Theorem

Towards Generation of High resolution Multimodal Synthesis for arbitrary identities in the Wild

Sai Niranjan Ramachandran (BS Mathematics)

Advisors: **Dr CV Jawahar**, IIT, Hyderabad
Dr Vinay Namboothiri , University of Bath, United Kingdom
Dr Venkatesh Babu,Department of Computational and Data Sciences (CDS), Indian Institute of Science
Dr Vamsi Pingali ,Department of Mathematics,Indian Institute of Science

Presentation Outline

1. Introduction
2. Background
3. Method
4. Results
5. Unresolved problems and Scope for Future Work
6. Thanks and Appendix

Introduction

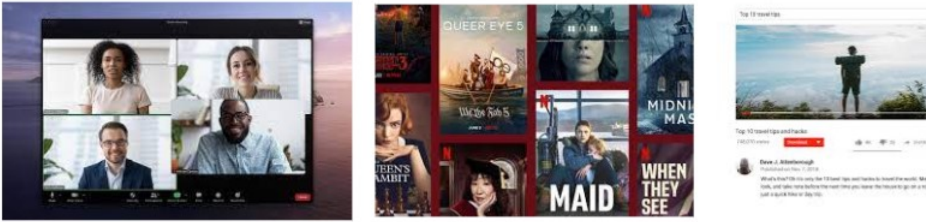


Figure 1: Netflix, Zoom and Youtube are increasingly common in our lives

Introduction

- AI and Deep learning has accelerated this trend
- Tasks such as generation of talking face videos, speech/text based lip synthesis and movie generation have witnessed a tremendous growth.

Figure 2: The actor Mark Hamill being de-aged using AI for the part of Luke Skywalker in the Mandalorian series.

Some more Examples

Figure 3: A SOTA model (FOMM) used to animate a "source image" as per a "driving video"

Figure 4: An Editing framework (STIT) editing a video featuring Kamala Harris

Current Problems with existing methods

1. Existing works at low quality $128x128 - 256x256$. HD quality is common nowadays.
2. Unable to synthesize fine features (teeth) or fail for unseen data.
3. Multiple tasks cannot be solved by a single method. Crucial for real world applications

What is the aim?

1. Though these problems *appear to be very different* we can utilize an alternate geometry inspired viewpoint to solve them.
2. Proof of concept using **Audio Visual Sync** task i.e, *generating videos that are in sync with an audio segment*. Has a wide range of real world applications; dubbing in the movie industry to real time video translation and motion capture.

Problem Construction

1. Consider a video V of resolution $M \times N$. If this video has T frames then $V = \{F_1, ..F_T\}$ where each F_i is an image of the same resolution.
2. Each image is a tensor from $\mathbb{R}^{ch \times M \times N}$ (ch =1 for grayscale,3 for RGB)
3. Given set of conditions C learn $\mathbf{\Phi_C} : \mathbb{R}^{\mathbf{T \times ch \times M \times N}} \mapsto \mathbb{R}^{\mathbf{T \times ch \times M' \times N'}} \ni \mathbf{\Phi_C(V)} = \mathbf{V}'$ for desirable V'.
4. Frames are *correlated*; we can utilize this.
5. Instead of single $\mathbf{\Phi_c}$ learn T maps $\{\phi_{k,C}\}$; easier, more general!

Manifold Hypothesis

1. Having a broad domain and range for ϕ can give undesirable results such as *blurred images* or worse, nonsensical images! (Figure)
2. We need to constrict the domain and range further!

Figure 5: Some results encountered when solving the problem in an unconstrained high dimensional space

Manifold Hypothesis

- Assume $\exists$ a manifold $\mathcal{M}$ termed as the **Natural Image Manifold** that contains all possible desirable images.
- The existence of $\mathcal{M}$ follows from the **Manifold Hypothesis** that states that real world data are often found in low dimensional manifolds in a high dimensional space.

But where is the Natural Image Manifold $\mathcal{M}$?

- As long as $\mathcal{M}$ is known, all we need is a *differentiable parameterization* of $\mathcal{M}$ and then we can simply *progress* along those regions that map our input frames to correct outputs using a suitable loss function.
- **But do we have the differentiable paramterization?**
- **Answer : NO!**
- To solve this use **Generative Modelling**!

The power of Generative Modelling

- Learn a mapping between a known lower dimensional distribution (such as isotropic normal distribution $\mathcal{N}(\mathbf{0},\mathbf{I})$) and unknown higher dimensional distribution of interest.
- Use the *properties* of the known distribution (**latent space**) to manipulate the higher dimensional distribution.
- Let $\mathcal{G}$ (the Generative Space) be the Range of the model. For a *good generative model* $\mathcal{G}$ would be *dense* or *very nearly dense* in $\mathcal{M}$

Finalizing the Problem Statement

Let $\exists$ a Generative Model $G : \mathcal{Z} \mapsto \mathbb{R}^{ch \times M' \times N'}$ where $\mathcal{Z}$ is a suitable latent space such as the isotropic multivariate normal. Given a video V of resolution $M \times N$ with T frames, let there be a suitable loss function L and let $Z = \{z_1, ...z_T\}$ be an *ordered* set of T elements each drawn from $\mathcal{Z}$. Then we have the following minimization problem :

$$argmin_{Z \in \mathcal{Z}^T} L(G(Z), V) \tag{1}$$

where $G(Z) = \{G(z_1), ..G(z_T)\}$ represents a "reconstructed video" V'.

Choosing the Generative Model

- Choose GAN which has highest quality of generated images.
- GAN has Generator (G), Discriminator (D) trained in an *adverserial* manner
- G generates *realistic* data
- D distinguishes real data and G's data.
- Training is a *zero sum* game

$$min_G max_D L(G, D) = \mathbb{E}_{x \sim p_{data}}[log(D(x)] + \mathbb{E}_{z \sim \mathcal{Z}}[log(1 - D(G(z)))] \quad (2)$$

StyleGAN

- Choose Karras et.al's pretrained Face StyleGAN trained on the FFHQ Dataset.
- This can gnerate large diversity of generated images including diverse faces in various poses and natural images such as animals, cars etc.

Figure 6: The diverse set of images capable of being generated through a StyleGAN

StyleGAN Architecture

Architecture gives us *flexibility* in choice of Latent Space as well!

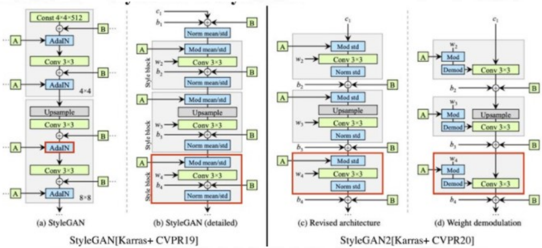


Figure 7: Architecture of a StyleGAN

StyleGAN Architecture

The algorithm of the generator is as follows:

1. $z \sim \mathcal{N}(\mathbf{0},\mathbf{I}), z \in \mathbb{R}^{512}$
2. Get $w = m(z), w \in \mathbb{R}^{512}.m(.)$ is a learnt network termed as the **mapping network**
3. Pass w through learnt affine transformations $A_1, , , A_n$
4. Add to convolution layer corresponding to successive resolution $(8x8 - 1024x1024)$
5. Obtain HD Image

StyleGAN Architecture

Therefore we get multiple choices of latent spaces.

- $\mathcal{Z} = \mathcal{N}(\mathbf{0},\mathbf{I})$
- $\mathcal{W} = m(\mathcal{Z})$
- $\mathcal{S} = [A_1(\mathcal{W}), ..A_n(\mathcal{W})]$

We choose $\mathcal{W}$ which is more expressive than $\mathcal{Z}$ while being less complex than others.

How to traverse the Latent Space?

We need to perform curve traversal in our chosen latent space.There are broadly two kinds of methods employed for this purpose:

1. **Optimization** - Involves optimization for a single image or a set of images using latents initialized in the latent space. Unstable for larger sets
2. **Geometry Based** - Traversal along directions corresponding to a *local* or *global basis* of directions in the latent space.Stable for larger sets. We shall use this

Geometry Aware Traversal

Two kinds of curve traversal methods.

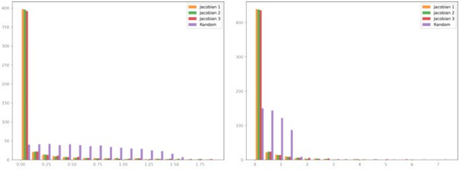
1. **Local Basis**: Define for every $w \in \mathcal{W}$ separately as the basis of the Tangent Space at w ($\mathcal{T}_w$). As $m(.)$ is a black-box neural network, to find this we need $z \in \mathcal{Z} \ni m(z) = w$. Then dm_z, (**pushforward**), the jacobian of m at z can be computed.

$$dm_z : \mathcal{T}_{\mathbf{z}} \mapsto \mathcal{T}_{\mathbf{w}} \tag{3}$$

To find the basis perform a dimensional reduction such as PCA/SVD on dm_z (**Manifold Hypothesis**)

Geometry Aware Traversal

Validity of the Manifold Hypothesis in this case been seen through Fig 7. which shows the distribution of singular values for 3 random z, a very significant chunk ($>$ 400 out of 512) is 0 or slightly larger)



(a) Singluar values $\sigma_i^{\mathbf{z}}$ with $0 \leq \sigma_i^{\mathbf{z}} \leq 2$

(b) All Singluar values $\sigma_i^{\mathbf{z}}$

Figure 8: Distribution of singular values of the jacobian of the mapping network

Local Basis method in practice

- Very stable traversal method that has the advantages of preserving facial identity and smooth transitions
- It needs to know the z at every stage that maps to an intermediate w which greatly reduces it viability.

Geometry Aware Traversal - Global Basis

2. **Global Traversal**: Assume $\exists$ *global basis* of directions for the manifold $\mathcal{W}$; algorithm to find this:
    1. Sample N vectors $z_{1:N}$ from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ where N is necessarily very large
    2. Get $w_{1:N}$ corresponding to these from the mapping network
    3. Perform a dimensionality reduction algorithm such as PCA or SVD
    4. Take the top k components to be the basis V
    5. Traversal is now given by the equation

$$w^{'} = w + Vx \tag{4}$$

where x is a control vector used to control the extent of traversal along the basis directions.

Geometry Aware Traversal - Global Basis

- Global Traversal is less accurate but *more general*
- x can be learnt as $f(w)$ depending on task
- Use this method based on video generation framework - **MocoGANHD**

MocoGANHD

- MocoGANHD (Tian et.al , *ICLR 2021* generates synthetic videos on top of a pretrained StyleGAN
- Given $w \in \mathcal{W}$ returns *ordered set* of latents, $W_T = \{w_1, ...w_T\}$ (**Video with same identity**)
- Learns control vector per frame for global traversal
- Learns *temporal consistency* between frames
- Learns *realistic motions* i.e head pose and expressions

MocoGANHD-Architectural Innovations

- Utilizes a conditional **LSTM** a sequential architecture defined for an ordered series of time steps $1 : T$ (during inference can work for arbitrary T)
- Output of network at time t; $x_t = f_t(x_{t-1}, \epsilon_t)$; ϵ_t is random vector corr. to motion at t, f_t is learnable
- t^{th} frame of video obtained as:

$$w_t = w + Vx_t \tag{5}$$

MocoGANHD

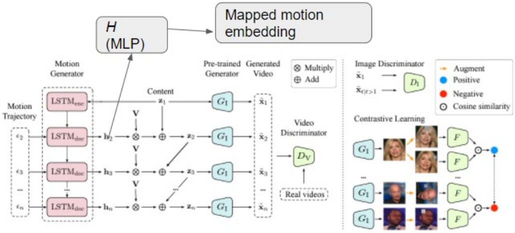


Figure 9: Architecture of MocoGANHD

Figure 10: Sample videos generated by MocoGANHD

Our innovations for the Audio Sync Setting

Generalizing MocoGANHD!

1. Substitute Audio conditioning in-place of motion
2. Starting w needs to be known always. Finding w for arbitrary image is *inversion*. Figure shows how hard problem is (SOTA). **This problem is resolved as well**.

Figure 11: Inversion of sample Video Frames

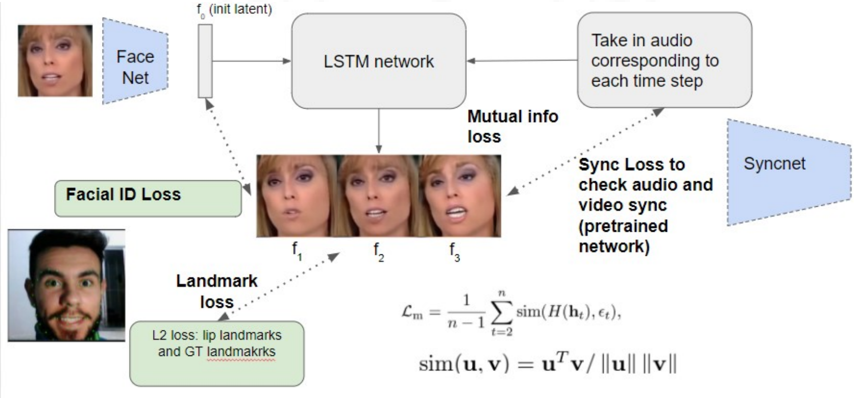
Solving the Inversion Problem

- Inversion is in fact a *separate* problem.
- For audio-sync by default datasets pair audio segements with frames belonging to different timestamps from the same video
- This can give a reconstruction prior for supervision
- **Not needed here** as generator can synthesize high quality images
- Therefore we can work with generated images!

Generality is built from the ground up!

- **Our method uses no ground truth image supervision**
- If our method works for generated images we can solve for arbitrary real images!
- The onus is now on the inversion method - **modularity attained!**

Our Architecture



$$\mathcal{L}_{\mathrm{m}} = \frac{1}{n-1}\sum_{t=2}^{n} \mathrm{sim}(H(\mathbf{h}_t), \epsilon_t),$$

$$\mathrm{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^T\mathbf{v} / \|\mathbf{u}\| \, \|\mathbf{v}\|$$

Figure 12: Our architecture

Results

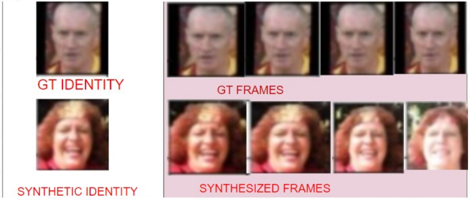


Figure 13: Expression transfer at low resolution 128×128

Shows expression transfer from a given identity to a synthetic identity made from randomly sampling in the latent space. Done in low resolution to show the consistency.

Results

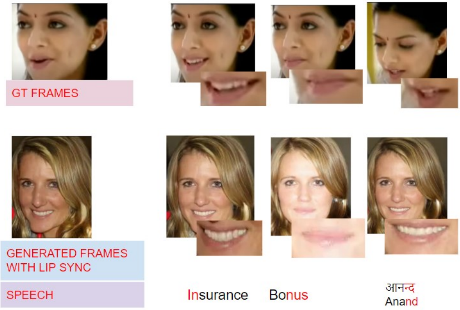


Figure 14: Generated Lip shapes and GT Lip shape matches for HD Video

Performing audio video sync on a video **unseen** to the model.Lip shapes match very well.

Unresolved problems

Problems yet to be resolved:

1. *Identity Jitters* : Unstable with respect to facial identity for large T.
2. *Motion Consistency* : Audio conditioning seems to be **insufficient** for motion.Explicit motion priors may be needed
3. *Improving Traversal*: More accuracy needed.Can utilize recent works utilizing Differential Geometry such as Riemannian CNFs (Mathieu et.al,Neurips 2020), Moserflow (Best paper, Neurips 2021)
4. *Inversion*: Open problem. Theoretically feasible for a large class of images but real world solutions are still limited.

Scope for Future Work

- Our method shows feasibility of approach; quality and resolution problems already solved!
- Generality of methods allows tackling of a wide range of Audio-Visual problems
- Possible to work towards a *holistic* solution

THANKS!

Local Basis Validity

Theorem 6.1

Jacobian dm_z is sufficient to give local basis of $\mathcal{T}_\mathbf{w}$

Proof.

Let the common dimension of $\mathcal{Z} = \mathcal{W} = n$ (For StyleGAN $n = 512$). We know, $\mathcal{Z}$ is $\mathcal{N}(\mathbf{0},\mathbf{I})$ in this dimension.
It is known further that in higher dimensions, due to *concentration of volume*, sampling $z \in \mathcal{Z}$ is equivalent to sampling $z \sim Uni(\sqrt{n-1}S^{n-1})$ i.e, in effect our vector comes from the surface of an hypersphere. Now, $\mathcal{T}_\mathbf{z}$ is therefore the plane passing through z and normal to the hypersphere $\sqrt{n-1}S^n$ at z whose $n-1$ dim basis in $\mathbb{R}^n$ can be found easily (just the $n-1$ dim standard euclidean basis with -1 appended to last dim).

We know that $dm_z : \mathcal{T}_\mathbf{z} \mapsto \mathcal{T}_\mathbf{w}$ is a **linear transformation**. Therefore, the basis of $\mathcal{T}_\mathbf{w}$ is mapped by the basis of $\mathcal{T}_\mathbf{z}$ and the proof follows.