Project Description To Decide

Antoine Buisson
Graduate Student
Department of Computer Science
Seoul National University
Télécom SudParis
Email: antoine.buisson@telecom-sudparis.eu

Mads Emil Marker Jungersen
Undergraduate Student
Department of **********
Seoul National University
****
Email: madsjungersen@snu.ac.kr

Steve Suard
Graduate Student
Department of Computer Science
Seoul National University
Télécom SudParis
Email: suard_st@snu.ac.kr

Jonathan ke Eun Woo kesson
Graduate Student
Department of **********
Seoul National University
****
Email: j.akesson99@gmail.com

This paper proposes a low bandwidth talking-head video synthesis model (**MODEL NAME**) for video conferencing. **MODEL NAME** builds on the previous paper "One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing" (face-vid2vid). We decided to work on the Source Image. The main idea is to send through the internet a low resolution resized version of the source image and upscale it at the other end with a super resolution model from the paper "Enhanced Deep Residual Networks for Single Image Super-Resolution" (EDSR). The goal is to either to reduce the amount of bandwidth needed, or be able to send more Source Images without impacting the bandwidth.

1 Introduction

Due to the COVID-19 Pandemic and the rise of new technology, videoconference is becoming a standard in every industry and in education. As Humans, we have a need for facial communication, this explain why it is now a standard to share your face through your webcam during such videoconferences. But this lead to two major problems. First of all, video data is the data type that use the internet bandwidth the most. Secondly, some workers and student may not have the best internet connection and will sometime experience connection drops when many people are sharing their webcams. That is why we should find a way to optimize the amount of data needed to share your face for videoconferences.

One of the proposed methods is the Talking-Head Video Synthesis. The idea is to take a high quality source image. Only use the driving video from the webcam to extract a number of facial feature. And finnaly reconstruct the video by morphing the features on the source image. This work extremely great to reduce the bandwidth. But reconstructing a whole video from a single image can lead to erroned results when the driving position from the video is becoming too different from the original Source Image position.

Another proposed method is the use of Super resolution models. The idea would be to only send a low resolution driving video through the internet, and upscale it back to its original resolution before being displayed. The main issue with this is that Super resolution models are usually pretty slow for good looking results and that it is sometimes hard to keep a temporal stability between each frame, which lead to a lot of flickering.

Having look at these two solutions, our idea is to use the best of both worlds by adding a super resolution block in the Talking-Head Video Synthesis Model. This could lead to more Sources Images being sent and thus, bring a solution to the previous erroned results.

2 Litterature Survey

2.1 One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing

This is the main building block for our solution. This paper was written by Ting-Chun Wang, Arun Mallya and Ming-Yu Liu from NVIDIA Corporation.

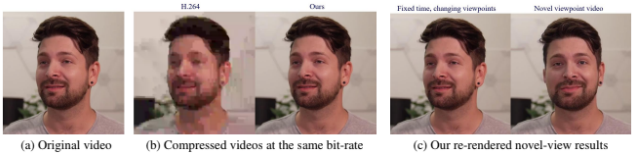


Fig. 1. I'm the figure 2 from part 6-1

Their methods allowed similar video results with H264 but for only one tenth of the bitrate. This paper is also able to re-render novel-view, meaning it can change the head position and rotation if needed. Finnaly, all those functions allow the model to do deepfake by simply choosing 2 different persons for the Source Image and Driving Image.

Here is the link to their paper : https://nvlabs.github.io/face-vid2vid/

2.2 Enhanced Deep Residual Networks for Single Image Super-Resolution

Part 2-2 ...

3 Data Pipeline

Part 3 ...

4 Our Solution[1]

Part 4 ...

5 Implemented Baselines

5.1 Talking-Head Video Synthesis Model

5.1.1 baseline 1******

5.1.2 baseline 2******

5.2 Single Image Super Resolution Model

6 Part 6

6.1 Part 6-1

Part 6-1 withFigure ‘2

7 Part 7

Part 7 with Table ??

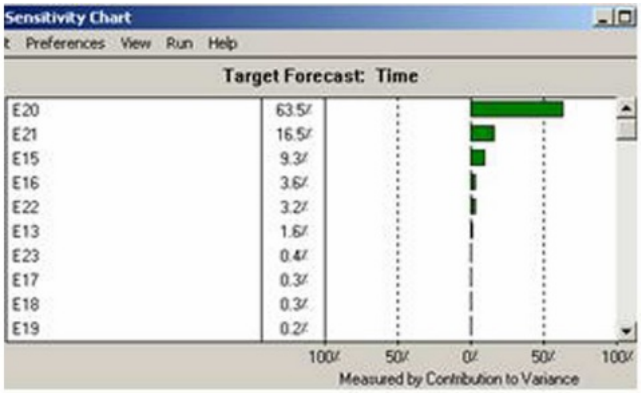
8 Part 8

Text Citation. This is a text citation.

Part 8 ...

9 Part 9

Part 9 ...



(a) Time

(b) Cost

Fig. 2. I'm the figure 2 from part 6-1

10 Part 10

Part 10 ...

1. Item 1
2. Item 2

...

Acknowledgements

Acknwledgments ...

Appendix A: Head of First Appendix

Appendix A

Appendix B: Head of Second Appendix

Subsection head in appendix

Appendix B with Eq. (??)

$$a = b + c. \tag{1}$$

[1]Still need all the training and merging of the two solution