Exploring 3D Lip Sync Generation

Kushagra Pundeer
University of Massachusetts, Amherst
kpundeer@umass.edu

Harshul Shukla
University of Massachusetts, Amherst
hshukla@umass.edu

Abstract

Lip-syncing in Computer Vision refers to the synchronization of lips in any video to a target audio or speech. Most work in this field today is aimed at lip-syncing 2D videos or fixed 3D animations. We will employ Wav2Lip, a state of the art lip-sync generative adversarial network, as the base architecture for our approach. In our paper, we will explore two different approaches to 3D lip-sync content - generating two stereoscopic pairs or a 3D model per frame. We assume that for stereoscopic generation, extending Wav2Lip to process two stereo pairs or a 3D movie independently will result in a higher loss of Left-Right depth consistency after reconstruction. To resolve this, we will propose using a combined CycleGAN and Wav2Lip architecture that lip-syncs a 3D stereo video to a target audio while maintaining depth consistency between left and right frames. For 3D model generation, we will employ Wav2Lip and PRNet, a pretrained encoder-decoder network that is able to create a 3D facial reconstruction given single RGB image. This model's performance will be evaluated with an alternative loss function that emphasises the reconstruction of the lips. With the first approach that combines CycleGAN's and Wav2Lip, we will find that our model has inferior results than just using Wav2Lip model but gives slightly better depth consistency albeit evaluated qualitatively. We hypothesize our results were inferior because we combined L1 reconstruction loss from both CycleGAN and Wav2Lip in order to get better disparity, which ended up over-constraining our model. For our second approach that uses Wav2Lip and a modified PRNet, we found no noticeable visual difference in the output results. We believe that the model was not able to incorporate significant changes due to inadequate training runway and a low learning rate.

1. Introduction

Following other trends of globalization and the increase in use of streaming services such as Netflix and Amazon Prime Video [13], there is more and more content that is inaccessible to much of the world due to linguistic lines. The consumer today has only two options - watch dubbed films with inconsistent lip or face movements or be forced to listen in a foreign language and read subtitles. Watching dubbed content can often be perceived as unpleasant because the lip movements are not in sync with the dubbed audio and doesn't evoke the same feeling of watching videos in their original language. Additionally, the subtitles are often hard to keep up with for many viewers. We hypothesize that through computer generated lip syncing, we can resolve this issue by creating synchronized lip movements corresponding to the target audio that can be dubbed in any language. This topic has received considerable attention [2, 6, 7, 8, 17] in the research community.

Additionally, as the years have passed there has been an increasing interest and need for 3D content. This is in part due to increase in commercial virtual reality (VR) and augmented reality (AR) systems. In 2023, the virtual reality market is projected to hit 34 billion globally. In 2019 both the AR and VR industries combined were estimated to be valued at 16.8 billion. [15]. As an example, consider the acquisition of Oculus Rift [10], one of the most popular VR gaming headsets today, by Facebook for 2 billion dollars back in 2014 [14]. As this technology and the media that support it continue to creep their way into the mainstream, it is reasonable to assume that we will see it applied to a host of different settings and be available to people of various backgrounds. In order to ensure that the media displayed by VR is accessible across cultural and linguistic lines, it will be helpful to be able to lip-sync this media to different languages or audio altogether.

An example of lip-syncing being a boon to 3D data is the education sector. By lip-syncing educational videos to various languages in addition to just plain dubbing, it will make lessons easier to follow and understand for those who speak a different language than the original lesson. A survey by Common Sense Media revealed that more than 60% of parents believe that VR can offer valuable educational content for their children [15]. This could

be a valuable asset to replace or supplement traditional education, especially in a remote-learning COVID-19 reality. Advertising and shopping is another industry that could benefit from this technology. The immersiveness of a 360 degree experience for shopping makes it hard for plain images to compete. In fact, video ads that use 360 degree technology see a 7% increase in click through purchase intent, likely due to the videos being more engaging [16]. If this content starts to be produced in mass, it is logical that corporations would want to create an ad in one language and be able to easily lip-sync actor into any desired language for global marketing campaigns. While these are a few examples of where this technology could be disruptive and helpful, there are many more, some of which have yet to be discovered.

For computer generated lip-sync technology to succeed, it must be extremely accurate and precise in its reconstructions. In fact, most viewers can recognize an out-of-sync video segment as small as just 0.05-0.1 seconds [12] in duration, which can be distracting and creates a feeling that the video might be fake. If these representations are then to be extended to the 3D case, it is important to ensure that the synchronization between frames, audio and the reconstruction of the lips is retained. To this end, the key contributions and claims of our work are as follows. We will propose a new architecture that leverages the use of stereoscopic pairs by combining the efforts of Wav2Lip [12] and Cycle-GAN [11]. This will produce results that could be viewed through each lens of a VR goggle. The other contribution of this work is the use of Wav2Lip[12] and PRNet[19] to generate self-contained 3D models that represent a talking face frame by frame. PRNet will be modified in order to improve the 3D reconstruction of the lip region.

2. Background

2.1. Wav2Lip

The Wav2Lip architecture proposed in [12] builds upon previous works like [7][8] to lip sync videos with talking faces of any identity and voice. Their core architecture can be summed up as "Generating accurate lip-sync by learning from a well-trained lip-sync expert" [12]. They incorporate a powerful lip-sync discriminator that can enforce the generator to consistently produce more accurate and realistic lip motion when compared to networks that use sync-loss. They identify L1 reconstruction loss used in [7][8] and discriminator loss used in [8] to be inadequate to penalize incorrect lip-sync generation. To resolve this they proposed a new evaluation framework that consists of new benchmarks and metrics that enable fairer lip synchronization evaluation for unconstrained videos.

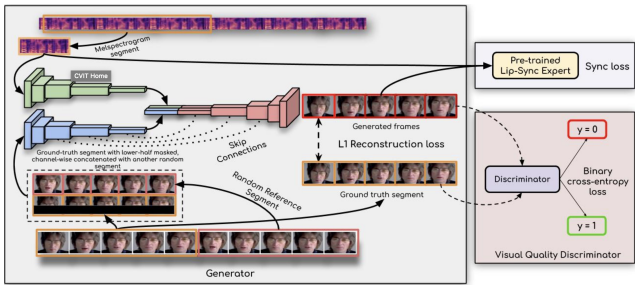

Figure 1. Wav2Lip[12]

For the discriminator the authors propose the following changes to SyncNet [3]. They feed the discriminator colored images instead of gray scale, deepen the discriminator model by adding residual skip connections, and make use of cosine-similarity[12]. The architecture of the model is best explained by the authors themselves:

> The generator in the architecture consists of three blocks (1) Identity Encoder, (2) Speech Encoder, and (3) Face Decoder. The identity encoder consists of a stack of residual convolutional layers that encode random reference frame concatenated with pose-prior (target-face with lower section masked). The speech encoder consists of a stack of 2D convolutions that encode input speech segment concatenated with face representation. Finally the decoder consists of stack of convolutional layers alongside transpose convolutions for up sampling. The generator is finally trained to minimize L1 reconstruction loss between the generated frames and ground truth frames. [12]

2.2. PR-Net

PRNet (Position-map Regression Network) is an end-to-end method that can predict dense alignment and reconstruct the shape of the face from a single RGB image [19]. The architecture of the model is a simple encoder-decoder network. The encoder portion uses a single convolution layer with 10 residual blocks, mapping the RGB image into a 8x8x512 feature space. The features are then used in the decoding blocks with 17 transposed convolutions to create the position map. The position map takes the same shape as the input image and contains information for both the dense alignment and 3D reconstruction [19].

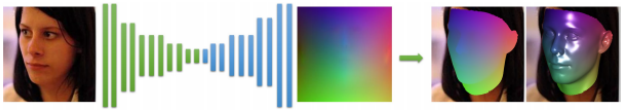

Figure 2. PRNet workflow. RGB to UV Position to 3D Model [19]

The novel improvements of PRNet include the use of the UV position map as the target of the network and a weighted

MSE as the loss function. As shown in Figure 2, the UV position map (a 2D image with 3 channels) stores the 3D coordinates of each point. At any point of the position map, the RGB values across the 3 channels can be directly interpreted as XYZ coordinates. As explained by Feng in the original paper, this representation allows for the use of a much simpler architecture that is lightweight and fast [19]. The weighted MSE allows for the network to return a large penalty when incorrectly predicting regions of the UV Position map that are important and a small penalty when they are not. This is useful because only some parts of the output space are important when reconstructing a face. The network heavily penalizes mistakes on the eyes, nose, mouth, and the rest of the face, while giving no penalty for incorrect predictions in the neck area [19]. The exact penalties and their effect on the reconstruction will be discussed in the approach section.

2.3. VisemeNet

VisemeNet is a deep learning network proposed in [20] that makes use of a three stage LSTM network to predict and generate generic facial images, referred to as animated visemes. The lip movements of these visemes are synchronized to target speech audio in real time. The first stage of the LSTM network makes predictions of phoneme-groups from audio. The second stage generates predictions on the geometric location of important facial landmarks with respect to the audio. Finally, the third and last stage uses the predicted phoneme-groups and facial landmarks to generate JA-LI parameters and sparse speech motion curves, which are then used to create face animations.

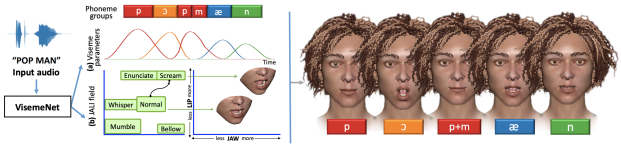

Figure 3. VisemeNet[20]

While this is an impressive and useful network, the problem it is trying to solve is a bit different than that of this paper. This network can generate a face given just the audio. If applied to our case, it would completely regenerate the entire face, not just the lip region. The goal is to sync the lips while leaving as much of the other human attributes the same. Furthermore, this network creates animated faces while our work tackles lip-syncing real faces. Additionally, syncing the lips of a real person (who still retains the rest of their facial features) is a task that requires greater accuracy and precision in comparison to animated visemes in order be unnoticeable to the human eye. Regardless, the architecture and lessons learned of VisemeNet are still valuable to the concept of lip-sync generation in the wild.

3. Approach and Methodology

Without regard for any particular commercial implementation of virtual reality or 3D media, the approach to reconstructing lip-synced media can take two viable paths. The first is to create two lip-synced stereoscopic pairs that can be viewed through each eye, likely in the form of some headwear. This would render the media in such a way that it would appear to the user that they have a 3D perspective due to two different angles that are aligned and accurate, similar to how humans infer depth. The other option involves reconstructing the media completely as a 3D model, where the process for visualizing the output media is left to the user. One can imagine an augmented reality framework, where the media is reconstructed and the user can move around and examine the lip-synced media from any angle of their choosing. Since the focus of this work is the viability of these methods after or during lip-sync generation, there will not be a focus on 3D reconstruction in the abstract case.

3.1. Stereoscopic Approach

In the Stereoscopic approach we use a combination of Wave2Lip [12] for lip-sync and CycleGAN for depth-estimation [11] as shown in Figure 4. We hypothesized that using CycleGAN alongside Wav2Lip would be best suited for generating 3D stereoscopic lip-synced content. Since CycleGAN takes stereo pairs as input to estimate depth in unsupervised manner, we thought it might be a good starting point to extend Wav2Lip's architecture [12].

[11] proposes a novel approach for unsupervised adversarial depth estimation using cycled generative networks. They estimate depth disparity in an indirect means through image synthesis using different views with an adversarial learning strategy. In order to make accurate stereo image reconstructions, the synthesized stereo pairs are constrained by each other in the form of a combined loss function.

3.1.1 Problem Statement

The problem can be formally defined as follows: given a left image I_l and a right image I_r from the camera, we are interested in generating $\hat{I}_l$ and $\hat{I}_r$ that are lip-synced to some target audio and are stereo-consistent. Here, stereo-consistent means that the left and right images generated remain stereo pairs of each other and therefore can be used to create a 3D movie or a VR scene. The visual input at each time-step are left-right stereo pairs of the current face crop concatenated with the same face crop with the lower half masked in order to be used as a pose prior for the generative network. This reduces the need of changing the pose significantly in the

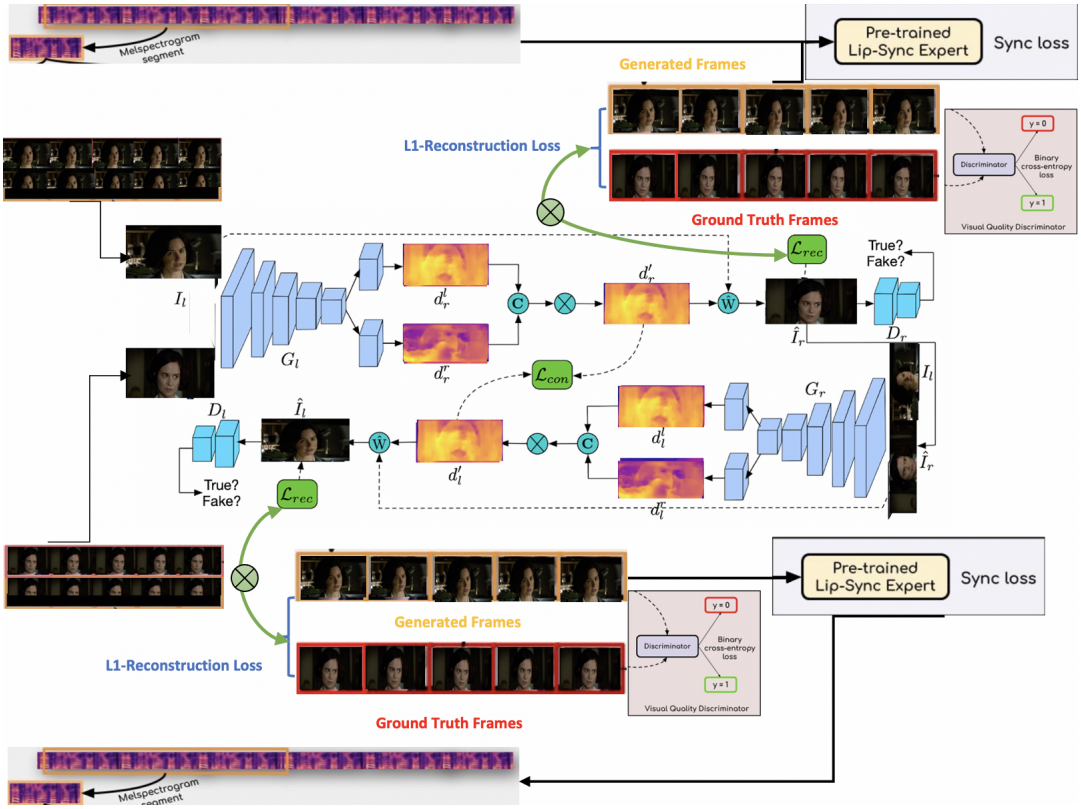

Figure 4. Wav2Lip inside CycleGAN. Our model uses pre-trained Wav2Lip and CycleGAN models and trains on stereo image pairs collected from our 3D movie dataset to synthesize lip-sync. The symbol c in the architecture denotes a concatenation operation; L_{rec} represents the reconstruction loss for different generators; L_{con} denotes a consistence loss between the disparity maps generated from the two generators. [12][11].

model during inference and helps in reduction of artifacts. The corresponding audio segment is also provided as input to the speech sub-network and the network generates stereo faces which were cropped with the mouth region morphed.

3.1.2 Stereo Lipsync + Depth Estimation

We assume that that left-to-right disparity $d_r^{(l)}$ is produced from a generative network G_l with left image I_l as input. To synthesize the right image we sample from I_l and use a warping function $f_w(.)$ as shown below.

$$\hat{I}_r = f_w(d_r^{(l)}, I_l)$$

We use a L1 reconstruction loss $\mathcal{L}_{rec}^r$ for the right stereo frame as a combination of two L1 reconstruction losses from Wav2Lip and CycleGAN as shown below and in the Figure 4 with green cross.

$$\mathcal{L}_{rec}^r = \mathcal{L}_{rec}^{Wav2Lip} + \mathcal{L}_{rec}^{CycleGAN}$$

$$\mathcal{L}_{rec}^r = ||I_r - f_w(d_r^{'}, I_l)||_1 + \frac{1}{N}\sum_{i=1}^{N}||L_g - L_G||_1$$

This reconstruction loss for right image $\mathcal{L}_{rec}^r$ is used to optimize the generator G_l. For the synthesized images $\hat{I}_r$, a discriminators D_r outputs a scalar value which is used to discriminate if the image $\hat{I}_r$ or I_r is fake or real.

3.1.3 Dataset for Stereoscopic Approach

The Wav2Lip model is originally trained on LRSA2 [12] for lip-syncing. It consists of videos of thousands of spoken sentences from BBC television. CycleGAN is originally trained on KITTI and CitySpace datasets [11] which are both stereo datasets developed for autonomous driving application. For our combined model we used the pre-trained models trained on the above datasets and used transfer learning to augment and evaluate on new datasets taken from 3D stereo movies alongside dubbed target audio files. For pre-processing, the movie was cut and cropped into sec-

tions that have visible faces talking in both stereo frames. For training the images were resized to 180x80x3.

3.2. 3D Model Approach

In the 3D Model approach we use a combination of Wave2Lip [12] for lip-syncing and PRNet [19] for 3D Reconstruction of the face. The pipeline for a complete 3D lip-synced reconstruction would be as follows. Wav2Lip would be provided with a movie clip that has one or more speaking characters and a new segment of spoken audio that will be used to lip-sync the talking subjects. This will generate a 2D movie with the new input audio. This movie must be split and passed into PRNet frame by frame. PRNet will generate a single 3D model for every frame in .obj format, a popular 3D file format that stores vertices, colors, vector normals, and faces. This will be the endpoint of the pipeline as the actual rendering of this data is left to the user. For the purpose of this paper, they will be visualized as 2D movie angles using MeshLab to capture screenshots of the 3D model and Windows Movie Maker to assemble each 3D model's screenshot as a movie with the initial input audio.

Note that the pipeline for PRNet also includes a pre-processing and post-processing step outside of the CNN. Initially, the framework used DLib [9], a modern C++ machine learning library, to detect and bound the facial region of the image with a CNN. The UV position map is generated when this cropped image is processed by the network. The framework then utilizes 68 different key facial landmarks as well as 45,000 unique points (which is about 2/3 of the possible capacity) [5]. At each point, the RGB value of the UV position map is used to create a vertex in the output 3D Model.

3.2.1 Problem Statement

Creating a 3D model from a single 2D image is a common area of research in Computer vision. Since we can directly infer the 3D model given the depth of each pixel, we can consider this a Monocular Depth Estimation problem. Many CNNs in the past have used 2D images and their corresponding depth maps to solve this problem. Unfortunately, these data sets are often limited in size and do not generalize well enough to the real world. There are a plethora of different shapes, objects, and orientations in the real world so this does not come as a huge surprise. This is an ill-posed problem because there is not any special information in an RGB image that would directly lend to inferring the depth, especially when generalized to the abstract depth estimation case. This is not the case when utilizing classic feature-based computer vision techniques such as Structure from Motion, which uses multiple images from the same point over time, or Stereoscopic Matching, which uses two or more images at the same time from different angles. Despite this, there are still some deep learning networks that have shown promising and impressive results. An overview of different methods can be found at [1].

Luckily for the purpose of our paper, we do not need to estimate the depth for any and all objects in the universe. Our problem is constrained to make depth estimates and the subsequent 3D models thereafter for facial imagery only. This vastly condenses the space of the types of images we need to consider, because there is a pattern to the shape, size, and orientation of human heads. While we may not be able to correctly infer the depth in every case, it is reasonable to think that we can achieve acceptable performance. The specifics of how these constraints are met are discussed in the implementation details further below. If we were trying to get the most accurate reconstruction, we would minimize an error over our predictions and the real values. Consider the equation below where H is the height of the image, W is the width, f is our predictor function, and $d(i, j)$ is the ground truth depth at the pixel (i,j):

$$Error = \sum_{i=0}^{W} \sum_{j=0}^{H} |d(i, j) - f(i, j)|$$

3.2.2 Dataset for 3D Model Approach

The dataset used to further tune PRNet will be 300W-LP [18], a dataset that contains various subject with large pose changes for each for a total of 62K images. This dataset was expanded from the original 300W dataset by taking an image in question, fitting the face to 3D Morphable Model parameters using a base face model with 68 key landmarks, and rotating the Morphable Model created to generate differently posed images[18]. Consider the images below as different poses of the same subject.

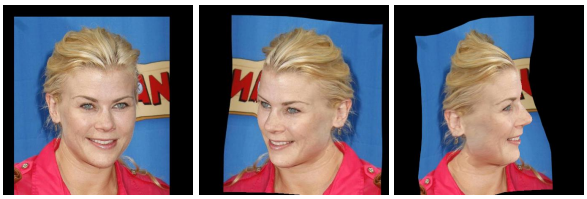

Figure 5. Original picture, left pose, right pose

This dataset was then further processed using a script from the Face3D[4] library that crops and generates the ground truth UV Position Map based on the Morphable Model for that image. The full instructions for Face3D

can be found on the Face3D Github and additional materials needed and their explanations can be found in this README

4. Experiments and Results

4.1. Stereoscopic Model

First we collected our own stereo image data-set consisting of 7000 images from 3D movies. We then pre-processed and resized the images to 180x80x3(RGB). For dataset preparation, we clipped sections of the movies that have people talking with visible lip movements in both stereo pairs. For target audio, we used the dubbed audio from the same movies.

We then incorporated the Wav2Lip architecture (both generator and discriminator) inside of the CycleGAN architecture layout while also combining the reconstruction losses of both Wav2Lip and CycleGAN. The network was fed left-right stereo pairs of the current face crop concatenated with the same face crop but with the lower half masked. The lower half in concatenated images are masked in order to be used as a pose prior for our cycled generative network. The cycled generative network generates disparity maps from the stereo pairs as intermediate feature representations that can be used to evaluate how well our model constraints for depth estimation and consistency. The output of the network is a series of synthesized stereo images that are lip-synced to the target speech and maintains depth consistency. The network is trained for 12 epochs with same learning rate = $1e^{-4}$ and other hyper-parameters as in the original Wav2Lip paper [12].

4.1.1 Stereoscopic Quantitative Results

Since our stereoscopic movie dataset was unlabeled and without any ground truth depth disparity data, we were not able to quantitatively check for disparity and stereo-image synthesis. But we were able to get quantitative results for Wav2Lip lip-syncing.

Table 1. Results for random stereoscopic videos

Method	LSE-D	LSE-C	FID
Wav2Lip	6.987	7.376	14.92
Wav2Lip+CycleGAN(Ours)	10.873	2.981	16.17

To measure results for lip-syncing we evaluated our model on the three quantitative metrics used by Wav2Lip [12] namely Lip-Sync Error Distance (LSD) (lower the better), Lip-Sync Error Confidence (LSE) (higher the better) and "Fréchet Inception Distance" FID score (lower the better). We can see from the table that our model (CyclGAN + Wav2Lip) does not perform as well as Wav2Lip. This can be attributed to the fact that we constrain our model with another L1-reconstruction loss from CycleGAN to maintain depth consistency and disparity between the generated stereo pairs.

4.1.2 Stereoscopic Qualitative Results

For qualitative results we evaluated the synthesized left-right stereo pairs through a VR Headset and also 3D glasses by creating anaglyphs which are color coded to create a 3D effect. Figure 6 shows the ground truth anaglyphs without lip sync and the synthesized anaglyph with lip sync for comparison.

Figure 6. Ground Truth Anaglyph (left), Synthesized Anaglyph (Right)

Next, we evaluate disparity maps of synthesized lip-synced stereo images by our network. Disparity is defined as the difference in image location of the same image pixel, when projected under perspective to two different cameras.

Figure 7. Left Frame (Top), Right frame (Middle), Disparity map (Bottom) (CycleGAN+Wav2Lip)

Due to GPU constraints, we resized our stereo inputs to reduce computation . Hence the intermediate disparity maps as shown in Figure 7 (Bottom) are a bit noisy, although the face and neck section is quite visible in the synthesized disparity maps which is the only part we care about for our model.

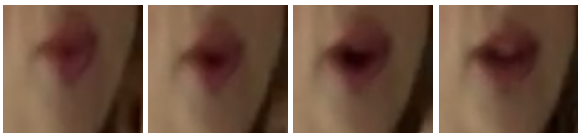

Figure 8. Lip Synchronization for the target word "Bec" of 'Because'. Full lip-syncing uploaded in the link provided.

In Figure 8 above, we have a series of pictures (cropped to show the mouth) showing the synchronization of lips to the target audio. This series of clips are for the target word "because". Because the clip series is long, we have the only included images corresponding to the letters b, e and c. An example of lip-sync generated by our model with a target dubbed audio can be seen in the link provided `https://youtu.be/zgwB4CgXny8`

4.2. 3D Model

Since, there are no known datasets with 3D lip-sync ground truth for 2d movies, we will try to qualitatively improve our final output results. In order to do so, we fine-tune the pretrained PRNet with a modified loss function to see if the 3D generation of the lips can be improved. The dataset for this approach will be the same as the original training. It is expected that much of the model's capacity will already be specialized to the weights of the original loss function.

As stated in the PRNet paper, the weighted MSE can be written as,

$$Loss = \sum ||P(x,y) - \hat{P}(x,y)|| * W(x,y)$$

where W(x,y) gives the weighting at a point (x,y) [19]. The original loss function uses a weighting ratio of 16:4:3:0 based on location. The 16 weight is reserved for 68 key points that define the edges of the face as well as important face structures. The 4 weighting is for the eye, nose, and mouth region while the 3 weighting is for the rest of the face. The 0 weighting is for the neck and hair region as they do not contribute to the reconstruction of the face. If we consider the key points as retaining their contribution to the loss and reducing other weights proportionally, the weighting can be written simply as 100%, 25%, 18.75%, and 0%.

We propose to use a different weighting scheme that puts additional emphasis on the mouth region. We will use 16:10:4:3:0 where all of the weights are the same except the 10 weighting for the mouth region and the mouth not being included in the 4 weighting. This will mean that the key points that define the face are still kept intact, while the importance of the mouth is more than doubled. This brings the ratio of weights to 100%, 62.5%, 25%, 18.75%, and 0%

Compare the old and proposed weight mask in the figure below. The key points are shown as white, the nose/eye/mouth region as grey, and the rest of the face as dark grey. The brightness of the pixel represents how much that particular pixel is weighted in the MSE loss calculation. Note that each is multiplied by the leftmost image, the binary face mask, in order to zero out the contributions of the neck area to the loss.

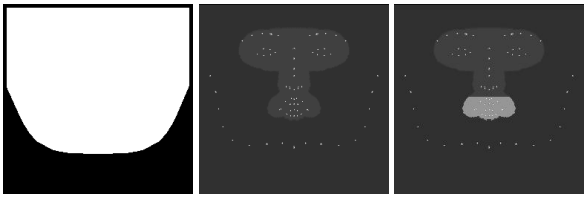

Figure 9. Binary Face Mask [5], Original Weight Mask [5], Proposed weight mask

Following the format of the original paper[19], the model was trained using Adam with a learning rate of .00001 and a half-life of 5 epochs. We also included some pre-processing augmentation techniques including rotating randomly between -45 and 45 degrees and scaling each color channels by 0.6 to 1.4 (clipping values as needed). During training we used a batch size of 16 images and our modified weighted MSE function. The model was trained on around 5,000 images from the Face3D-processed 300W-LP dataset, which was discuss earlier, for 10 epochs.

4.2.1 3D Model Qualitative Results

Unfortunately, our training procedure was not long enough to generate a noticeable visual difference in the 3D models. This is for a variety of reasons, including GPU computation restrictions. Some ideas for improving this fine-tuning will be discussed in the conclusion.

As an end-to-end example of the capabilities of this pipeline, we have composed this YouTube video `https://youtu.be/y7xDDLJx97k`. The video first shows the input to our Wav2Lip + PRNet combination model - a movie clip of a famous Indian actor, Shah Rukh Khan, talking to someone on the phone. As one can see, the lips are not matched to the audio because this is only a dub. The next two segments show a lateral and front view of our 3D model as it lip-syncs to the English audio. Note that the latter two clips have been sped up a little due to video editing software constraints. While these results are not perfect and there is blue color-shifting to a degree, the reconstruction is quite good.

For a frame-to-frame comparison of our suggested method and the original, refer to the images below. On

the left we have the results of the original model with the same frame for the modified model. Out of these three samples, only the first is visually distinguishable from its partner and the modified result is not necessarily better. One could argue that the wider space between the lips is the network reconstructing the lips more accurately, but that is a weak argument because it does not hold for all or even most frames. Another thing that is worth noting is that there is not a degradation of the quality of the generation for the nose and eyes, meaning that the network was able to retain the construction of those features.

Figure 10. 3rd frame: Orig, Modified

Figure 11. 100th frame: Orig, Modified

Figure 12. 205th frame: Orig, Modified

5. Conclusion

For the first approach we couldn't either confirm or refute the hypothesis that a combined CycleGAN and Wav2Lip is necessarily a better approach than two independent Wav2Lip on the each stereo pairs despite small qualitative improvements in depth consistency. We attributed the inferior results of our combined approach to over-constraining of the L1 reconstruction loss which can made it difficult to balance the quality of synthesized images for Lip-Syncing and depth-consistency. For future work, we can look into changing the weights of both L1 reconstruction losses to prevent the CycleGAN from over-constraining losses from disparity. We could also explore further changes to the generator in the CycleGAN, as it is possible having two different generators might overpower the training process. To resolve this we can experiment with a new architecture that is able to balance both depth consistency and lip-syncing tasks for better synthesis of stereoscopic lip-synced content. Finally we can work on developing labeled stereo training data-sets like KITTI and CityScapes specifically for Lip-Syncing. With some additional depth information we can both better evaluate and constrain our models. Increasing frames per second for both the training dataset collection and our network is another possible method for improving upon the results from our model.

For the second approach we also were not able to confirm or refute our hypothesis that training with a modified weighted MSE in order to put more emphasis on the lip reconstruction would yield better results. This is likely due to the fact that the network was originally trained for much longer with the original weighted MSE and would require more training than we conducted to properly learn this altered objective. We recommend trying this approach by both training from scratch as well as fine-tuning for more epochs with more data and comparing which results are more accurate. Additionally, it would make sense to use a larger learning rate with a longer half-life decay. This would allow for more intense fine-tuning of the network's weights. In terms of other changes that could be helpful, we recommend altering PRNet's encoder/decoder network to also process the frames directly before and after the frame in question. This would be a framework that is more appropriate for movie data and could improve results by providing temporal context on either side of the image. This could be even further extended by adding a frame-to-frame synchronization loss than penalizes when the depth at a pixel changes too rapidly. Hopefully, this would create an output that is more seamless from frame to frame and creates more accurate reconstructions.

While this project aimed to create accurate 3D reconstructions, the methods deployed were a conglomerate of different models that were able to produce results together, but were originally designed for their specific individual tasks. Despite having accepted our null hypothesis, the combinations of the models produced reasonable results even if our modifications did not improve them. This is in part due to the above as well as GPU and training time limitations. Instead of this approach, it could be beneficial to train a network from scratch for this exact problem outright. This would require a large labeled dataset as well as auxiliary discriminators that could accurately evaluate 3D reconstructions with frame-to-frame context. We urge other researchers to also investigate this problem.

References

[1] C. Z. Y. T. F. Q. Chaoqiang Zhao, Qiyu Sun. Monocular depth estimation based on deep learning: An overview, 2020.

[2] L. Chen, H. Zheng, R. K. Maddox, Z. Duan, and C. Xu. Sound to visual: Hierarchical cross-modal talking face video generation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition workshops*, 2019.

[3] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016.

[4] Y. Feng. Face3D Github. `https://github.com/YadiraF/face3d`, accessed 2020-12-07.

[5] Y. Feng. PRNet Github. `https://github.com/YadiraF/PRNet`, accessed 2020-12-07.

[6] O. Fried, A. Tewari, M. Zollhöfer, A. Finkelstein, E. Shechtman, D. B. Goldman, K. Genova, Z. Jin, C. Theobalt, and M. Agrawala. Text-based editing of talking-head video. *ACM Trans. Graph.*, 38(4), July 2019.

[7] A. Jamaludin, J. S. Chung, and A. Zisserman. You said that? : Synthesising talking faces from audio. *International Journal of Computer Vision*, 2019.

[8] P. K R, R. Mukhopadhyay, J. Philip, A. Jha, V. Namboodiri, and C. V. Jawahar. Towards automatic face-to-face translation. *Proceedings of the 27th ACM International Conference on Multimedia*, Oct 2019.

[9] D. E. King. Dlib C++ Library. `http://dlib.net/`, accessed 2020-12-07.

[10] Oculus. Untether your expectations. `https://www.oculus.com/quest-2/`, accessed 2020-12-07.

[11] A. Pilzer, D. Xu, M. Puscas, E. Ricci, and N. Sebe. Unsupervised adversarial depth estimation using cycled generative networks. In *2018 International Conference on 3D Vision (3DV)*, pages 587–595, 2018.

[12] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar. 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*, pages 484–492, 2020.

[13] G. V. Research. Video Streaming Market Size. `https://www.grandviewresearch.com/industry-analysis/video-streaming-market,accessed2020-12-07,`.

[14] B. Solomon. Facebook Buys Oculus, Virtual Reality Gaming Startup, For 2 Billion. `https://www.forbes.com/sites/briansolomon/2014/03/25/facebook-buys-oculus-virtual-reality-gaming-startup-for-2-billion/?sh=213709832498`, accessed 2020-12-07.

[15] I. Stevanovic. 30 Virtual Reality Statistics for 2020. `https://kommandotech.com/statistics/virtual-reality-statistics`, accessed 2020-12-07.

[16] E. Sweeney. 360 video ads boost purchase intent by 7%, study finds. `https://www.marketingdive.com/news/360-video-ads-boost-purchase-intent-by-7-study-finds/51895,accessed2020-12-07,`.

[17] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner. Neural voice puppetry: Audio-driven facial reenactment, 2020.

[18] X. L. H. S. S. Z. L. Xiangyu Zhu, Zhen Lei. Face alignment across large poses: A 3d solution, 2015.

[19] X. S. Y. W. X. Z. Yao Feng, Fan Wu. Joint 3d face reconstruction and dense alignment with position map regression network, 2018.

[20] Y. Zhou, Z. Xu, C. Landreth, E. Kalogerakis, S. Maji, and K. Singh. Visemenet: Audio-driven animator-centric speech animation, 2018.