CROSS-LANGUAGE SPEECH DEPENDENT LIP-SYNCHRONIZATION

Abhishek Jha[1], *Vikram Voleti*[1], *Vinay Namboodiri*[2], *C. V. Jawahar*[1]

[1]CVIT, IIIT Hyderabad, India, [2]Department of Computer Science and Engineering, IIT Kanpur, India

{abhishek.jha@research, jawahar@}.iiit.ac.in, vikram.voleti@gmail.com, vinaypn@iitk.ac.in

ABSTRACT

Understanding videos of people speaking across international borders is hard as audiences from different demographies do not understand the language. Such videos are often supplemented with language subtitiles. However, these hamper the viewing experience as the attention is shared. Simple audio dubbing in a different language makes the video appear unnatural due to unsynchronized lip motion. In this paper, we propose our system for automated cross-language lip synchronization for re-dubbed videos. Our model generates superior photorealistic lip-synchronization over original video in comparison to the current re-dubbing method. With the help of a user-based study, we verify that our method is preferred over unsynchronized videos.

Index Terms— Lip-synchronization, visual-dubbing

1. INTRODUCTION

Speech videos in many ways have affected socio-political and cultural construct of modern human civilization. Massive Open Online Courses (MOOCs), are prime examples of how online instructional videos can help skill development beyond the boundaries of conventional classrooms. Yet we find limited penetration for these speech videos when they cross international boundaries. A major reasons for this is a cultural gap between the linguistics of the audience and the content producer. This results in slow learning curves as well as dropouts from such online courses. Subtitles in different languages do not lend enough help since they divert the attention of the audience.

One way to alleviate this problem would be to re-dub speech videos in the accent or language of the audience. The current process of dubbing involves *translation* to the target language to resemble the lip motion of the source language as much as possible, *recording* of the dubbed content in pace with the original performance, and *editing* of the dubbed soundtrack and lip motion to be temporally close. This process is performed by production companies, and is both time-consuming and expensive. Moreover, this creates a clear visual discrepancy between the lip motion and the audio track which arises due to differences in correspondence of phoneme sequences and lip motions [1], which causes a strong discomfort for viewers due to alteration in the perceived sound [2].

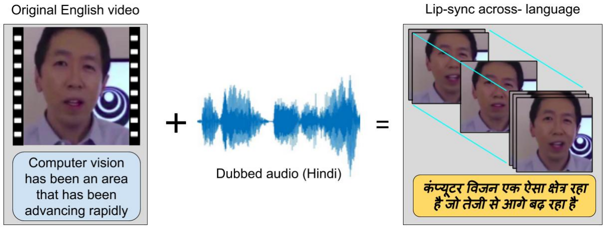
Original English video
Computer vision has been an area that has been advancing rapidly
+
Dubbed audio (Hindi)
=
Lip-sync across- language
कंप्यूटर विजन एक ऐसा क्षेत्र रहा है जो तेजी से आगे बढ़ रहा है

Fig. 1. Lip synchronization on Andrew Ng Machine learning tutorial video based on dubbed audio: Cross-language lip-sync to synchronize lip-motion of the original English video (left) into a different language (Hindi)(right).

This is also a huge distraction for those who are hearing-impaired, as they rely significantly on lip reading [3, 4]. A similar problem exists in the field of computer animation, where lip-motion of the animated characters are constrained upon the textual script of the character. This usually requires a human in the loop to manually lay the visemes for animated characters. Hence, such a system cannot be scaled for photo-realistic lip-synchronization (lip-sync). These findings are the motivation for our work to solve the problem of lip-sync.

In this paper, we propose a model for lip-syncing a target video based on the audio dubbing in a different language, for instance English video with Hindi audio dubbing, as shown in Figure 1. The inputs to our model are speech video where the lip-motion of a speaker is clearly visible, and the dubbed audio. The output is a video generated with synchronized lip motion. We also propose a scalable pipeline for dataset creation, which will later be used to train our models. Unlike audio-dubbing which requires professional dubbing artists to give their voices, our method does not depend on human visual input for lip-synchronization of a target video. Figure 1 shows the lip-synchronization for a video clip of an Andrew Ng MOOC video [5]. We evaluate our generative model based on the structural similarity (SSIM) index of the lip-synchronized videos with respect to the original English videos. Lastly, through a user-based study we show that lip-synchronization makes the speech video more engaging while preserving photorealism.

2. RELATED WORK

In the past few years, a number of work have appeared in the field of visual speech recognition, like lipreading [6, 7], word-spotting [8], and speech reconstruction [9]. However, very few recent work appears in the inverse domain of lip-synchronization. The earliest work related to ours could be animation of facial movements in avatars modeled either from audio [10] or text [11]. They mostly used HMM for sequential lip trajectory generation [11]. One of the first systems for animating a virtual avatar's face directly from speech [12] models joint distribution of acoustic and visual speech.

The work by Bregler *et al* [13] learns a mapping between the visemes and phonemes for one specific actor and language, and synthesizes new lip movements through image warping. However, the results of this method fail to dub between different languages and different individuals. Some of the recent work focus on synthesizing photo-realistic lip-motions and facial expressions. Face2Face [14] morphs the facial landmarks of a person in a target video based on those of another actor. However, these kind of models require visual feed from a human in the loop, which can be quite expensive to scale, and require a stronger supervision.

The advent of Recurrent Neural Networks, especially LSTMs [15], gave way to better sequence learning, which made generation of features from speech more efficient. Pham *et al* [16] used CNN followed by an LSTM to generate face parameters from input audio waveform. Karras *et al* [17] proposed a network consisting of a spatial convolution layer followed by a temporal convolution network on top of fully connected layers to convert speech audio into facial expressions. Chung *et al* [18] proposed an encoder-decoder convolutional network to jointly embed face and audio.

Most similar to our work are [19, 20] which use speech audio represented as Mel Frequency Cepstral Coefficients (MFCC) features [19] and text [20] to train an LSTM to produce a sequence of lip landmark points. The lip landmarks are then used to generate mouth texture. Finally this mouth texture is merged with the face in the original frame. Our work is different from [19, 20] in that our method synchronizes lip motion across two different languages, in contrast to just English-to-English. Hence, our challenges include learning higher-level viseme-phonemic relations across languages.

3. METHOD

Instructional videos provide a controlled framework for our problem, since the speakers usually speak scripted dialogues, in good lighting, facing the camera. The challenge is to model the lips, and generate new lip movements for the same speaker, given the dubbed audio in another language. This section discusses our proposed methods to address these challenges.

3.1. Cross-language lip-sync

Major challenges in lip-syncing audio of a foreign language (e.g. Hindi) on video of original language (e.g. English) are the differences in their grammatical structure and set of phonemes. One way to solve this is to directly generate lip-images conditioned upon the foreign language audio and target video. But such end-to-end systems require large amount of training data to learn the complex audio-visual relation between the two modalities [18]. At the same time, recent developments in generative networks [21, 22] have yielded impressive results [23]. Considering these factors, we first learn an embedding between Hindi audio and lip-landmarks. This allows us to predict lip-landmarks from a relatively smaller speech corpus. From these predicted lip-landmarks, we generate mouth images over the original English video to match the Hindi audio. This entire two-step pipeline can be seen in Figure 2 (which also includes an intermediate processing step, discussed in Section 4.2).

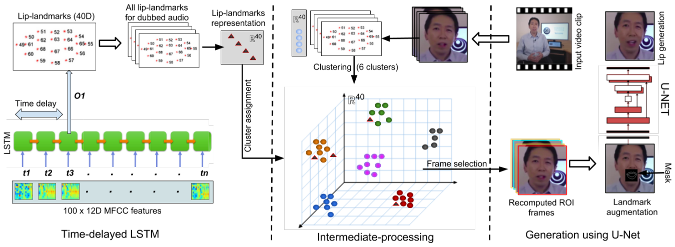

Fig. 2. Cross-language lip-sync: (left) Pipeline for training LSTM with Hindi speech and lip landmarks, (center) shows reassignment of frames for each predicted lip-landmark using intermediate-processing step, (right) pipeline for inferring using U-Net on frames from English video.

3.1.1. *Audio to Lip Landmarks*

The first step is to encode audio into lip-landmarks. For each phoneme there exists a viseme, and the lip-motion responsible for the transition between two different visemes depends on the its location in the viseme-sequence that constitutes the spoken word. This makes audio to lip-landmarks a sequence modeling problem. Hence, similar to [19, 20], we use an LSTM [15], which takes 25ms audio MFCC features at 10ms time step as input, and predicts lip-landmarks at a time step 't' = 20 (or 200ms) after first time step, and is, therefore, called Time-Delayed LSTM (TD-LSTM) [24], as shown in Figure 2 (left). During inference, given a new audio sample, we use the predicted lip-landmarks as the prior to generate the mouth (lip-region) in the second step.

3.1.2. *Lip Landmarks to Generated Faces*

Once lip landmarks are predicted from the audio in foreign language, in the second step the lips of the speakers in the original video must be modified to match these landmarks. We, hence, use a U-Net similar to [20] to generate mouth of the speaker conditioned on an encoded prior. During training, the input to the network is the $256X256$ RGB face image of the speaker, with the mouth masked by a black box of constant size, and the original landmarks in the face drawn as a white polygon, see Figure 2 (right). The output of the network is the original face image. This allows the network to learn to generate actual face of the speaker with the lip-region conditioned upon the lip-polygon on the masked face.

As L1 loss is commonly used while generating images, we use this to train the U-Net. In addition, since our main focus is on correctly generating mouth region of the speaker, we add another loss term to penalize wrongly predicted pixels in that region. Considering the mean of the black mask as the center of the mouth region, we add a Gaussian weight kernel G_{loss} to the L1 loss such that the weight of this loss decreases radially from the center of the mouth to the face extremities. Formally, for a ground truth $\hat{y}$ and the predicted face frame y, where any pixel location is represented by (i, j), our loss is defined as:

$$L = (\sum_{i,j} \|\hat{y_{ij}} - y_{ij}\|) * (1 + G_{loss}) \qquad (1)$$

$$G_{loss} = \sum_{i,j} c * \exp \frac{(i - u_i) * (j - u_j)}{v_{ij}} \qquad (2)$$

In equation 2, c is a normalization constant, u_i and u_j represent the mean pixel location of the black mask (mouth region), and v_{ij} represents the covariance.

During inference, the mouth is augmented with lip landmarks predicted by the TD-LSTM network. Thus, the U-Net will then generate faces according to the Hindi dubbed audio. Unlike [19, 20], we train U-Net on multiple sources, allowing the network to generalize over multiple speakers.

3.2. Dataset

We require two different datasets, one for each language: (i) Hindi speech dataset, for training time-delayed LSTM, and (ii) English videos dataset, to train U-Net for lip-generation.

Hindi speech dataset: As we wish to learn an encoding from Hindi audio to lip landmarks, we require a dataset consisting of Hindi audio to train a time-delayed LSTM. Since parallel audio speech corpus is difficult to find, and because dubbing is mostly a post-production phenomenon, we record 5 hours of audio-visual data of a native Hindi speaker narrating articles from Hindi newspapers and stories. Using voice activity detection [25], the video clips are segmented to give continuous segments of speech clips. For 5 hours of speech data, we get 5000 video clips of average length 2 seconds. Each such video clip is then sampled at 25 frames per second (fps). To obtain landmarks, we use a HOG-based face detector to find the speaker's face in the clip, and predict 68 face landmarks using Dlib [26]. We then choose the landmarks corresponding to mouth region (landmarks 49 to 68) a.k.a. lip-landmarks, and normalize them. For each video clip segmented by voice activity detection, these normalized lip-landmarks are saved. Similarly, for each video clip we extract audio and sample it at 100Hz. We extract MFCCs for each sampled segment of the extracted audio clip. The training set consisted of 90% of total dataset, validation was done on rest of 10% of the dataset.

English speech dataset: As our aim is to generate lip-synced Andrew Ng's machine learning videos with Hindi dubbing, we use 20 Andrew Ng videos to create a dataset of English speech videos. The input to our U-Net is frames from instructional video clips. For each frame where the face of the speaker has been detected, a square region around the face with 1.5 times the bounding box of the 68 landmarks [26] is extracted. This results in images with full visibility of the instructor's face. The mouth region of each face is considered as the bounding box around the mean of the mouth landmarks (49 to 68), and of width 0.25 times the width of the face region. It is then replaced with a black mask and a white polygon connecting the lip landmarks, and resized to the input shape of the U-Net. The output of the U-Net is the original face image. It is important to not let the network overfit on the input images. We therefore used multiple image sources to pre-train U-Net — frames extracted from 1) Telugu movies, 2) videos of Andrew Ng's deep learning.ai lectures, 3) GRID corpus [27], 4) Hindi Speech dataset Table 1 details the number of frames from each source.

Source	Train	Validation
Telugu movies	37130 frames	4159 frames
English speech	24359 frames	16035 frames
GRID	13350 frames	1500 frames
Hindi speech	37790 frames	3714 frames

Table 1. Dataset distribution for training U-Net.

4. IMPLEMENTATION

4.1. TD-LSTM

Our proposed TD-LSTM model consists of a single layer, 60 neurons, LSTM followed by a 40D dense layer. We also up-sample each video clip at 100Hz, matching the sampling rate of audio, to compute lip-landmarks. In each forward pass, the network takes 100 time-steps MFCC features (100 x 12D) and predicts the 20th up-sampled lip-landmark frame (1x40D). This is done densely for each Hindi audio-visual clip. This results in an offset in the prediction of lip-landmarks of 200ms at the beginning, and 800ms at the end of the video. We compensate for this by replicating the first and the last frame's predicted landmarks respectively for the appropriate number of frames. We also implemented TD-LSTM with 500ms and 800ms time-delays. But we found very little perceptual difference between the results, and therefore chose 200ms delay. Also, using a Bidirectional TD-LSTM did not perceptually affect the results.

We implemented the network in Keras deep learning framework [28], with a batch size of 64, mean square loss, and Adam [29] as the choice of optimizer. We trained our network for 20 epochs, with a total time of around 4 hours on Nvidia GTX 1080 Ti, till the loss started plateauing. Pre-training with 10% videos randomly sampled from GRID corpus resulted in faster saturation of loss.

4.2. U-Net

We use U-Net architecture similar to [20] and trained the it on 4 NVIDIA TitanX GPUs, using a batch size of 16, counting 4 batches per iteration, until $\approx$ 5000 iterations. This took $\approx$ 2 seconds per iteration, and occupied $\approx$ 3.3GB of memory including model weights and images kept in the buffer. As U-Net has been trained on Andrew Ng's lip-landmarks to predict original frame, it also learns an undesirable mapping between jaw location and the shape of lip in the output. To overcome this, we introduce an *intermediate-processing* step between TD-LSTM and U-Net. We normalize the lip-landmarks from all the frames in the target instructional video in English, and group them into 6 clusters. We empirically found that restricting the number of clusters to be 6 provides the optimal quality and speech trade-off. All the lip-landmarks predicted by TD-LSTM are then assigned to a cluster based on their distances from the centroids frames. This allows the predicted lip-landmarks to be assigned to the appropriate face frames. Only the set consisting of these 6 frames are then fed to the U-Net, hence generating 6 distinct facial poses corresponding to the centroid frames. This results in a jittery face video but with proper lip-synchronization.

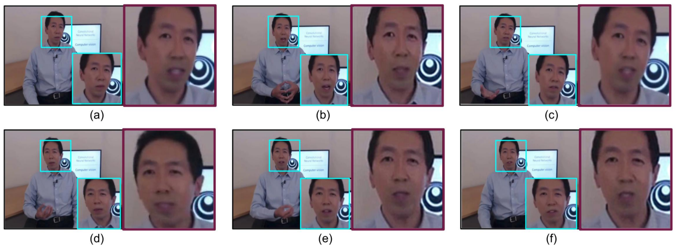

Fig. 3. Qualitative results for cross-language lip-sync: Each of the 6 images depicts original English video (left) along with its enlarged ROI, (right) shows our generated Hindi lip-synced video.

The frames generated from the U-Net are slightly blurred, therefore we use a pre-trained CNN deblur network trained on facial images for sharpening. We then compute the pairwise homography between each generated video frame and that of the original instructional video clip using all the 3D face-landmarks [30]. This gives us a transformation matrix between the frame pairs. We then crop a rectangular mouth region in the generated video frames which is augmented over the original video using the computed transformation matrix. All these ROI augmented frames along with Hindi dubbed audio are then used to create the final Hindi lip-synced video.

5. RESULTS

User-based study: To check the quality of our proposed models, we conducted a user based study on 10 Andrew Ng's ML video clips of length upto 1 minute. We randomly selected five videos with Hindi audio naively overlaid (un-synced), and their corresponding Hindi lip-synced versions, and showed them to each of the 20 users. We then asked them to rank the videos between 1 (hard to understand) to 5 (easy to understand) based on their comfort. Since comfort can be subjective and ill-defined, we asked users to rate the percentage of lip-synchronization perceived by them for each pair. As shown in Table 2, the means of the comfort score and percentage lip-synchronization were higher for our cross-language lip-synced videos. The average comfort rating across users for each video pair can be seen in Figure 4 (a), where as average percentage lip-synchronization can be seen in Figure 4 (b). We also show qualitative results of cross-language lip-synchronization in Figure 3.

Quality of generation: We also compare SSIM index [31], a widely used metric to evaluate the quality of generated images. To evaluate the frame generation performance of our cross-language model, SSIM scores was computed between the generated frames and the original frames for the four datasets used to train U-Net. The average SSIM score for each of these dataset can be seen in Figure 4 (c), with mean average SSIM score of 0.58 with the overall standard deviation of 0.05.

	C-US	C-S	LS%-US	LS%-S
Mean	2.51	**3.1**	23.86	**45.95**
Std-dev	1.07	0.6	25.9	24.1

Table 2. Mean scores and standard deviation for Cross-language lip-sync on Hindi: (C) comfort level for (US) un-synced speech overlay, and (S) lip-synced version; (LS%) Lip-Sync percentage for (US) un-synced and (S) lip-synced versions.

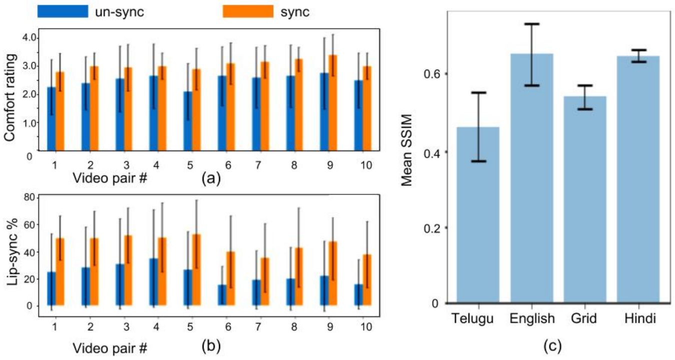

Fig. 4. (a), (b) User feedback for cross-language lip-sync corresponding to 10 video pairs - showing average comfort rating and, average percentage of perceived lip-synchronization; and their respective standard deviation for lip-unsynced (blue) and lip-synced videos (orange). (c) Mean and standard deviation of SSIM scores for various datasets used to train U-Net

6. CONCLUSION

We propose a lip-synchronization methods for dubbed educational videos to improve instructor-student engagement during online video lectures. We detail our pipelines for dataset creation for audio-to-lip-landmarks as well as lip-landmarks-to-mouth-generation. Our user-based-study shows that lip-synchronization can improve effectiveness of content delivery through dubbed speech videos.

7. REFERENCES

[1] William H Sumby and Irwin Pollack, "Visual contribution to speech intelligibility in noise," *The journal of the acoustical society of america*, vol. 26, no. 2, pp. 212–215, 1954.

[2] Harry McGurk and John MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746, 1976.

[3] Elmer Owens and Barbara Blazek, "Visemes observed by hearing-impaired and normal-hearing adult viewers," *Journal of Speech, Language, and Hearing Research*, vol. 28, no. 3, pp. 381–393, 1985.

[4] Kricos P.B. Lesner S.A., "Visual vowel and diphthong perception across speakers," *Journal of the Academy of Rehabiitative Audiology 14*, pp. 252–258, 1981.

[5] "Machine Learning Course, Andrew Ng," https://www.deeplearning.ai.

[6] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Lip reading sentences in the wild," in *CVPR*, 2016.

[7] Themos Stafylakis and Georgios Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," *arXiv preprint arXiv:1703.04105*, 2017.

[8] Abhishek Jha, Vinay P Namboodiri, and CV Jawahar, "Word spotting in silent lip videos," in *WACV*. IEEE, 2018.

[9] Ariel Ephrat and Shmuel Peleg, "Vid2speech: speech reconstruction from silent video," in *ICASSP*. IEEE, 2017.

[10] P Kakumanu, R Gutierrez-Osuna, A Esposito, R Bryll, A Goshtasby, and ON Garcia, "Speech driven facial animation," in *Proceedings of the 2001 workshop on Perceptive user interfaces*. ACM, 2001, pp. 1–5.

[11] Lijuan Wang, Wei Han, Frank K Soong, and Qiang Huo, "Text driven 3d photo-realistic talking head," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[12] Matthew Brand, "Voice puppetry," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 1999, pp. 21–28.

[13] Christoph Bregler, Michele Covell, and Malcolm Slaney, "Video rewrite: Driving visual speech with audio," in *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 1997.

[14] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *CVPR*, 2016.

[15] Sepp Hochreiter and Jrgen Schmidhuber, "Long short-term memory," vol. 9, pp. 1735–80, 12 1997.

[16] Hai X Pham, Yuting Wang, and Vladimir Pavlovic, "End-to-end learning for 3d facial animation from raw waveforms of speech," *arXiv preprint arXiv:1710.00920*, 2017.

[17] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 94, 2017.

[18] J. S. Chung, A. Jamaludin, and A. Zisserman, "You said that?," in *BMVC*, 2017.

[19] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 95, 2017.

[20] Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brébisson, and Yoshua Bengio, "Obamanet: Photo-realistic lip-sync from text," .

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," *arxiv preprint arXiv:1611.07004*, 2016.

[23] Arnab Ghosh, Viveka Kulharia, Vinay Namboodiri, Philip HS Torr, and Puneet K Dokania, "Multi-agent diverse generative adversarial networks," in *CVPR*, 2018.

[24] Alex Graves and Jrgen Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *NEURAL NETWORKS*, pp. 5–6, 2005.

[25] "WEBRTC, VAD," https://webrtc.org/.

[26] Davis E King, "Dlib-ml: A machine learning toolkit," *JMLR*, vol. 10, no. Jul, pp. 1755–1758, 2009.

[27] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[28] François Chollet et al., "Keras," 2015.

[29] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[30] Adrian Bulat and Georgios Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *International Conference on Computer Vision*, 2017.

[31] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, Nov 2003, vol. 2, pp. 1398–1402 Vol.2.