 3, 7

[18] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 1

[19] Xun Huang, Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Multimodal conditional image synthesis with product-of-experts gans, 2021. 1

[20] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015. 4

[21] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators, 2023. 2

[22] Jeongho Kim, Gyojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on, 2023. 3

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 5

[24] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks, 2017. 3

[25] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping, 2020. 3, 6

[26] Tao Li, Gang Li, Jingjie Zheng, Purple Wang, and Yang Li. Mug: Interactive multimodal grounding on user interfaces. In *arXiv*, 2022. 2, 6

[27] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics, 2019. 3

[28] Momina Masood, Marriam Nawaz, Khalid Mahmood Malik, Ali Javed, and Aun Irtaza. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward, 2021. 3

[29] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18444–18455, 2023. 1, 2, 3, 4, 7

[30] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks, 2015. 2

[31] Gaurav Parmar, Dacheng Li, Kwonjoon Lee, and Zhuowen Tu. Dual contradistinctive generative autoencoder, 2020. 3

[32] Yogesh Patel, Sudeep Tanwar, Rajesh Gupta, Pronaya Bhattacharya, Inno Davidson, Royi Nyameko, Srinivas Aluvala, and Vrince Vimal. Deepfake generation and detection case study and challenges. *IEEE Access*, 11:143296–143323, 2023. 3

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 7

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[35] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019. 1, 3, 6, 7

[36] Jia Wen Seow, Mei Kuan Lim, Raphaël C.W. Phan, and Joseph K. Liu. A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing*, 513:351–371, 2022. 3

[37] Cuifeng Shen, Yulu Gan, Chen Chen, Xiongwei Zhu, Lele Cheng, Tingting Gao, and Jinzhi Wang. Decouple content and motion for conditional image-to-video generation, 2023. 1

[38] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019. 4, 7

[39] Simranjeet Singh, Rajneesh Sharma, and Alan F. Smeaton. Using gans to synthesise minimum training data for deepfake generation, 2020. 3

[40] Farhana Sultana, Abu Sufian, and Paramartha Dutta. Evolution of image segmentation using deep convolutional neural network: A survey. *Knowledge-Based Systems*, 201–202: 106062, 2020. 2

[41] Shubham Tandon, Aryan Vig, Murli Kartik, and Harish Chandra Kumawat. Real-time face transition using deepfake technology (gan model). In *2023 International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication, and Computational Intelligence (RAEEUCCI)*, pages 1–5, 2023. 3

[42] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 3, 6

[43] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos, 2020. 3, 6

[44] Vrizlynn L. L. Thing. Deepfake detection with deep learning: Convolutional neural networks versus transformers, 2023. 3

[45] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020. 1

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 2

[47] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018. 2, 4

[48] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models, 2023. 2

[49] Tianyi Zhang, Zheng Wang, Jing Huang, Mohiuddin Muhammad Tasnim, and Wei Shi. A survey of diffusion based image generation models: Issues and their solutions, 2023. 2

[50] Zijun Zhang, Ruixiang Zhang, Zongpeng Li, Yoshua Bengio, and Liam Paull. Perceptual generative autoencoders, 2020. 3