Centre For Innovation

DeepZoom

Video Conferencing Reimagined

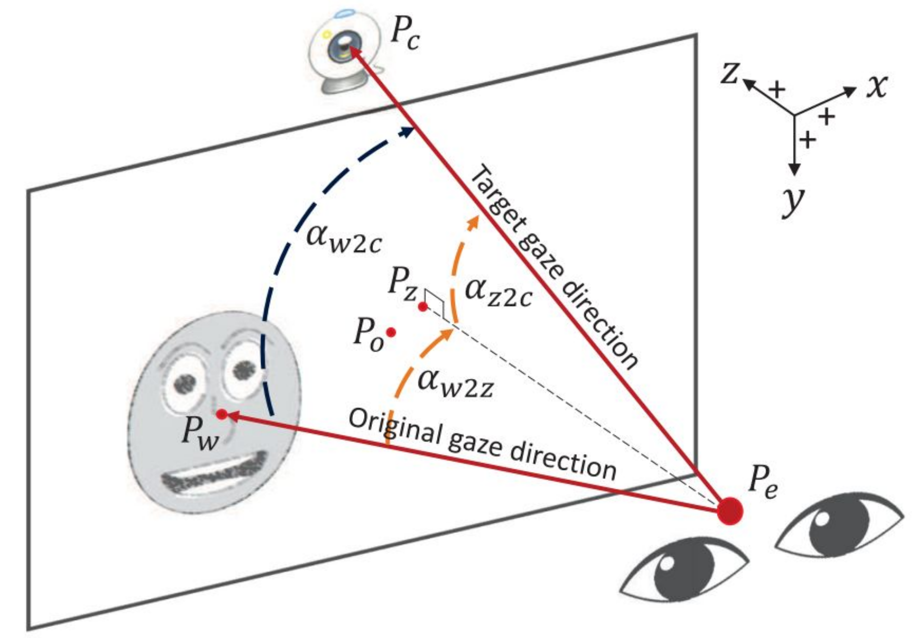
DeepZoom - An Intelligent VideoCon Solution

- In the recent times, due to the pandemic video conferencing tools have become an integral part of our lives.
- We feel that the current solutions in the market don’t solve most of the problems faced by the users and require high bandwidth to work.
- We reimagine video conferencing tools, with innovative features using the latest advancements in Deep Learning.
- DeepZoom’s vision is to go beyond the usual communication space, giving the users a rich experience as close to the real world as possible.

Gaze Redirection

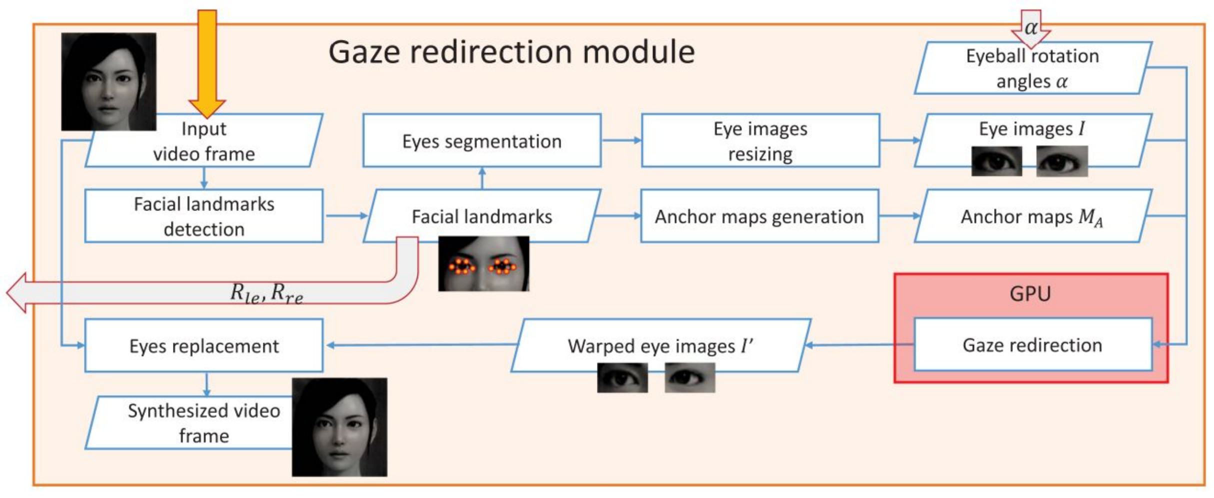
What it does?

- Changes the original gaze direction of the eyes towards the position of the camera.



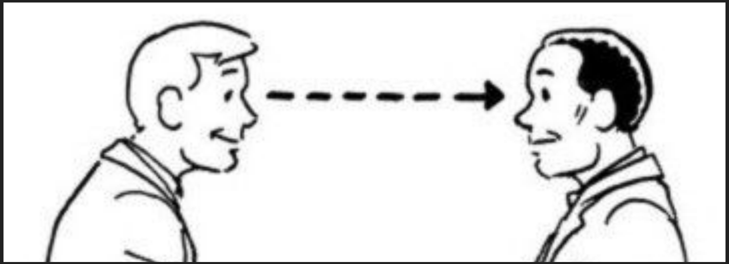
Implementation

- Eyeball rotation angles are dynamically estimated based on the positions of the camera and local and remote participants' eyes.
- Then, a warping-based convolutional neural network is used to relocate pixels for redirecting eye gaze.



Impact

- **Personal Connectivity :** Creates an eye-to-eye contact which gives social, emotional, and haptic feedback to the speaker as well as listener.
- **Better Reach among target audience :** Increases the attentivity and alertness of the listener
- Instills motivation to the speaker as well.



Demo:

Gaze_Redirection_Demo

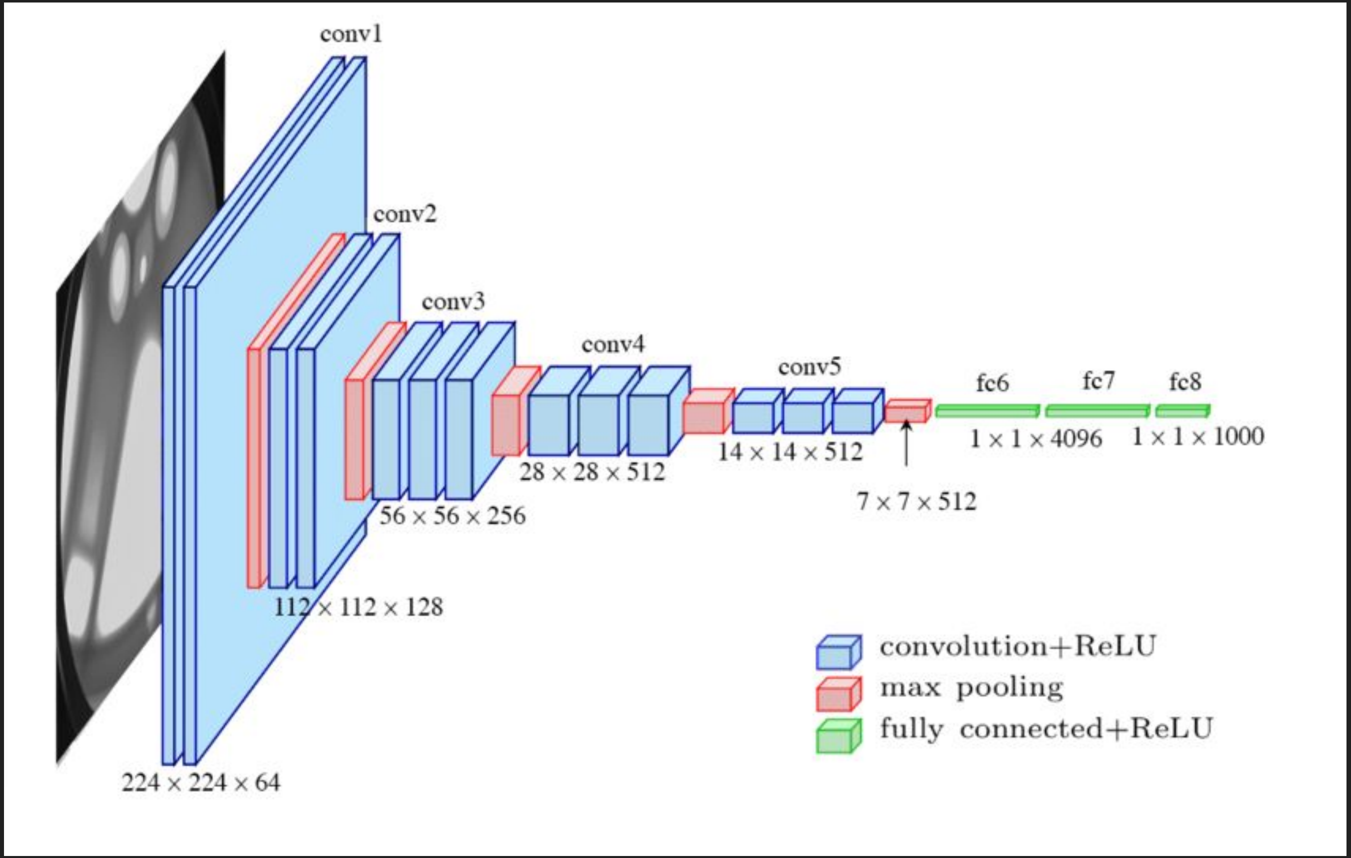
Facial Detection & Recognition

What it does?

- **Face Detection** : Detect the human faces inside the frame of the video
- **Face Recognition :** Finds a known face in the video and correctly recognizes all the known faces.
- **Face Verification :** Verifies the face pairs as same person or different persons.
- **Facial Attributal Analysis :** Analyzes the entire facial attributes of the face that's detected and gives out the following metrics :
  - **Estimated Age**
  - **Facial Expression (Like Angry, Fear, Neutral, Sad, disgust and happy)**
  - **Gender**

How it does ?

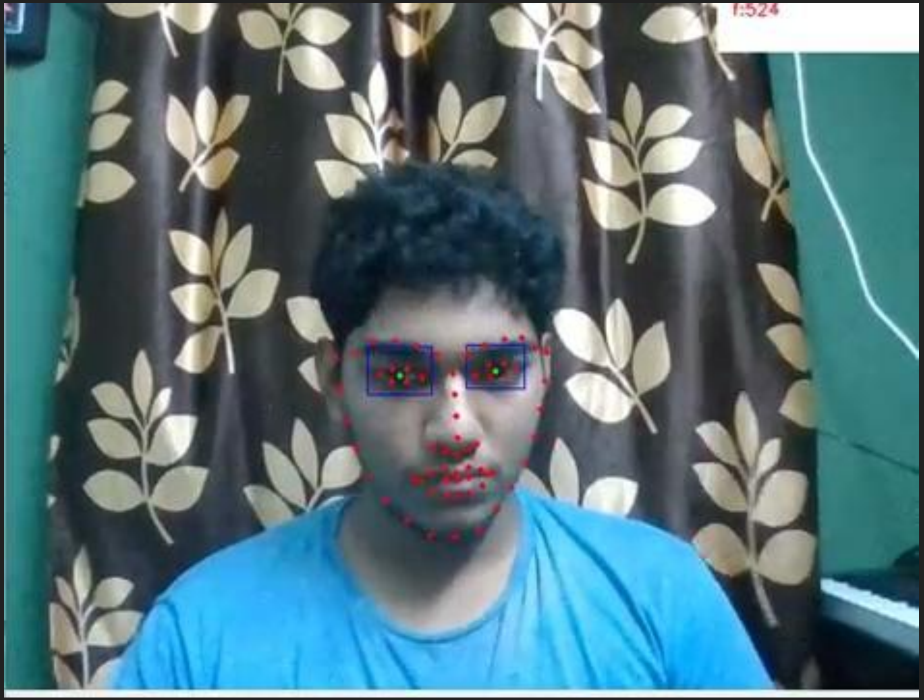
- The Face detection algorithm is based on a convolutional neural network which is pre-trained on a huge set of human faces.
- The hidden layers learn the various features of the human face and is able to detect it irrespective of the spatial location of the face in the frame.
- The rationale behind the emotion recognition algorithm lies on the universality of facial expression and body language.
- Going a step further, by performing the facial recognition multiple times, The model can output the closeness of two faces and thus enabling us to verify identity of the person in the frame.



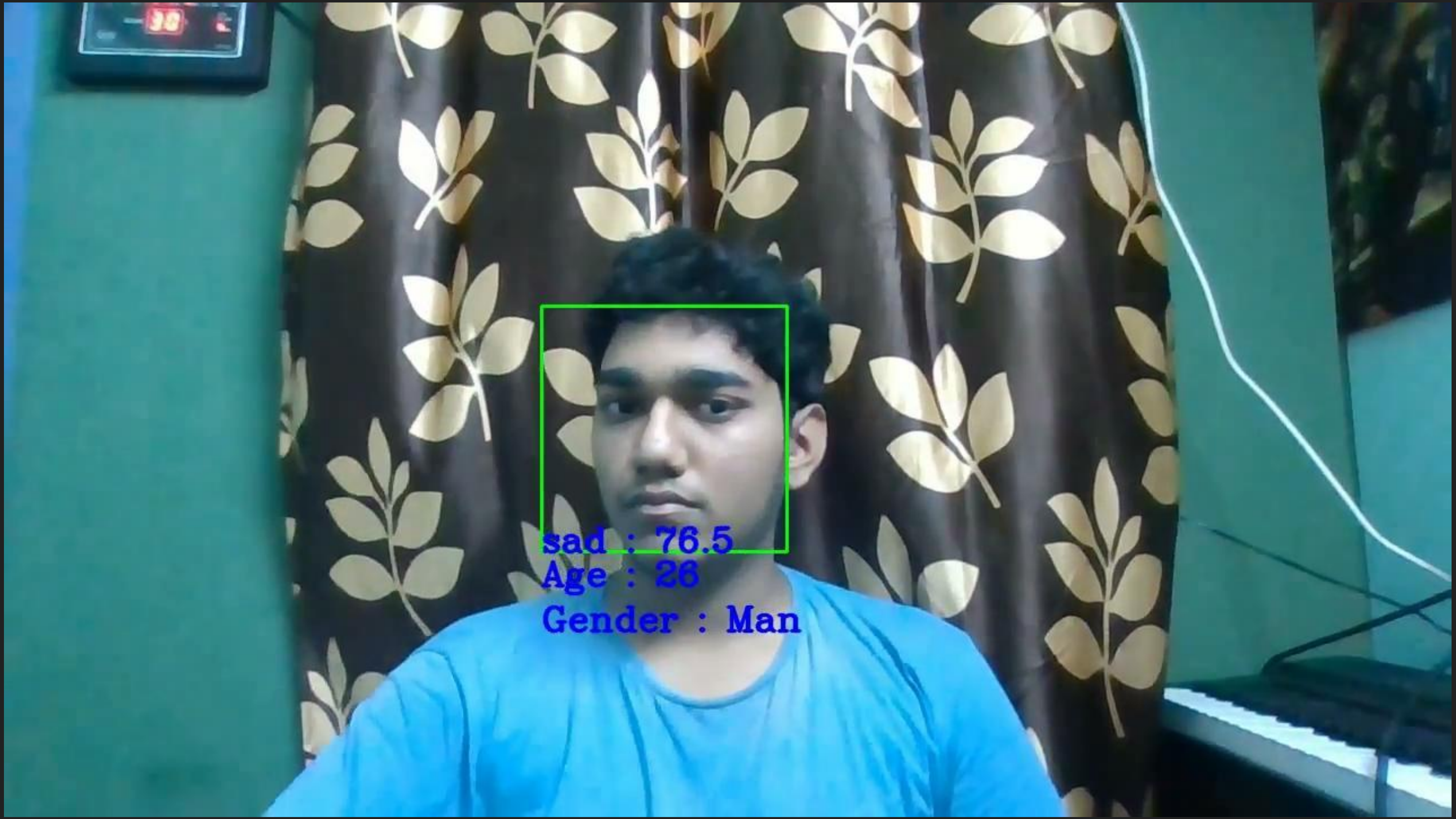
The VGGNet Architecture

The Impact of what it does?

- **SOCIAL INTERACTION MONITORING :** Our DeepZoom monitors the emotions of user throughout their usage and gives an analysis on the social interactions of the person.
- **BETTER ATTENTION TRACKING :** The tool implements facial verification and detection features to track the attention of audience members, which could prove highly useful to schools and similar organizations.
- **DEEPER INSIGHT INTO AUDIENCE :** Using our Facial Attributal analysis we gather data of gender, estimated age and the emotional response of the audience throughout the event, which are vital metrics for the organizer.



Detecting Facial Features (demo)



Emotion Recognition (demo)

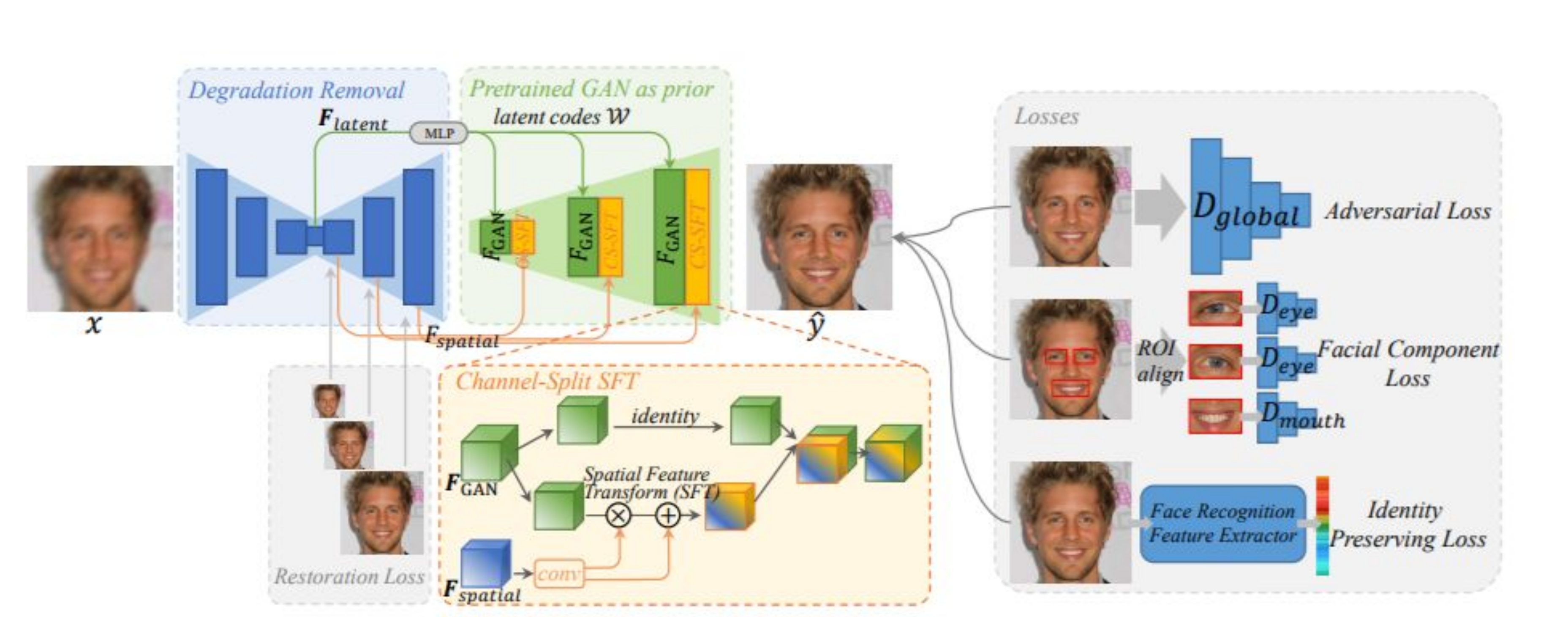
Super-Resolution

What it does?

- The image super-resolution is the process of recovering high-resolution images from low-resolution images.
- Taking in a low-resolution image from one side, We place our attention on the face and proceed to restore realistic and faithful details.
- Can be used to enhance image quality sent over low bandwidth networks.

Implementation

- We take help of Generative adversarial networks incorporating a Generative facial prior (GFP).
- Here the Blind face restoration framework tries to take a facial image suffering for unknown degradation and aims to estimate a high quality image as similar as possible to the ground truth image.
- We also try to minimize the identity preserving loss in order to get an image as close to the ground truth as possible



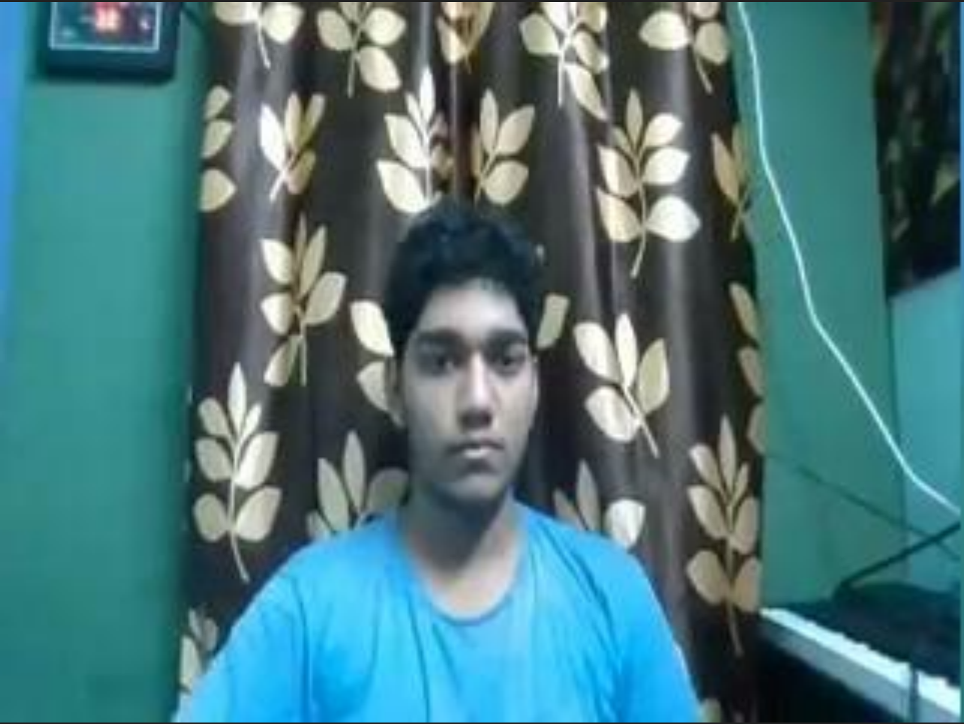
GFP-GAN Architecture

Impact of what it does?

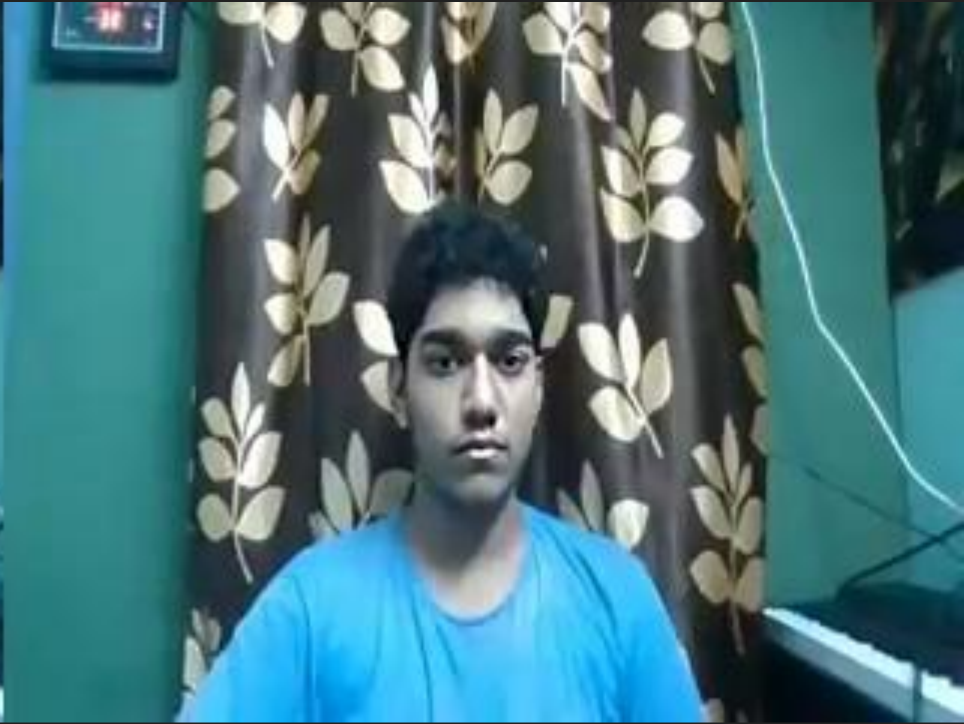
- **HIGH QUALITY EXPERIENCE :** Super resolution technique enables the users to experience a high quality video in spite of the other having a poor internet connection.

VS

Demo super_res_demo:



Input

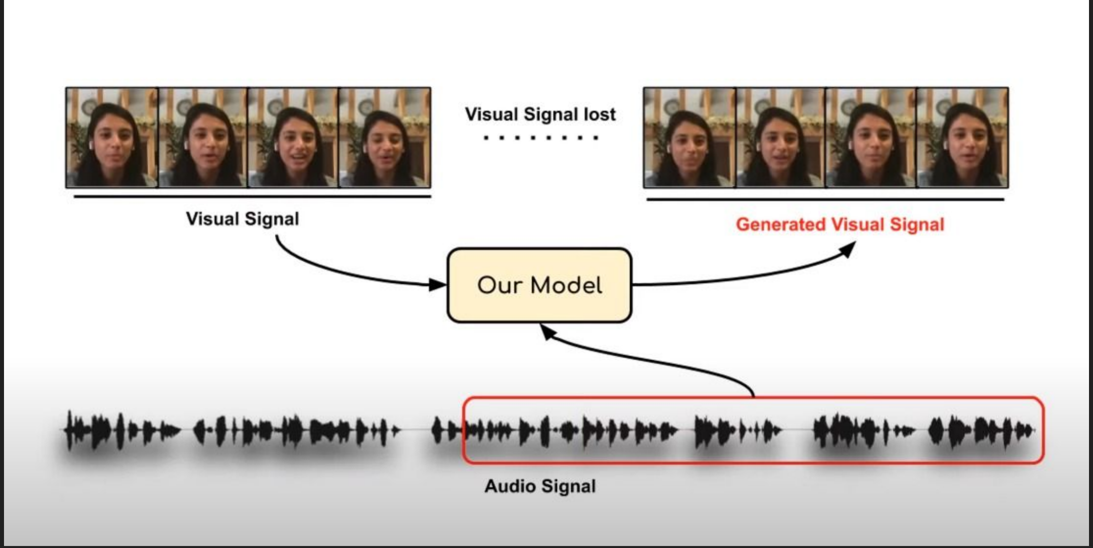


Output - More clear face

Lip-sync

What it does?

- Synchronizing lip movements with audio and also reconstructing missed part of the video with existing audio and previous frames

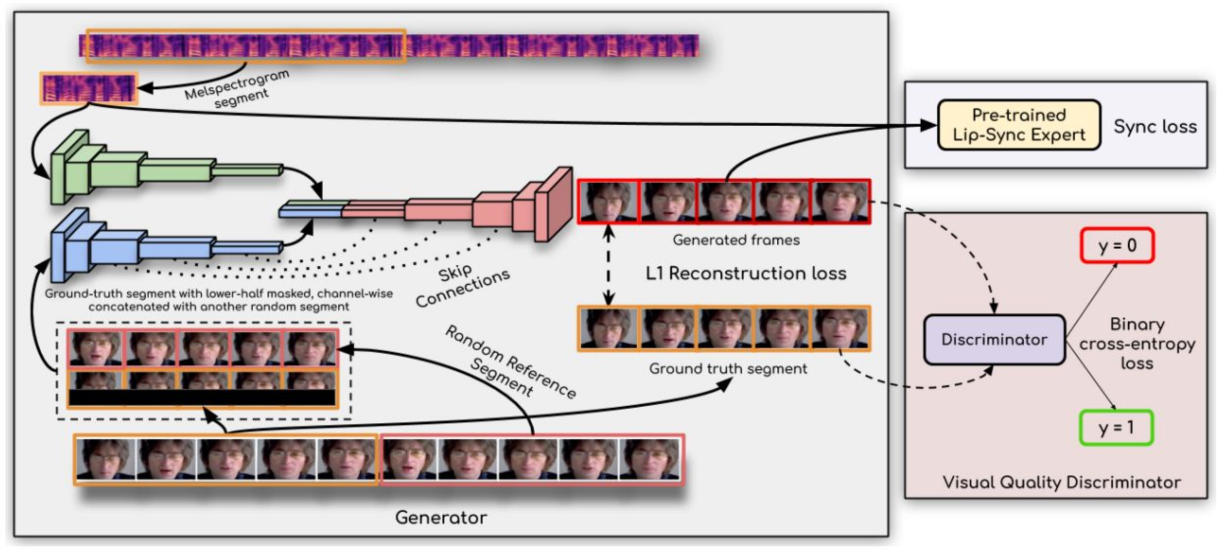


Impact of what it does?

- **Realistic Experience:** Video with correct lip sync will provide a realistic feeling of conversation.
- **Easier understanding:** the listener will find lip synced videos more understandable otherwise it can be annoying.
- **Lower Bandwidth:** Only audio can be transmitted and can be lip synced with a digital avatar which would save bandwidth.

Implementation

- Generator generates accurate lip-sync by learning from an “already well-trained lip-sync expert" which is the Discriminator.


Melspectrogram segment
Pre-trained Lip-Sync Expert
Sync loss
Generated frames
L1 Reconstruction loss
Skip Connections
Ground-truth segment with lower-half masked, channel-wise concatenated with another random segment
Random Reference Segment
Ground truth segment
Generator
y = 0
Discriminator
Binary cross-entropy loss
y = 1
Visual Quality Discriminator

Audio Enhancement

What it does?

- Denoises the noise while speaking or hearing in the virtual meet

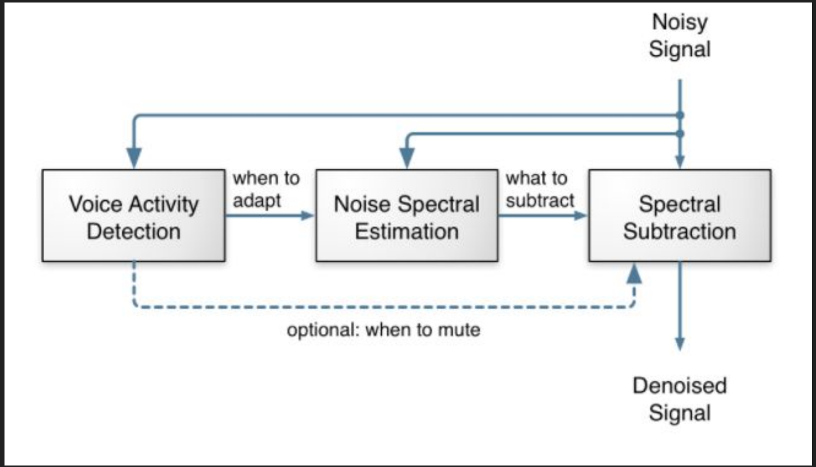
- Mutes all sound if person is not found to talk .

Impact of what it does?

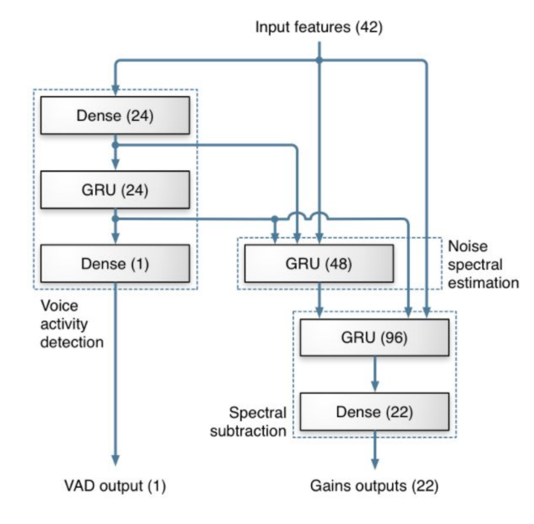
- **Prevention of unwanted noise due to accidental unmutes:** It’s the most common mistake in video conferencing. Unwanted noise can be avoided in large meets.
- **Noise Reduction:** Reduces noise in the background while talking which increases the quality of the sound.
- **Easy understanding:** Easy to capture the words of the speaker and the speaker need not repeat the words.

Implementation

- Start from conventional DSP (Signal Processing) approach
- Replace complicated estimators with an RNN (Recurrent Neural Network)



Conceptual view of a conventional noise suppression algorithm



Topology of the neural network used

Video Compression

What it does?

- Takes in the video footage from one end and compresses it locally and transmits to the other end.
- This gets extracted back and the quality is restored at the other end.
- In the process, vital information like the face are preserved and the background gets relatively blurred.

Impact of what it does?

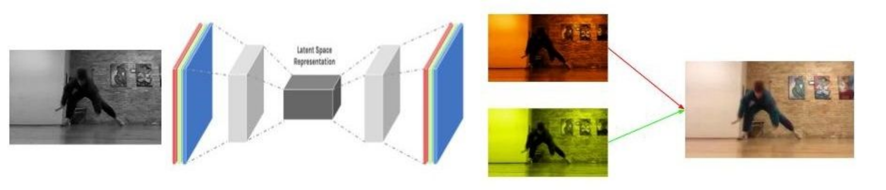
- **LOW BANDWIDTH SUPPORT :** This tool helps our users to get the maximum advantage even if having a poor connectivity or low-bandwidth.
- **DATA SAVER :** An additional feature in our software , “Data saver mode”, when activated minimizes the video traffic between devices thus saving the data of the user.

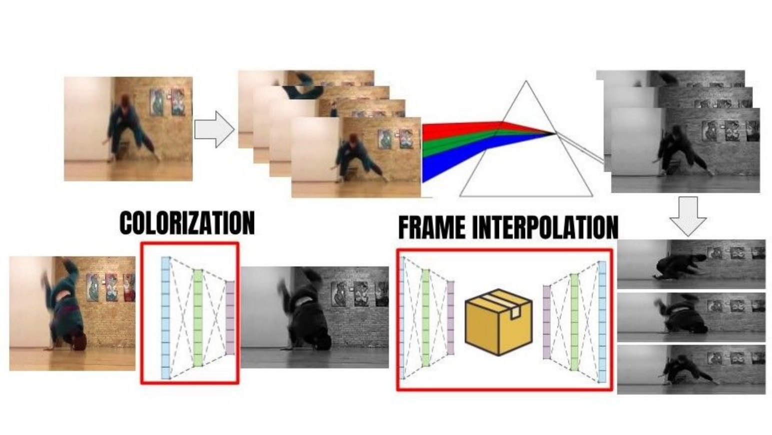
Implementation

Video compression consists of two sub-tasks:

1. Frame Interpolation
   - Reconstruction which attempts to construct the original image from scratch
   - The basic idea is to transfer hidden states both forward and backward (in time) in order to inform compression and decompression.
   - Residual which attempts to refine the scaled encoded image to achieve decompression

2. Colorization:





Knowledge Distillation
(Model Compression)

Model Compression

- **Challenges**:

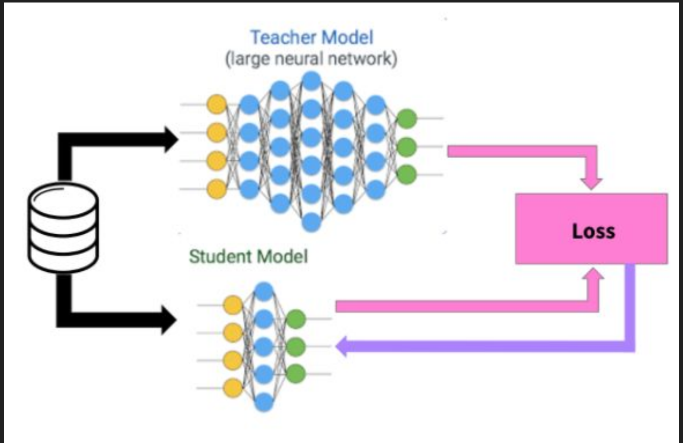
1. Most of the architectures that exist for the task mentioned in the slides introduce heavy latency due to high computational requirements.
2. Video meets happen live that is in real time. Therefore even a latency of 1-2 seconds might cause unpleasant experience to the users.

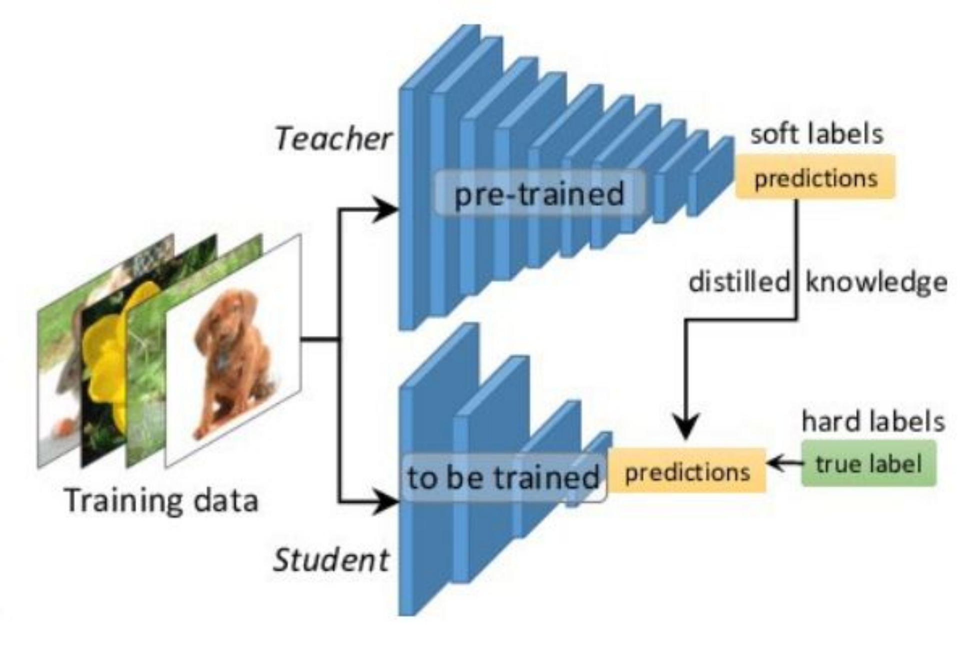
- **Solution**:

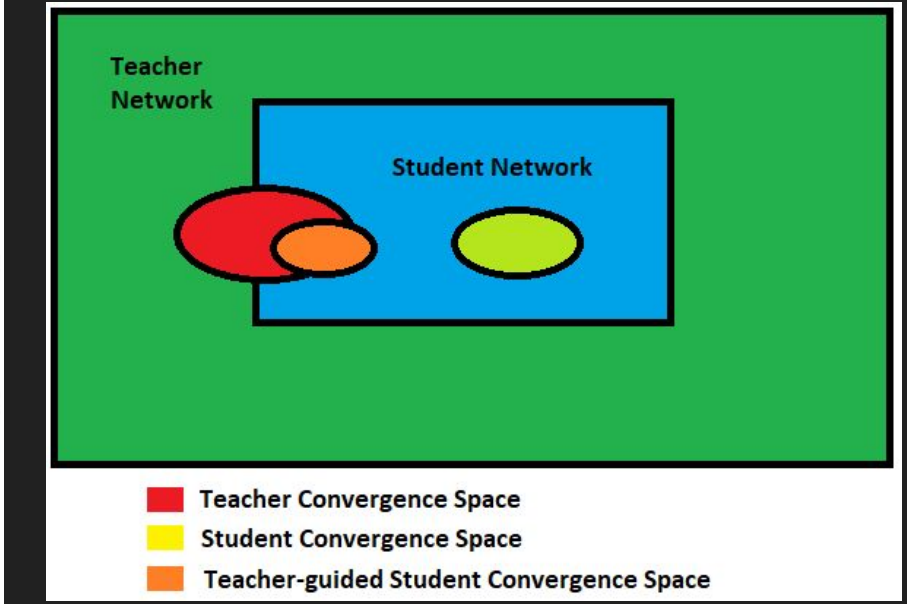
1. Using model compression, we decrease the number of parameters without compromising on image quality which results in a powerful solution for Smart Video Conferencing.

What is Knowledge Distillation?

- It refers to the idea of model compression by teaching a smaller network, step by step, exactly what to do using a bigger already trained network.
- The smaller network is then trained to learn the exact behavior of the bigger network by trying to replicate its outputs at every level (not just the final loss)




Teacher
pre-trained
soft labels
predictions
distilled knowledge
Training data
hard labels
true label
to be trained
predictions
Student


Teacher Network
Student Network
🟥 Teacher Convergence Space
🟨 Student Convergence Space
🟧 Teacher-guided Student Convergence Space

Future Goals:

- Aiming for a better learning experience.

❑ General Goals:

- Read and implement research papers as small models first.
- Move to complex models with adding unique features by the members

❑ Specific Goals:

- Implement “***Visual Speech Enhancement Without A Real Visual Stream***” and “***Speech Denoising with Deep Feature Losses***” research papers.

Thank You

Project Members: Sailesh, Harish, Awik, Nisharg, Rushill, Shrihari

Mentors: Neham Jain, Nihal JG

Bibliography

Y. Taigman, M. Yang, M. Ranzato and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701-1708, doi: 10.1109/CVPR.2014.220.

Chih-Fan Hsu, Yu-Shuen Wang, Chin-Laung Lei, and Kuan-Ta Chen. 2019. Look at Me! Correcting Eye Gaze in Live Video Communication. ACM Trans. Multimedia Comput. Commun. Appl. 15, 2, Article 38 (June 2019), 21 pages. DOI:https://doi.org/10.1145/3311784

Xintao Wang and Yu Li and Honglun Zhang and Ying Shan. Towards Real-World Blind Face Restoration with Generative Facial Prior. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2021