TOWARDS GENERATION OF HIGH RESOLUTION MULTIMODAL SYNTHESIS FOR ARBITRARY IDENTITIES IN THE WILD

A Thesis submitted for the completion of
requirements for the degree of

Bachelor of Science
(Research)

by

R Sainiranjan
Undergraduate Programme
Indian Institute of Science

Under the supervision of

Prof CV Jawahar
IIIT, Hyderabad

Prof Vinay Namboothiri
University of Bath,United Kingdom

Prof Venkatesh Babu
Department of Computational and Data Sciences
Indian Institute of Science

Prof. Vamsi Pingali
Department of Mathematics, Indian Institute of Science

Acknowledgements

I would like to acknowledge my supervisors Profs Jawahar, Vinay and Venkatesh for their hands on guidance, support and encouragement in every step of the way. Without you this work would not be possible. A big thanks are in hand to Rudrabha Mukhopadhyay from CVIT, IIIT Hyderabad for being a large part of this project. Your opinions and expertise were truly invaluable. Thanks are also in hand to Harsh, Rishubh and Tejan from the VAL Lab in IISc. Thanks guys, for being amazing sounding boards and giving some good ideas! A special thanks to Vamsi sir and the math department for being a great help! Lastly, I would like to thank my parents, my source of inspiration who stood by me every step of the way.

ABSTRACT

In this work, we tackle the **problem of high resolution multimodal synthesis for arbitrary identities**.The goal here involves a subset of related problems such as the *generation* of high resolution videos constrained through priors such as pose or motion, bringing a given video in *sync* with a provided audio segment and performing *modifications* on an existing video sample that may change attributes such as gender or age whilst preserving the core facial identity(ies) of the subject(s) present. The aim here is to work towards a solution which is as *general* as possible i.e, being invariant of the choice of data which is being attempted here through the paradigm of *Generative Modelling*. The specific task that shall be challenged here is that of *Audio Visual Sync* i.e generating videos that are in sync with an audio segment.

Index Terms— Multimodal synthesis, Generation, Generative Modelling

1. INTRODUCTION

With the tremendous growth in digital communication across the globe, we have witnessed an explosion in visual content. From video calls in platforms such as Zoom or Meet to streaming our favourite shows on Netflix, to watching and creating content on Youtube, videos have become a part and parcel of our daily lives.With the advent of the era of Deep Learning and Deep Computer Vision it has been possible to work on a wide range of solutions involving videos that have been utilized in the real world as well to improve the quality of our experiences. Naturally, understanding and enabling applications using talking face videos [1–4] has been an active area of re- search in recent years. Specifically, tasks such as speech/text- based lip synthesis [1,5] have witnessed tremendous advancements. In parallel, there has been a growth in works tackling video generation such as [6,7], movie generation and 3D understanding. However, there are several drawbacks of existing works which we list below.

1. Almost all Video related works work at low visual quality ($128^2 - 256^2$). This is problematic given the fact that improvements in internet have rendered HD streaming (Resolution $> 720^2$) to be the de-facto visual quality for videos.

2. Existing solutions are often brittle having problems such as being unable to synthesize *fine features* such as teeth [1] ,failing for identities that are not in their training data and thus have to be *retrained* to work on a new dataset.

3. Further while a method may be capable of syncing frontal videos to a given audio segment it will often fail to transfer poses appropriately and vice versa,i.e, there exists no *generality* in current methods which lowers their utility for downstream tasks which may involve the solution of a diverse set of problems.

The aim shall be to resolve these problems by providing an alternative viewpoint that allows these problems to be tackled together in a holistic fashion.To show a proof of concept, the problem of Audio Visual sync shall be addressed here.

2. BACKGROUND

The Manifold Hypothesis: Consider an arbitrary video of resolution $M \times N$. If this video has T frames then we are dealing with a set of T images $V = \{F_1, , , F_T\}$ where $F_i \in R^{M \times N}$ and $i \in N$ is a valid integral index. Let the final desired video be V^*, note that V^* might necessarily *not* have the same resolution as V which is a valid assumption in cases when the original video is itself at a low resolution say 128^2 while the desired resolution could be at say, 720^2. We do however,assume for now that $|V^*| = |V|$ i.e, they have the same number of frames. Observe then that our problem amounts to learning a mapping $\Phi : R^{M \times N} \mapsto R^{M^* \times N^*}$ that performs a *desired* transformation on each frame.

Both the domain and range of Φ are effectively too large in this case. A lot of values might yield blurry images or worse meaningless images!This becomes even more important when we consider *generalizability*. Among the many candidates for Φ that can be learned through a dataset we would like to choose those that are as *general* as possible i.e, work for images outside of the dataset used. Therefore Φ has to be constricted further, Consider the effective subsets for each resolution that encode images of desirable properties such as high visual quality, we term these as *natural image manifolds* $\mathcal{M}$, $\mathcal{M}^*$.To show the validity of our definition, we use a popular assumption in theoretical computer science and machine learning the *Manifold hypothesis* which states that real-world high-dimensional data lie on low-dimensional manifolds embedded within the high-dimensional space. Now we can constrain Φ further, not only should it perform the *desired transformation* on each individual frame but each transformed image *must* also lie on the natural image manifold.

Generative Modelling:How might we find regions of the natural image manifold $\mathcal{M}^*$ that map the input frames to the correct outputs? If we had a differentiable parameterization of the manifold, we could progress along the manifold to these regions by using a suitably constructed loss function to guide our search. In that case, images found would be guaranteed to be high resolution as they came from the high resolution image manifold, while also being correct as they would satisfy the given condition. In reality however, we do not have such convenient, perfect parameterizations of these manifolds. But, we can approximate such a parameterization

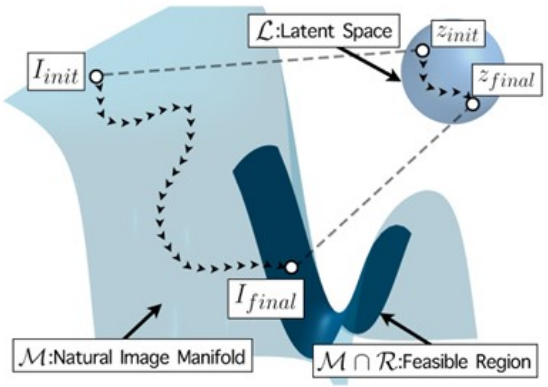

Fig. 1. Latent Space Optimization to generate candidate videos. [10]

by using techniques from unsupervised learning. In particular, much of the field of **Deep generative modeling** (e.g. VAEs, flow-based models, and GANs) amounts to creating models that map from some latent space to a given manifold of interest. By leveraging advances in generative modeling, we can even use pretrained models without the need to train our own network! This goes a long way towards solving problems *2* and *3* that we had mentioned earlier.

3. RELATED WORK

Some prior work have aimed to find vectors in the latent space of a generative model to accomplish a task; [8] for creating embeddings,[9] in the context of compressed sensing and [10] for super resolution. In this work, we focus on GANs, as recent work in this area has resulted in the highest quality image-generation among unsupervised models. Regardless of its architecture, let the generator be called G, and let the latent space be $\mathcal{W}$. Ideally, we could approximate $\mathcal{M}^*$ by the image of G, which would allow us to rephrase the problem above as the following: find a set of T latent vectors $Z \subset \mathcal{W}$ such that for a loss function L we perform the following optimization;

$$argmin_{\mathcal{W}} L(G(Z), V) \tag{1}$$

We can go further and define additional constraints that make sense for videos these include conditions such as *Identity Preservation* and *Temporal consistency* which state the fact that Identities have to be preserved throughout all frames and motion and poses have to transition naturally. Let these constraints define a different manifold $\mathcal{R}$ in the image space thus effectively we would be searching in $\mathcal{M}^* \cap \mathcal{R}$. This leads us to an important consideration; *Bringing temporal behaviour into the latent space*.

MocoGAN-HD[18], *Stitch it in Time(STIT)*[19] are some of the works that show how this can be done.MocoGAN-HD relies on training a LSTM based *Motion Generator* on top of the latent space of a pretrained StyleGAN to generate videos while *STIT* relies on *inversion* which is the problem of finding a vector in the latent space that on passage through the Generator shall yield a desired image. For reasons that will be explained in the later section, *Inversion* is a harder problem to tackle and thus we choose to base our architecture on MocoGAN-HD which we modify to suit our needs.

4. METHOD

Generation in the Latent Space We use Karras et al.'s pretrained Face StyleGAN (trained on the Flickr Face HQ Dataset, or FFHQ) [11,12,13] as our choice of Generator. Existing works such as [14,15,16]have shown the large diversity of generated images which include natural images such as animals and facial images from a large range of identities and in distinct poses. Sample images are shown in **Fig 2**. Therefore, we use the manifold learnt by the model as our image manifold. The architecture of StyleGAN (**Fig 3.**) allows for a diverse choice of Latent Spaces. This is because once a vector $z \in R^{512}$ is sampled from the *standard normal* ($\mathcal{N}(\mathbf{0}, \mathbf{I})$) distribution it is passed through a learnt mapping network *m(.)* which transforms it to a vector $w \in R^{512}$. This vector is then passed through a learnt affine transformation before being added to the convolution layer corresponding to each resolution. At each stage of this procedure we may define an appropriate latent space such as Z the standard normal, $\mathcal{W}$ the space defined by the outputs of $m(.)$, S the outputs of the affine transformations and so on. Each space has its own set of properties that make them worthy of consideration as several works show[14,15,16,17]. We chose our latent space to be $\mathcal{W}$ as it is more *expressive* than Z while being much less complex than the other candidates in the fray.

Temporal Generation of Latents We now consider the problem of bringing temporal behaviour into the latent space of a pretrained StyleGAN.We consider the MocoGAN-HD archtecture for this purpose. One advantage we now have is that due to the LSTM architecture arbitrarily many frames can be generated during inference which is not possible in inversion based methods whose complexity scale poorly with an increase in the length of videos.Some sample frames generated using MocoGAN-HD are shown in **Fig 4.** While its architecture is shown in **Fig 5.** MocoGAN-HD takes in an initial latent and at each time step t of the LSTM the hidden state $\mathbf{h_t}$ computed depends on the previous hidden state $\mathbf{h_{t-1}}$ and a randomly sampled vector ϵ_t of the same dimension which signifies the *motion transition* at that time step. To obtain $\mathbf{w_t}$ the latent vector corresponding to the t^{th} timestep we do the following.First conduct principal component analysis (PCA) on k randomly sampled latent vectors from Z passed through the mapping network $\mathcal{W}$ to get the basis $\mathbf{V}$. Then, we estimate the motion direction from the previous frame $\mathbf{w_{t-1}}$ to

Fig. 2. Images that can be generated from a pretrained StyleGAN. [14,15]

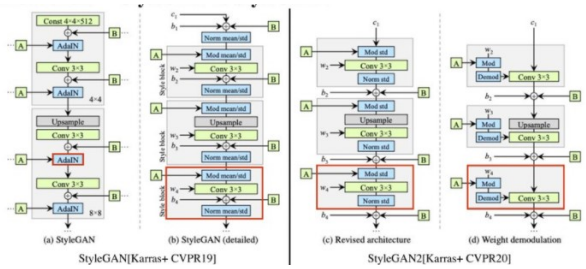

Fig. 3. Architecture of a StyleGAN. [11,12] (Zoom for clarity)

the current frame $\mathbf{w_t}$ by using $\mathbf{h_t}$ and $\mathbf{V}$ as follows:

$$\mathbf{w_t} = \mathbf{w_{t-1}} + \lambda \mathbf{h_t} \cdot \mathbf{V} \tag{2}$$

Here λ represents the step size, $\mathbf{h_t} \in [-\mathbf{1}, \mathbf{1}]$. The logic of using this update is that while motion by itself may be a complex attribute and not dependent on one single direction we may assume it to be a linear combination of some set of *basis directions* that encode arbitrarily complex concepts. Such a basis can be found by a simple dimensionality reduction approach such as a PCA alternatives to this could be using SVD or Matrix factorization.

Furthermore, two discriminators are used in training, a *content discriminator* which signifies if the video is consistent and an image discriminator which checks the valditiy of generated images. A *mutual information loss* is added to ensure that ϵ_t is used by the network at every time step t and a facial identity loss is added for identity consistency.

Fig. 4. Sample video frames generated using MocoGAN-HD[18]. Shows features such as blinking and talking movements.

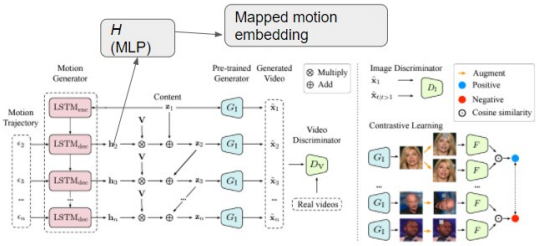

Fig. 5. Architecture of MoCoGAN-HD

The Inversion Problem The problem with using a FFHQ based StyleGAN2 is that its underlying dataset FFHQ is one of images and not of videos. One way to resolve this is to take a dataset of Videos such as *AV-Speech*[20] a dataset made of curated youtube videos and to *invert* the dataset i.e, find the latents in $\mathcal{W}$ that correspond to the images in this dataset. The issue here is that inversion is very sensitive to factors such as *pose* and *alignment* which often vary extensively in videos. Bad quality of inversion will invariably skew the results and make any optimization infeasible. We show sample results of inversion of video frames using a start of the art inversion method *Pixel2Style2Pixel(PSP)*[21] in **Fig 6**.

Fig. 6. Inversion of sample video frames using a state of the art (SOTA) method. Top row is original frames while bottom has the inverted ones.

This shows that inversion *cannot* be relied on as a means in this case.

Solving the Inversion Problem: However, it turns out that a "dataset" can still be curated for a variety of video based tasks. Consider the task of *Audio-Visual Sync* for example,

this involves syncing a given video to a given audio segment. Usually audio segments are paired with frames belonging to different timestamps from the *same video* that they belong to as additional reconstruction and identity priors can be used to act as a supervision [1]. In our case however, we can ideally dispense away with this. As we have a good generative model, quality of reconstruction is not much of an issue. The only issues here are *preservation of identity* and *syncing of video*. If we have good enough supervision for both of these problems, then we might as well work with generated images! Another advantage of using generated images is the fact that our method actually uses *little to no* supervision, what this means is that *generalizability* has been build into the model from the ground up. Assuming our method works solves the problem for generated images then it can solve the problem for arbitrary *real images* as well, i.e, images that have been obtained from the real world as long as the inversion means works, i.e, the onus is now on the inversion method and not on our method which goes a great way to add some modularity into the picture.Therefore we now use this breakthrough to construct our novel method.

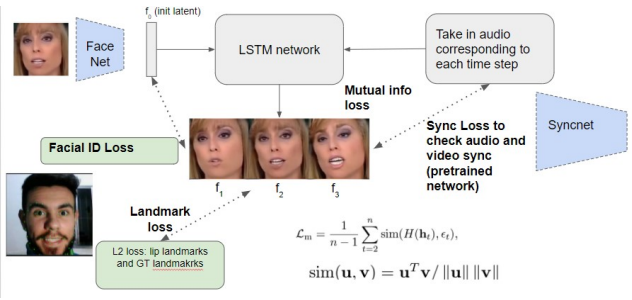

Fig. 7. Our proposed Architecture

Architecture and Losses:We model our architecture based on the MoCoGAN-HD framework. We too use a LSTM network to bring in explicit temporal dependency between the frames that will be generated. Here as well, an initial vector is taken and passed through the LSTM at each time step t however we take in an *audio embedding* a_t instead of a random vector ϵ_t. The audio embedding corresponding to a timestep is obtained from a convolution based audio encoder that may or may not be pretrained.

We use a Facial ID loss among all generated frames. This loss is computed identically as in MocoGANHD by taking cosine similarity between the embeddings of a pretrained Face Recognition network(Facenet). To guide the model to learn *Audio Visual Sync*, we use a pretrained *Syncnet Discriminator* [1], capable of distinguishing audio and video sync trained on the AV-Speech Dataset. Further, we use a *landmark loss* i.e, we take the Mean squared Error (MSE) between the landmarks of the lip region of the original and the generated videos to encourage the network to learn poses invariant of the identity. Lastly, we use a Mutual Information Loss similar to MoCoGAN-HD to ensure that the model learns to use the audio embeddings. Our proposed architecture is shown in **Fig.7** **Training Procedure:** We train on 2 Nvida Tesla V100 GPUs of 11GB each. The total number of iters set is around 50k images. To construct our dataset we randomly sample 10k vectors from the 512 dimensional standard gaussian and pass it through the mapping network to obtain our latent datasets. We concurrently use the AV-Speech dataset and set both datasets to have a common batch size of 1 i.e, 1 video per latent. We train our LSTM on 8 frames. The weights for the losses are 0.1 for identity loss, 1 each for the syncnet, landmark and mutual information loss.

5. INFERENCE AND RESULTS

During inference time we simply unroll our trained LSTM to the desired number of frames of the given audio given an initial latent. **We do not use the Ground Truth videos only the audio is used by our model**.**Fig 8** shows the results at inference time from a video in the dataset at the default resolution of the dataset (128^2) **Fig 9.** shows the results at a higher resolution on a new HD youtube video that does not belong to the dataset. Closeup of the lips and the speech represented by the frames have been added for this case.

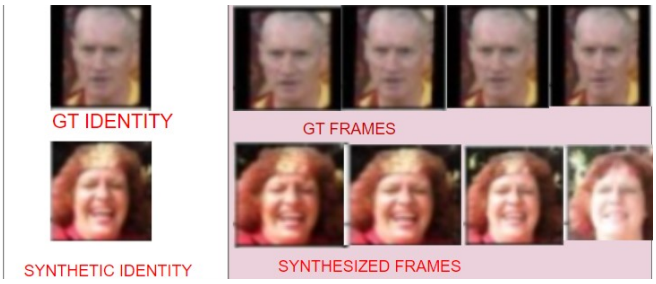

Fig. 8. Results of Audio Visual Sync on a Generated Video

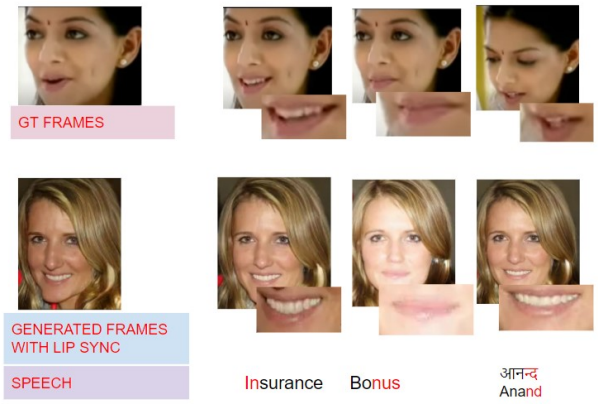

Fig. 9. Generated Lip shapes and GT Lip shape matches for HD Video

As can be seen while sync is present in both cases, it is being accompanied by jitters in identity. The jitters are

higher for the case when the video is not from the dataset. This shows that while the method is promising it is not stable enough to work for a large number of frames. This could be caused due to the fact that our chief source of getting the sync right, the Syncnet network is trained on the AVSpeech dataset while the images being generated here have properties similar to the FFHQ dataset.To improve this further work is being done in this direction.

Bibliography

[1] Prajwal K R et al. "A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild". In: *Proceedings of the 28th ACM International Conference on Multimedia*. MM '20. New York, NY, USA: ACM, 2020.

[2] Prajwal K R et al. "Towards Automatic Face-to-Face Translation". In: *Proceedings of the 27th ACM International Conference on Multimedia*. MM '19. Nice, France: ACM, 2019. ISBN: 978-1-4503-6889-6. DOI: 10.1145/3343031.3351066. URL: http://doi.acm.org/10.1145/3343031.3351066.

[3] Rithesh Kumar et al. "ObamaNet: Photo-realistic lip-sync from text". In: *ArXiv* abs/1801.01442 (2018).

[4] Ariel Ephrat et al. "Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation". In: 37.4 (2018). DOI: 10.1145/3197517.3201357.

[5] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. "The Conversation: Deep Audio-Visual Speech Enhancement". In: *ArXiv* abs/1804.04121 (2018).

[6] Aliaksandr Siarohin et al. "First Order Motion Model for Image Animation". In: *Conference on Neural Information Processing Systems (NeurIPS)*. Dec. 2019.

[7] Yurui Ren et al. *PIRenderer: Controllable Portrait Image Generation via Semantic Neural Rendering*. 2021. arXiv: 2109.08379 [cs.CV].

[8] Yipeng Qin Rameen Abdal and Peter Wonka. "Image2StyleGAN: How to embed images into the StyleGAN latent space?" In: *Proceedings of the International Conference on Computer Vision (ICCV)*. 2019.

[9] Ajil Jalal Ashish Bora, Eric Price, and Alexandros G. Dimakis. "Compressed Sensing using Generative Models". In: *International Conference on Machine Learning (ICML)*. 2017.

[10] Alexandru Damian Sachit Menon et al. "PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models". In: *Proceedings of the Conference on Computer Vision and Pattern Recognition*. 2020.

[11] Samuli Laine Tero Karras and Timo Aila. "A style based generator architecture for generative adversarial networks". In: *Proceedings of the Conference on Computer Vision and Pattern Recognition*. 2019.

[12] Tero Karras et al. "Analyzing and Improving the Image Quality of StyleGAN". In: *Proc. CVPR*. 2020.

[13] Tero Karras et al. "Training Generative Adversarial Networks with Limited Data". In: *Proc. NeurIPS*. 2020.

[14] Yipeng Qin Rameen Abdal and Peter Wonka. "Image2StyleGAN++: How to Edit the Embedded Images?" In: *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[15] Jonas Wulff Lucy Chai and Philip Isola. "Using latent space regression to analyze and leverage compositionality in GANs". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2021.

[16] Jonas Wulff and Antonio Torallba. "Improving Inversion and Generation Diversity in StyleGAN using a Gaussianized Latent Space". In: *Proc.Neurips*. 2020.

[17] Aaron Hertzmann Erik Härkönen, Jaakko Lehtinen, and Sylvain Paris. "GANSpace: Discovering Interpretable GAN Controls". In: *Proc.Neurips*. 2020.

[18] Yu Tian et al. "A Good Image Generator Is What You Need for High-Resolution Video Synthesis". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=6puCSjH3hwA.

[19] Rotem Tzaban et al. *Stitch it in Time: GAN-Based Facial Editing of Real Videos*. 2022. arXiv: 2201.08361 [cs.CV].

[20] A. Ephrat et al. "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation". In: *arXiv preprint arXiv:1804.03619* (2018).

[21] Elad Richardson et al. "Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021.