Invertable Frowns: Video-to-Video Facial Emotion Translation

Ian Magnusson*
Northeastern University
Boston, Massachusetts, USA
magnusson.i@northeastern.edu

Aruna Sankaranarayanan*
Media Lab (MIT)
Cambridge, Massachusetts, USA
arunas@mit.edu

Andrew Lippman
Media Lab (MIT)
Cambridge, Massachusetts, USA
lip@media.mit.edu

ABSTRACT

We present Wav2Lip-Emotion, a video-to-video translation architecture that modifies facial expressions of emotion in videos of speakers. Previous work modifies emotion in images, uses a single image to produce a video with animated emotion, or puppets facial expressions in videos with landmarks from a reference video. However, many use cases such as modifying an actor's performance in post-production, coaching individuals to be more animated speakers, or touching up emotion in a teleconference require a video-to-video translation approach. We explore a method to maintain speakers' identity and pose while translating their expressed emotion. Our approach extends an existing multi-modal lip synchronization architecture to modify the speaker's emotion using L1 reconstruction and pre-trained emotion objectives. We also propose a novel automated emotion evaluation approach and corroborate it with a user study. These find that we succeed in modifying emotion while maintaining lip synchronization. Visual quality is somewhat diminished, with a trade off between greater emotion modification and visual quality between model variants. Nevertheless, we demonstrate (1) that facial expressions of emotion can be modified with nothing other than L1 reconstruction and pre-trained emotion objectives and (2) that our automated emotion evaluation approach aligns with human judgements.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning**; *Multi-task learning*; **Neural networks**; **Image manipulation**; Reconstruction; **Computer vision**.

KEYWORDS

Video-to-video translation; Video Manipulation; Emotion; Facial Expression; Computer Vision; Deep Learning; Evaluation

ACM Reference Format:
Ian Magnusson, Aruna Sankaranarayanan, and Andrew Lippman. 2021. Invertable Frowns: Video-to-Video Facial Emotion Translation. In *Proceedings of the 1st Workshop on Synthetic Multimedia – Audiovisual Deepfake Generation and Detection (ADGD '21), October 24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3476099.3484317

*equal contribution

ADGD '21, October 24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8682-1/21/10…$15.00
https://doi.org/10.1145/3476099.3484317

1 INTRODUCTION

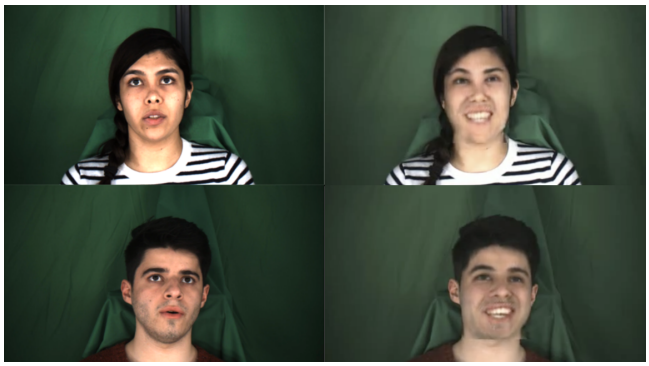

Figure 1: Wav2Lip-Emotion takes existing videos (real neutral inputs on left) and modifies facial expressions of emotion (generated happy outputs on right) while maintaining speakers' lip synchronization and pose. Examples here are from all-around best model on unseen faces.

The face is the window to the mind: the human face conveys information about our mental, physical, and of most of all, the emotional state. We can sense if someone is in emotional distress or happily animated. Using facial expression, voice, muscular tension, and other cues, we are also able to pick up on subtler emotions. Research has shown that our facial expressions can invoke an emotion in both expresser [9, 14] and the recipient of the expression [2]. Further, research on non-verbal behaviours has shown that facial displays are integral to how we signal and maintain social dominance [17] and other social cues. Even when we are tasked with maintaining a neutral facial disposition (e.g., newscasters), we often display subtle cues that convey our biases [21].

In several scenarios, the expression of facial emotion may be suppressed. For example, in studies inhibited children have lower facial expressiveness than uninhibited children [15]. Their baseline facial dynamics operate over a smaller range [26]. Facial emotion suppression is also inhibited by post-traumatic stress disorder (PTSD), as measured by the movements of certain facial muscles during a facial emotion recognition task [24]. Facial palsy interferes with expressing facial emotions as well; computer vision models that are trained to recognize emotions can often perceive less joy and greater negative emotions in the faces of facial palsy patients [8].

Thus, the artificial suppression or augmentation of facial emotion may be desirable at times. Individuals with or without inhibited facial expressions may benefit from tuning their own expressions to better fit their social circumstances. One may want to alter the expressions in videos shown to them. Speakers might be yelling at each other during a video conference, but nevertheless want

to gather the content in their exchange without the unpleasant expressions. Or a film director may want to augment or diminish the expressions of an actor.

Inspired by this idea, our primary contribution is the creation of a deep learning model that modifies the facial emotion of a speaker in a given video, while preserving the visemes, pose, and identity from the original video. Existing research in this area has produced models that modify emotion in an image [16, 22, 28], synthesize videos with a certain emotion from a single image [10, 31], or puppet expressions using reference input [30, 32]. However, to our knowledge no work exists that directly modifies the emotion in an existing video.

We build Wav2Lip-Emotion,[1] a multi-modal computer vision model that modifies facial emotion via video-to-video translation. This model adds emotion modification to Wav2Lip [27], a recent model that synchronizes face videos with speech audio by means of L1 reconstruction and pre-trained synchronization objectives. We also propose an automated emotion evaluation using the NISL2020 model [7] to examine continuous valued changes in valence and arousal in generated videos against baseline changes in pairs of emotions performed by human actors in the MEAD dataset [31]. Through the automated evaluation backed up by a user study, we demonstrate that facial emotion can be modified while maintaining lip sync and moderate visual quality by L1 reconstruction and pre-trained emotion objectives alone.

2 RELATED WORK

Image-to-image translation. One approach utilizes the StyleGAN architecture [16] to modify emotion in an image by first using optimization to recover a latent StyleGAN vector that approximates the input image and then shifting it along a learned subspace for the desired emotion shift [22]. Other researchers have employed 3D face synthesis techniques to generate various emotions on a neutral face [3] or used cycle consistency loss to overcome the availability of paired emotion data [28].

Image-to-video translation. Synthesizing a video from a single image is another context in which emotion manipulation has been explored. Fan et al. [10] utilize a dataset of short videos of faces moving from a neutral expression to a specified emotion expression to train a controlled image-to-image translator. The translator reconstructs the expression video frame by frame using the neutral expression in the first frame and the frame index offset of the frame to be generated. Most similarly to our work, the authors of the MEAD dataset [31] also make a baseline model that generates a "talking head" video. These videos are conditioned on input audio, a single neutral expression reference image, a desired emotion, and an emotion intensity. This baseline separates the task of lip sync into a module that generates the lower face based on the audio, while the task of emotion modification is addressed by a module that generates the upper face based on a desired emotion. These halves are composed together using a refinement network. While an interesting task in its own right, such talking head generation is more suitable for different use cases such as animating memes where the full reference video of speaker is not available.

[1]Code available at https://github.com/jagnusson/Wav2Lip-Emotion.

Video-to-video translation. To our knowledge no work in video-to-video translation directly tackles the problem of emotion manipulation. Other video-to-video translation work maps the pose and facial keypoints of an input driving video onto a different identity provided by a single image or video [30]. This enables the puppeting of emotional expressions and lip movements on the basis of a separate video. One clever work uses a large repository of pre-annotated reference videos to quickly look up instances of relevant phonemes with desired facial expressions which can then be used with such a puppeting technique to edit an input video [32]. While highly efficient, this approach is limited to repeating already recorded expressions. Instead we adapt the approach of Wav2Lip which utilizes an L1 reconstruction loss and a pre-trained discriminator derived from SyncNet [5] to translate out of sync lip videos to synchronized ones [27]. We extend their model to emotion modification.

3 APPROACH

The Wav2Lip architecture [27] has been shown to synthesize high quality lip synchronization given a set of video frames and unsynchronized audio. We extend this architecture to modify emotion via its L1 reconstruction loss and an additional pre-trained emotion objective, and call our architecture Wav2Lip-Emotion. It penalizes the generation of frames where the facial expression diverges from a specified emotion. Meanwhile the lip synchronization and visual quality discriminators present in the original Wav2Lip help preserve these traits while facial emotion is synthesized.

3.1 Datasets

The Wav2Lip architecture is trained on the LRS2 dataset [4] which contains thousands of spoken sentences from BBC television. To introduce emotion modification, we fine tune the Wav2Lip architecture on MEAD [31], a dataset of actors performing a set of utterances with several emotional variations that span arousal and valence.

The MEAD dataset (see Table 1) is a controlled emotion dataset consisting of 40 hours of videos of 60 actors reading the same sentence with different facial expressions. The dataset is extensive – it includes a diverse set of actors from different continents spanning 15 different countries. The speakers therefore have several regional characteristics in their faces and their manner of speaking.

As of May 2021, only the first part of the dataset containing videos of 47 individuals has been released. Each individual performs the following emotions at 3 levels of intensity: angry, disgust, fear, contempt, happy, sad, and surprise. Neutral is used as an eighth emotion but only has one intensity level. The recognizability of the emotions is validated by a user study that finds that labellers can accurately identify the intended emotion performed by the actor. All videos are framed around the head, from identical angles for all actors, with controlled lighting.

In this paper, we describe results achieved from training our model with only the happy and sad level 3 and neutral level 1 emotions. We limit our analysis to these 3 emotions with 1 intensity level each and only frontal videos due to the large amount of compute required to pre-process the data. Happy and sad emotions have highest levels of recognizability, while neutral emotion on

Dataset	Hours	Individuals	Emotion	Intensity
Unreleased				
MEAD full	40	60	8	3
Released				
MEAD train	~ 25	37	3	1
MEAD val	~ 3	5	3	1
MEAD test	~ 3	5	3	1

Table 1: The full MEAD data has not been completely released. On May 5th, 2021, a dataset of 47 actors was made available. We divide these 47 actors presently available into three splits, and only include happy, neutral, and sad emotions at one intensity due to pre-processing compute constraints.

the other hand was much lower. We include these emotions to capture a selection of easier and harder emotion modifications in our analysis.

We randomly separate the actors in the dataset into 37, 5, and 5 actor splits for training, validation, and testing. We also ensure that the validation and test splits contain at least 2 actors from the smaller of the two reported genders, and ensure that there is a spread in the representation of different ethnic backgrounds. In our test and validation set, we have 2 female and 3 male actors, while our training set has 27 male and 20 female actors.

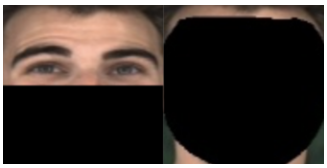

Figure 2: Examples of half and full masking strategies.

3.1.1 Pre-processing. The Wav2Lip architecture trains on windows of video frames and corresponding audio. Each frame in the window contains the face of the speaker alone. Videos of a speaker are broken down into their constituent frames, the faces are detected, cropped face frames are extracted, and stored along with the corresponding audio. As Wav2Lip is trained on 25 FPS LRS2 data we resample the MEAD data to 25 FPS as well.

In the original Wav2lip, the bottom half of these face frames are masked with zeroes. This enables the generator to focus on synthesizing the lips alone without worrying about the facial features on the top half. Since the expression of emotion involves modifying other facial features besides the lips, one version of Wav2Lip-Emotion also experiments with a masking approach that replaces the entire face with zeroes. In order to do this, we use the *dlib* library [18] along with an off-the-shelf facial landmark detector [6] that identifies a set of 81 facial keypoints. We then mask a convex hull along the boundary of the face using the list of facial boundary keypoints. See Figure 2 for examples of both masking strategies.

3.2 Method

Our Wav2Lip-Emotion approach extends Wav2Lip to modify emotion via L1 reconstruction and pre-trained emotion objectives.

(1) For our pre-trained emotion objective, we employ a DenseNet model [13] trained by [25]. This emotion classifier detects 6 emotions and neutral expressions and achieves accuracy of 73.16% on the FER2013 dataset [11]. Due to pre-processing compute constraints, we utilize only happy, neutral, and sad. We modify the Wav2Lip architecture to use this pre-trained emotion classifier as another objective. We train to maximize the class probability of the desired emotion, as predicted by the emotion classifier on the generated outputs.
(2) We fine-tune Wav2Lip-Emotion using input videos from MEAD having a specified *source emotion* and we sample the target video frames from a different *destination emotion* to allow the L1 reconstruction loss to also encourage emotion modification. For the scope of this work we only produce models that modify emotion between a pair of specific emotions.
(3) The Wav2Lip architecture masks the lower half of target image frames to create a pose prior input that does not reveal lip information. In Wav2Lip-Emotion, we experiment with an additional masking approach that replaces the entire facial area with zeroes to conceal emotion changes to the eyes, eyebrows, and other facial features besides the lips. Likewise in all of our variants, we modify Wav2Lip's adversarially trained visual quality discriminator to scrutinize the whole rather than bottom half of the frame. This change required the addition of residual connections to this discriminator to overcome vanishing gradients.
(4) At inference, unlike the original Wav2Lip architecture which synchronizes the lip movements in a set of video frames to an unsynchronized audio input, Wav2Lip-Emotion retains the original audio and simply modifies the emotion while ensuring the consistency of lip movements.

3.3 Model

We extend the Wav2Lip architecture with an additional emotion objective and modify the visual quality discriminator. Thus the model is composed of a generator and 3 discriminators: 1) lip synchronization and 2) emotion pre-trained objectives, as well as 3) an adversarially trained visual quality objective. The model operates on inputs composed of short windows of audio and face-cropped video frames, and outputs generated face frames.

3.3.1 Generator. The generator, G contains 3 blocks. An identity encoder, speech encoder, and face decoder. The architectural details for these blocks are outlined in the Wav2Lip paper. The identity encoder concatenates a random frame with a pose prior consisting of a masked version of the target face. The speech signal is also encoded as a stack of 2D convolutions which are concatenated with the frame. The decoder is also a set of convolutional layers that have been modified for upsampling.

3.3.2 Original Training Objectives. Wav2Lip follows other image translation work in using the L1 reconstruction loss between the

real ($\mathbf{L_G}^i$) and generated frames ($\mathbf{L_g}^i$):

$$L_{recon} = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{L_g}^i - \mathbf{L_G}^i\right\|_1 \tag{1}$$

The pre-trained lip synchronization discriminator used by the authors of Wav2Lip is a modified version of SyncNet [5], a model that detects lip synchronization errors in videos, which is pre-trained on the LRS2 dataset. Note that this lip synchronization discriminator is not trained further during the Wav2Lip or Wav2Lip-Emotion training process. The modified SyncNet model uses RGB images instead of grayscale images, contains a deeper network with residual skip connections, and introduces a new loss function that uses the cosine similarity between the speech audio ($\mathbf{s}^i$) and the face video ($\mathbf{v}^i$) with binary cross-entropy loss:

$$P^i_{sync} = \frac{\mathbf{v}^i \cdot \mathbf{s}^i}{max(\left\|\mathbf{v}^i\right\|_2 \cdot \left\|\mathbf{s}^i\right\|_2, \epsilon)} \tag{2}$$

$$E_{sync} = \frac{1}{N}\sum_{i=1}^{N} -log(P^i_{sync}) \tag{3}$$

Wav2Lip also uses a visual quality discriminator, which penalizes unrealistic faces. Unlike the pre-trained lip synchronization discriminator, the visual quality discriminator learns by discriminating the generated images and real images during training. The discriminator is trained to maximize the L_{disc} objective function.

$$L_{gen} = \mathbb{E}_{\mathbf{x}\sim\mathbf{L}_g}[log(1 - D(\mathbf{x}))] \tag{4}$$

$$L_{disc} = \mathbb{E}_{\mathbf{x}\sim\mathbf{L}_G}[log(D(\mathbf{x}))] + L_{gen} \tag{5}$$

Originally visual quality discriminator only examines the lower half of the face. But since emotion modification also takes place in the upper half, we modify the architecture to discriminate the whole image. We also introduce residual connections to overcome problems with vanishing gradients that occurred during the process of retraining the model on LRS2 data with this modified discriminator architecture.

3.3.3 Pre-trained Emotion Objective. In addition to the two discriminators listed above, we add a pre-trained emotion objective. We utilize a DenseNet classifier trained by [25] to predict emotion labels for each generated frame. We put the logits, $\mathbf{z}$, output by the final layer through softmax to form the likelihood of each emotion class $e \in E$:

$$g_e = \frac{\exp(z_e)}{\sum_{k\in E}\exp(z_k)} \tag{6}$$

We minimize the deviation of the desired emotion class likelihood, g_d from the maximum value:

$$L_{emotion} = \frac{1}{N}\sum_{i=1}^{N} 1 - g^i_d \tag{7}$$

3.3.4 Total Loss. The overall loss minimized by the generator is given by the weighted sum,

$$L_{total} = s_r \cdot L_{recon} + s_w \cdot E_{sync} + s_g \cdot L_{gen} + s_e \cdot L_{emotion} \tag{8}$$

where the weights s_r, s_w, s_g, and s_e are tuned as hyperparameters.

3.3.5 Variants.

Masking. We present two model variants with respect to our masking strategy. These are the proposed **full** masking approach (described in Section 3.1.1) which covers the full face, and **half** masking which preserves the original Wav2Lip masking strategy and enables the use of pre-trained Wav2Lip checkpoints.

Emotion Modification Strategy. Our work explores two avenues for encouraging emotion modification in generated videos: L1 reconstruction loss and a pre-trained emotion objective. Thus we explore 3 variants—**L1**, **Emotion Objective**, and **L1 + Emotion Objective**—in which one of each or both avenues for emotion modification are utilized. The impact of the emotion objective can be ablated by simply removing that objective. Since the L1 reconstruction loss is also critical for maintaining lip synchronization and visual quality, we cannot simply remove it. It is able to encourage emotion modification only when we provide target video frames drawn from the specified *destination emotion* which differs from the *source emotion* from which the input reference image is drawn (described in Section 3.2). So to ablate the emotion modification via L1 loss we use *source emotion* video frames for both inputs and targets.

3.3.6 Hyperparameters. All variants, except **Emotion Objective** without L1, use an L1 reconstruction loss weight (s_r) of 0.8, a synchronization loss weight (s_w) of 0.03, a visual quality discriminator weight (s_g) of 0.07, and an emotion objective weight (s_e) of 0.1. In the **Emotion Objective** without L1 variant, the emotion objective and L1 loss have contrary goals in that the target images whose reconstruction is measured by the L1 loss are from the *source emotion* rather than the *destination emotion* which is optimized for by the emotion objective. Thus in this variant we use an L1 reconstruction loss weight (s_r) of 0.6 and an emotion objective weight (s_e) of 0.3 to help the emotion modification compete with the L1 loss.

Meanwhile in the **L1** and **L1 + Emotion Objective** variants we utilize target videos from the *destination emotion* while using videos from the *source emotion* to provide the reference frames that inform the model of the identity of the speaker. However, while the MEAD data does contain emotion variations of the same utterances, the performances are not perfectly synchronized. Thus with the **full** masking models we also utilize the *destination emotion* video for the masked pose prior input, as this allows the greatest pose synchronization. The **half** masking models however do not fully obscure the emotion information in the pose prior input and thus we use the *source* instead of *destination emotion* for the masked pose prior, despite the imperfect pose synchronization.

We normalize inputs to the emotion objective by the channel mean and standard deviation of the MEAD data and convert to greyscale, as the emotion objective was pre-trained on greyscale only images. In order to further prevent vanishing gradient issues with the visual quality discriminator we clamp the gradient norm between 1e-2 and 1e10. All other settings are as specified by the original Wav2Lip.

4 EVALUATION

We propose an automated emotion evaluation approach that compares changes in valence and arousal in generated videos against

Figure 3: Input sad frame (left), generated happy outputs (grid center), and target ground truth happy frame (right). In the grid the first row is *Full* masking while the second row is *Half* masking. The columns are emotion modification strategies left to right as follows: *L1 + Emotion Objective, L1, Emotion Objective.*

changes in baseline human performances in MEAD. We also provide qualitative examples (see Figure 3) and a user study with findings that corroborate our automated emotion evaluation approach.

4.1 Wav2Lip Retraining

masking	LSE-D↓	LSE-C↑	FID↓
full	7.640	6.213	5.938
half	7.013	7.029	7.818
original model			
Wav2Lip + GAN	**6.469**	**7.781**	**4.446**

Table 2: Lip sync (*LSE-D, LSE-C*) and visual quality (*FID*) for re-trained Wav2Lip with modified visual quality discriminator necessary for emotion modification, reported on LRS2 dataset compared with original Wav2Lip results.

Since we redesign the architecture of the visual quality discriminator and introduce a new full face masking strategy in one of our variants, it was necessary to retrain Wav2Lip on the LRS2 data used by the original authors. For our **full** masking variants, we retrained Wav2Lip from scratch following the same procedure as the original authors except for changing the masking technique and the visual quality discriminator. For our **half** masking variants, we were able to start training from checkpoints provided by the Wav2Lip authors before the introduction of the visual quality discriminator objective.

In Table 2 we report the results of retraining Wav2Lip on LRS2 prior to fine-tuning for emotion modification. We give the three metrics employed by the original authors: lip synchronization error distance (**LSE-D**), lip synchronization error confidence (**LSE-C**), and Fréchet inception distance (**FID**).

LSE-D and LSE-C utilize SyncNet [5], the original architecture modified by Wav2Lip for their lip synchronization discriminator. The originally reported accuracy of this model is over 99%. The LSE-D measures the L2 distance between the SyncNet encodings of audio and video, where close vectors are trained to represent synchronization. LSE-C on the other hand represents the SyncNet confidence output, where a larger number indicates better synchronization.

FID [12] is a commonly used automatic measure of distance between image datasets. Feature representations of the images in the dataset are encoded by the pre-trained Inception network and then the Fréchet distance is calculated between two Gaussians fitted to these representations. Wav2Lip takes the FID of generated output frames against ground truth frames as a measure of overall visual quality, where a lower value indicates higher visual quality from greater fidelity to the distribution of ground truth frames. Same as Wav2Lip, we utilize the FID implementation by [29].

While our retrained Wav2Lip models do not achieve identical performance with the original, they are nevertheless sufficiently comparable given that our objective is simply to maintain lip synchronization rather than alter it.

Also notably the **half** masking variant comes out of retraining more performant than the **full** masking variant, likely because it was able to leverage the already high performance of the original authors *Wav2Lip* checkpoint as a starting point rather than being trained from scratch. Initial attempts to re-train the **full** mask variant suffered from vanishing gradients in the discriminator. Attempts to introduce batch normalization greatly harmed discriminator performance, so we instead opted to add residual skip connections to the discriminator architecture. From there we trained all models until convergence.

4.2 Qualitative Observations

Figure 3 provides examples of generated frames from all of our model variants on sad to happy emotion modification based on inputs from our test split of the MEAD data. The input sad image frame is shown on the left, the happy generated outputs are provided for each of the 6 model variations in the grid in the center, and a frame from the target ground truth happy performance is provided on the right for contrast. In the grid the top row is **Full** masking while the bottom row is **Half** masking. The columns are emotion modification strategies left to right as follows: **L1 + Emotion Objective**, **L1**, **Emotion Objective**.

masking	emotion modification strategy	LSE-D↓	LSE-C↑	FID↓	Δ valence ↑	Δ arousal ↑
full	L1 + Emotion Objective	11.357	1.378	103.016	0.729	0.400
full	Emotion Objective	10.993	1.855	73.046	0.044	0.200
full	L1	11.413	1.477	141.719	0.995	**0.803**
half	L1 + Emotion Objective	10.563	2.101	71.508	**1.000**	0.522
half	Emotion Objective	**10.443**	**2.328**	**48.819**	0.095	0.244
half	L1	10.673	2.000	82.709	0.945	0.595
Average Wav2Lip Original		10.032	2.751	31.813	-	-
Average Ground Truth		10.651	2.533	26.302	-	-

Table 3: Comparison averaged across emotions of Wav2Lip-Emotion variants and ground truth on lip sync (*LSE-D, LSE-C*), visual quality (*FID*), and our novel automated emotion evaluation approach that normalizes change in valence and arousal by a human baseline. Original model and ground truth shown for contrast, with FID averaged over emotion combinations.

Analysis of a selection of examples seems to indicate that the **L1 + Emotion Objective** emotion modification strategy has issues with blurry outputs but performs the task most faithfully by modifying the emotion while preserving pose. The **L1** emotion modification strategy on the other hand produces more crisp results with clear emotion modification but does a poor job of preserving pose and visual quality. Finally, **Emotion Objective** emotion modification does an excellent job of preserving pose and sync, but produces very little emotion modification.

4.3 Automatic Evaluations

We report automated evaluation results for Wav2Lip-Emotion variants in Table 3. We follow Wav2Lip's utilization of pre-trained automated evaluation metrics (**LSE-D**, **LSE-C**, and **FID**, explained in Section 4.1) and craft an approach for automatic emotion evaluation (Δ **valence** and Δ **arousal**).

To measure valence and arousal in videos we use the NISL2020 model [7], the winner of FG-2020's Competition in Affective Behavior Analysis in-the-wild (ABAW) [19]. Trained on the Aff-Wild2 dataset [20], this model outputs per-frame valence scores, ranging from [-1,1], that indicate how positive or negative a facial expression is. It also outputs per-frame arousal scores, ranging from [0,1], that indicate how active or calm the expression is. The model incorporates information from all the frames of a video in which a face is detected, and thus assesses the consistency of our generated videos over a temporal window of frames. We take the average of each value over all frames in a video to get a video level score.

In order to make our emotion metric comparable across different pairs of source and destination emotions, we devise the following normalization scheme. Our test split of the MEAD dataset contains performed *destination emotion* target videos associated with the *source emotion* input videos that can be used as a baseline for our evaluation. Thus we get average valence and arousal scores over all frames of an actor performing a given utterance for the generated outputs (v_g and a_g), *source* ground truth emotion (v_s and a_s), and *destination* ground truth emotion (v_d and a_d). We then take the ratio of the change in the generated video compared to its *source emotion* input against the change in the *destination emotion* video compared to the *source emotion* video:

$$\Delta \text{ valence} = \frac{v_g - v_s}{v_d - v_s} \tag{9}$$

$$\Delta \text{ arousal} = \frac{a_g - a_s}{a_d - a_s} \tag{10}$$

Thus intuitively positive values indicate change in the "right" direction, with a value of 1 indicating a change in emotion identical to the ground truth change performed by the actors in the MEAD dataset. Meanwhile emotion pairs like sad and neutral, instead of sad and happy, will have smaller ground truth changes and will be scaled so as not to be drowned out when aggregated with pairs that have larger ground truth shifts. Finally, scores greater than 1 are possible. While overshooting the valence change towards happy or sad emotion is desirable, overshooting the valance modification towards neutral is not. Thus, we further take $\Delta \text{ valence}_{*2n} = 1 - |1 - \Delta \text{ valence}|$ as our normalized score for neutral *destination* modifications to penalize overshooting.

The numbers presented in Table 3 are the averaged values over all videos in each condition, further micro-averaged over all 6 emotion pairs of {sad, neutral, happy}. This reveals that no one model variant excels in all cases. Unsurprisingly the **half** masking with **Emotion Objective** modification strategy variant performs best on lip synchronization and image quality metrics as this model is closest to the original Wav2Lip approach. It takes advantage of the half masked model checkpoints provided by the authors and needs to retrain only the visual quality discriminator unlike our **full** masking models which also retrain the generator. This model also achieves better metrics on the LRS2 as shown in Table 2. However, this model performs relatively poorly on our automated emotion evaluation.

The best performing models on the emotion evaluation appear to come from the **L1** and **L1 + Emotion Objective** emotion modification strategy variants rather the **Emotion Objective** variants. This indicates that the L1 reconstruction loss along with the novel use of *destination emotion* videos for target frames produces most emotion modification, while only a small amount of modification is achieved with the emotion objective alone. Nevertheless the **Emotion Objective** only variants do still move emotion (particularly arousal) in the right direction and achieve better synchronization and visual quality metrics. The **L1 + Emotion Objective** variant

	emotion modification	LSE-D↓	LSE-C↑	FID↓	$v_d - v_s$	$a_d - a_s$	Δ valence ↑	Δ arousal ↑
All Variants	happy -> neutral	11.543	1.52	95.127	-0.746	-0.213	0.649	0.767
	happy -> sad	10.715	1.986	77.768	-0.807	-0.069	**2.143**	0.121
	neutral -> happy	**10.56**	**2.277**	93.241	0.74	0.212	0.66	**0.808**
	neutral -> sad	10.774	2.113	98.442	-0.077	0.155	0.185	0.244
	sad -> happy	10.82	1.676	**65.007**	0.807	0.069	0.663	0.377
	sad -> neutral	11.03	1.566	91.232	0.084	-0.155	-0.486	0.427
Best Variant	happy -> neutral	11.017	1.757	72.889	-0.746	-0.213	0.661	0.826
	happy -> sad	**10.117**	2.312	77.642	-0.807	-0.069	**3.537**	0.455
	neutral -> happy	10.462	2.449	67.377	0.746	0.213	1.063	**1.032**
	neutral -> sad	10.375	**2.576**	69.902	-0.077	0.155	0.383	0.441
	sad -> happy	10.489	1.838	**68.397**	0.807	0.069	1.246	-0.444
	sad -> neutral	10.918	1.674	72.839	0.077	-0.155	-0.888	0.825

Table 4: Comparison over pairs of *source* and *destination* emotions for lip sync (*LSE-D, LSE-C*), visual quality (*FID*), and change of *valence* and *arousal* in generated outputs normalized by human baseline change between ground truth inputs and targets ($v_d - v_s$ and $a_d - a_s$). Top section averaged over Wav2Lip-Emotion variants, and bottom section best all-around variant.

that utilizes both approaches, appears to provide a balance between synchronization and visual quality against emotion modification.

We contrast our model performances on the synchronization and visual quality metrics against those same metrics run on the ground truth data as well as the generated outputs of the original Wav2Lip run on the synchronized ground truth video and audio. **LSE-D** and **LSE-C** numbers on the ground truth and Wav2Lip generated happy, neutral, and sad videos are averaged over emotions to produce the numbers at the bottom of Table 3. The **FID** scores meanwhile are the average over the FID between all combinations of the three emotions. While our models' synchronization and visual quality performance are worse than that of Wav2Lip on the LRS2 (see Table 2), our metrics are better aligned for Wav2Lip's outputs on MEAD as well as the metrics on the unmodified ground truth itself. This discrepancy may arise from slight audio desynchronization in the raw MEAD data. Meanwhile the use of FID to judge visual quality may be better suited for lip synchronization than emotion modification, since the latter makes much more salient changes to the images in the videos. Evidently even real differences in performed emotion made by the same speaker in the MEAD ground truth data shift the distributions of Inception encodings more than Wav2Lip's generated results on LRS2. The FID between original Wav2Lip outputs on MEAD and ground truth frames from a distinct emotion is somewhat higher than the FID between ground truth emotions only, suggesting that the FID nevertheless measures some additional differences in visual quality produced by video translation.

We present metrics separately for each emotion modification pair of *source* and *destination emotions* in Table 4. The top section is averaged over all model variants, while the bottom section presents the numbers for the best all around performing model, **half** masking with **L1 + Emotion Objective** emotion modification strategy. We also report the baseline change in valence and arousal between ground truth inputs and targets ($v_d - v_s$ and $a_d - a_s$) to reveal the NISL2020 model's sensitivities to changes in each pair of emotions as performed by humans. Notably the difference between neutral and sad appears to manifest more as a change in arousal than as a change in valence. These baseline changes also demonstrates how the normalized Δ valence and Δ arousal are sensitive to noise in generated changes when there are relatively small changes of the ground truth ($v_d - v_s$ and $a_d - a_s$) in the denominator, possibly explaining erratic scores for arousal change between sad and happy and valence change between sad and neutral.

More generally Table 4 shows how emotion results are highly dependent on the source and destination emotion while synchronization and visual quality are relatively unaffected. Some *destination emotions* may be easier to optimize for or some input *source emotions* may be more difficult to alter. Notably the Δ valence scores for happy to sad are much higher than all others including sad to happy, possibly indicating that the model is able to exploit traits of sadness as it appears in the training data more effectively than other emotions. Nevertheless the best all around performing model gets Δ valence and Δ arousal values around 1.0 on many emotion pairs, indicating a general ability to modify emotion comparably to the ground truth changes performed by the actors in the MEAD dataset.

4.4 User Study

We conduct a small crowdworker evaluation to rate the clips generated by the top 3 models on their visual quality, lip synchronization, and emotion. We randomly select 2 videos per emotion pair for each of the 3 best performing models shown in the top half of Table 3, giving us 2 * 6 * 3 or 36 videos overall. We create a simple form, that asks 20 workers to rate each of these 36 videos on their visual quality and lip synchronization aspects, and to judge the emotion expressed by the speaker in the video. We select 20 workers from the Prolific platform [23], who are fluent in English, have participated in at least 10 tasks on Prolific, and have a 95% acceptance score on earlier tasks. For the first two factors, we ask workers to rank the generated videos between 1 and 5, where 1 stands for "Very bad lip synchronization" and "Very bad visual quality" and 5 stands for "Excellent lip synchronization" and "Excellent visual quality". To judge the emotion expressed in the generated video,

masking	emotion modification strategy	sync ↑	visual quality ↑	Δ emotion ↑
half	L1 + Emotion Objective	2.467	2.138	**0.779**
half	Emotion Objective	**3.492**	**2.887**	0.379
half	L1	2.758	1.85	0.732
	emotion modification	**sync ↑**	**visual quality ↑**	**Δ emotion ↑**
	happy -> neutral	2.483	2.117	**0.833**
	happy -> sad	2.35	1.967	0.569
	neutral -> happy	**3.483**	**2.933**	0.454
	neutral -> sad	2.95	2.183	0.533
	sad -> happy	3.333	2.167	0.621
	sad -> neutral	2.833	2.383	0.771

Table 5: User score from 1-5 over *Sync* and *Visual quality*, as well as Δ *emotion* of generated outputs normalized by expected change. Top half averaged over 3 best variants and bottom half averaged over emotions.

workers scored a video between 1 and 5 where a score of 1 meant that the speaker in the video is "Very Sad", a score of 3 meant that the speaker in the video is "Neutral" and a score of 5 meant that the speaker in the video is "Very Happy"

In order to make our emotion metric comparable across different pairs of source and destination emotions, we utilize a similar normalization scheme as on our automated emotion evaluations. The raw per-video emotion scores for generated outputs (e_g) are normalized with respect to the ground truth emotion values of the *source* and *destination* emotions (e_s and e_d). So for example a modification from neutral to happy would have $e_s = 3$ and $e_d = 5$. We then normalize the raw scores as follows:

$$\Delta \text{ emotion} = \frac{e_g - e_s}{e_d - e_s} \tag{11}$$

Thus positive values indicate change in the correct direction. And again, we further take $\Delta \text{ emotion}_{*2n} = 1 - |1 - \Delta \text{ emotion}|$ as our normalized score for neutral *destination* modifications to penalize overshooting.

The user study results aggregated over model variants (top of Table 5) corroborate findings of our novel automatic emotion evaluation by indicating that **L1** and **L1 + Emotion Objective** are most effective at emotion modification but struggle with synchronization and visual quality. The per-emotion pair results (bottom of Table 5), show that we achieve meaningful emotion modification across emotion pairs along with moderate synchronization and visual quality.

5 CONCLUSION

To our knowledge, Wav2Lip-Emotion is the first approach to cast the synthesis of facial emotion as a video-to-video translation task. Our method extends an existing lip synchronization model, Wav2Lip, with a new task of modifying facial emotion in translated images through L1 reconstruction and pre-trained emotion objectives. In order compare results from all model variants on the MEAD dataset, we propose a novel automatic emotion evaluation approach and corroborate it with a user study.

Our evaluations support our ability to modify emotion, with both automatic metrics and human judgements rating the emotion modification in our best performing models as nearly comparable to the ground truth changes performed by the actors in the MEAD dataset. Our model also appears to maintain the level of lip synchronization present in the input videos, as was the original intent of building on the Wav2Lip architecture. However the visual quality of our models is only moderate. Greater performance may be possible from training on additional data with more variety of posses and speaker identities. While we have taken advantage of the paired emotion videos in the MEAD dataset for emotion evaluation, Wav2Lip-Emotion does not require such paired data. It can train on any videos in which a single speaker is found performing differing emotions. These could be labeled, as in the CMU-MOSEI dataset [1], or automatically inferred. Likewise while the present work has been limited to individual models for translating from a specific emotion to another one, our approach makes no fundamental obstacle to a multi-task approach that translates many emotions. By utilizing additional intensity labels, it may even be possible to dial the level of emotion modification through a hyperparameter, similarly to the approach of Fan et al. [10]. Such possibilities suggest the further promise of this proof of concept of video-to-video translation for modifying facial emotion.

ACKNOWLEDGMENTS

This research was made possible by the generous support by the MIT Media Lab Consortium and the CISCO gift to Research in Telecreativity. We are grateful to Dr. William T. Freeman and Dr. Phillip Isola for their guidance and advice on this project. We also thank the anonymous ADGD 2021 reviewers for their insightful and helpful comments.

REFERENCES

[1] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 2236–2246. https://doi.org/10.18653/v1/P18-1208

[2] Julianne Gold Brunson and P Scott Lawrence. 2002. Impact of sign language interpreter and therapist moods on deaf recipient mood. *Professional Psychology: Research and Practice* 33, 6 (2002), 576.

[3] Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. 2019. Photo-Realistic Facial Details Synthesis From Single Image. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision (ICCV).

[4] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip Reading Sentences in the Wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3444–3453. https://doi.org/10.1109/CVPR.2017.367

[5] Joon Son Chung and Andrew Zisserman. 2016. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*.

[6] codeniko. 2019. 81 Facial Landmarks Shape Predictor. https://github.com/codeniko/shape_predictor_81_face_landmarks.

[7] D. Deng, Z. Chen, and B. E. Shi. 2020. Multitask Emotion Recognition with Incomplete Labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG)*. IEEE Computer Society, Los Alamitos, CA, USA, 592–599. https://doi.org/10.1109/FG47880.2020.00131

[8] Joseph R Dusseldorp, Diego L Guarin, Martinus M van Veen, Nate Jowett, and Tessa A Hadlock. 2019. In the eye of the beholder: changes in perceived emotion expression after smile reanimation. *Plastic and reconstructive surgery* 144, 2 (2019), 457–471.

[9] Paul Ekman. 1993. Facial expression and emotion. *American psychologist* 48, 4 (1993), 384.

[10] Lijie Fan, Wenbing Huang, Chuang Gan, Junzhou Huang, and Boqing Gong. 2019. Controllable Image-to-Video Translation: A Case Study on Facial Expression Generation. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (Jul. 2019), 3510–3517. https://doi.org/10.1609/aaai.v33i01.33013510

[11] Panagiotis Giannopoulos, Isidoros Perikos, and Ioannis Hatzilygeroudis. 2018. *Deep Learning Approaches for Facial Emotion Recognition: A Case Study on FER-2013*. Springer International Publishing, Cham, 1–16. https://doi.org/10.1007/978-3-319-66790-4_1

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2018. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. arXiv:1706.08500 [cs.LG]

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2261–2269. https://doi.org/10.1109/CVPR.2017.243

[14] Carroll E. Izard. 1990. Facial expressions and the regulation of emotions. *Journal of Personality and Social Psychology* 58, 3 (1990), 487–498. https://doi.org/10.1037/0022-3514.58.3.487

[15] Jerome Kagan, Nancy Snidman, and Doreen Arcus. 1993. On the temperamental categories of inhibited and uninhibited children. *Social withdrawal, inhibition, and shyness in childhood* (1993), 19–28.

[16] Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4396–4405. https://doi.org/10.1109/CVPR.2019.00453

[17] Caroline F Keating, Allan Mazur, and Marshall H Segall. 1977. Facial gestures which influence the perception of status. *Sociometry* (1977), 374–378.

[18] Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (2009), 1755–1758.

[19] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. [n.d.]. Analysing Affective Behavior in the First ABAW 2020 Competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*. 794–800.

[20] Dimitrios Kollias and Stefanos Zafeiriou. 2018. Aff-Wild2: Extending the Aff-Wild Database for Affect Recognition. *CoRR* abs/1811.07770 (2018). arXiv:1811.07770 http://arxiv.org/abs/1811.07770

[21] Andrea Miller, Renita Coleman, and Donald Granberg. 2007. TV Anchors, Elections & Bias: A Longitudinal Study of the Facial Expressions of Brokaw Rather Jennings. *Visual Communication Quarterly* 14, 4 (2007), 244–257. https://doi.org/10.1080/15551390701730232 arXiv:https://doi.org/10.1080/15551390701730232

[22] Dmitry Nikitko. 2019. StyleGAN — Encoder. https://github.com/Puzer/stylegan-encoder.

[23] Stefan Palan and Christian Schitter. 2018. Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27.

[24] Sandra Passardi, Peter Peyk, Michael Rufer, Tanja S. H. Wingenbach, and Monique C. Pfaltz. 2019. Facial mimicry, facial emotion recognition and alexithymia in post-traumatic stress disorder. *Behaviour Research and Therapy* 122 (2019), 103436. https://doi.org/10.1016/j.brat.2019.103436

[25] Luan Pham, The Huynh Vu, and Tuan Anh Tran. 2021. Facial Expression Recognition Using Residual Masking Network. In *2020 25th International Conference on Pattern Recognition (ICPR)*. 4513–4519. https://doi.org/10.1109/ICPR48806.2021.9411919

[26] Rosalind W Picard. 2000. *Affective computing*. MIT press.

[27] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2020. 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* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 484–492. https://doi.org/10.1145/3394171.3413532

[28] Albert Pumarola, Antonio Agudo, Aleix M. Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2018. GANimation: Anatomically-Aware Facial Animation from a Single Image. In *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer International Publishing, Cham, 835–851.

[29] Maximilian Seitzer. 2020. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid. Version 0.1.1.

[30] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/31c0b36aef265d9221af80\protect\penalty\z@872ceb62f9-Paper.pdf

[31] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. 2020. MEAD: A Large-Scale Audio-Visual Dataset for Emotional Talking-Face Generation. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 700–717.

[32] Xinwei Yao, Ohad Fried, Kayvon Fatahalian, and Maneesh Agrawala. 2021. Iterative Text-Based Editing of Talking-Heads Using Neural Retargeting. *ACM Trans. Graph.* 40, 3, Article 20 (Aug. 2021), 14 pages. https://doi.org/10.1145/3449063