Imperial College London

Department of Computing

Background and Progress Report

Author:
Peter Robertson

Supervisor:
Dr Stefanos Zafeiriou

Submitted in partial fulfilment of the requirements for the MSc degree in Computing Science / Machine Learning of Imperial College London

Contents

1 Introduction

Visual speech prediction, often referred to as lip reading, is a very difficult task, partly due to ambiguities between phonemes such as '*p*' and '*b*', which look the same but sound different. Current methods of addressing this problem focus on training deep learning models using 2D temporal data of people speaking with the relevant video frames labelled with the correct spoken text [1, 2, 3, 4]. There currently do not exist any models which attempt to address this problem with the use of 3D temporal datasets made up of head scans of subjects, capturing depth and thus, more information of the mouth which may be of use to the models. However, there currently exists a severe lack of availability of such datasets due to the complexity and hardware requirements in capturing such data. Currently their exists two publicly available datasets: LRW-3D [5] and the VOCASET [6], both of which are relatively small in comparison to the models used in current lip reading models. To first attempt this problem, new datasets which capture 3D temporal models of subjects heads speaking must be established by the community. As well as capturing new data, expanding existing datasets with generated synthetic data with the use of generative adversarial networks may assist in the solution to this problem.

2 Lip Reading

This section shall discuss current lip reading models built with deep learning techniques and the availability of both 2D and 3D temporal datasets which lip reading models can currently be trained on. In recent years the problem of Lip Reading has seen huge advances due to the availability of new datasets and the use of Deep Learning models. Due to the large availability of video data, datasets such as GRID [7], LRW [1], LRS [3] and LSVSR [4] have been constructed by compiling the appropriate data.

Current lip reading models currently make use of 2D temporal data such as videos. This is convenient as video data is already widely available and relatively straightforward to obtain and process. Unlike 2D temporal data, 3D temporal data (3D video), requires multiple cameras to record simultaneously, which requires synchronisation, adding to the complexity of the system. The data must then be processed to produce the final product, whether this is in the form of video with depth information or more complex 3D scans [8].

2.1 Datasets

2.1.1 Controlled Conditions Datasets

The GRID dataset was released in 2006 [7] and contains 34,000 samples from 34 speakers, each with 1000 sentences. The corpus uses sentences with a fixed grammar: <command:4>, <colour:4>, <preposition:4>, <letter:25>, <number:10>, <adverb:4>, with a total vocabulary of 51 words. The primary limiting factor of the GRID dataset is that all data was captured directly for the use of the dataset.

2.1.2 In The Wild Datasets

To build larger datasets, subsequent datasets are commonly built with data "in the wild", meaning that the videos have not been captured with the intention of being used for this dataset. Variations in lighting, angles, speakers and a wide vocabulary are common, this does however make the datasets more challenging to learn from. LRW and LRS are both comprised of content from BBC broadcasts [1, 3] allowing for a far larger corpus size of over 1000 words for LRW and over 6000 words for LRS. Both of datasets are captured and processed with the same pipeline summarised as follows. Firstly, as the subtitles are not aligned to the video on the broadcasts, optical chapter recognition is used to obtain the text being spoken in the video clips, the audio and text are then aligned per frame using the HTK toolkit [9]. A HOG-based detection algorithm [10] is used for face detection for cropping the frames to the subjects head followed by facial landmarks for mouth localisation and speaker identification.

LSVSR is a dataset published by DeepMind and Google which makes use of the huge amount of videos on YouTube [4], resulting in a total length of 3886 hours of training data. The dataset is far larger than LRS, but aligns phonemes to frames, as opposed to words or characters. The pre-processing steps are similar as in LRW and LRS; alignment is performed with the algorithm laid out in previous work by DeepMind [11], faces are tracked to ensure the speaker is visible in frame.

2.2 3D Datasets

To the best of the author's knowledge, there are currently two datasets with 3D temporal data which are appropriate for training lip reading models. The first of which is LRW-3D [5] which has been captured from four subjects, two native English speakers and two non-native to increase variability in the dataset. The subjects have been captured speaking the corpus used in the LRW dataset [1], a vocabulary of 500 words. The resulting dataset comprises of 660 seconds of 3D meshes and audio per subject. The dataset is not comprised of full sentences but would be appropriate for word-level lip reading.

The second dataset is the VOCASET [6], captured from 6 male and 6 female subjects. Each subject was recorded speaking 40 sequences, each ranging from 3 to 5 seconds, resulting in a total time of 30 minutes. The recorded 3D meshes are registered to the FLAME model [8], a statistical 3D facial mesh with around 5000 vertices. Unlike the LRW-3D dataset, the sequences are grammatically correct sentences, chosen to maximise phonetic diversity. This makes the VOCASET appropriate for creating a model for sentence-level lip reading.

It should be noted that neither of these datasets were captured for the purpose of lip reading, but for synthesising realistic statistical facial models driven from an audio input, and thus the transcriptions and frames are not aligned. This alignment would have to be performed to the datasets before using the data to train lip reading models. In order to achieve this automatic speech recognition (ASR) systems such as DeepSpeech [12] could be used to solve this problem.

2.3 Lip Reading Models

Chung et al [1] used a convolutional model based on the VGG-M architecture which produced character level distributions with the LRW dataset, these distributions are then processed by a language model for text prediction.

The LipNet model [2] was produced to be able to predict sentence-level lip reading of varied length, while previous work by [1] predicted on a word level. The model used spatiotemporal convolutions to process multiple frames of video at once followed by a recurrent layer using Gated Recurrent Units (GRU) [13]. The model used the GRID dataset [7] in which all subjects are recorded under consistent conditions of good lighting and fixed angles. LipNet achieved the state of the art performance on the GRID dataset by using an model with increased complexity by incorporating 3D convolution and recurrent units.

As the LRW dataset could not be used to train a model such as LipNet for sentence-level lip reading, but the GRID dataset had limitations in the number of subjects and vocabulary, Chung et al created the LRS dataset [3]. The model presented was trained on both audio and video and made capable of taking either or both as the model inputs. To prevent the model from being dependent on a single input source, the inputs are systematically distorted or removed. The video input is passed through convolutional layers, followed by LSTM layers, while the audio is converted to MFCC, then input to LSTM layers. The two are combined with then attention mechanism and further LSTM layers and an output fully connected layer with softmax activation for character distributions. The model also makes use of curriculum learning by initially training the model on short sequences of single words. The length of training sequences are increased throughout training. It is stated by [3] that this accelerates training and reduces overfitting.

In 2018 DeepMind published their V2P model [4] along with the LSVSR dataset which is larger than all previous datasets, containing 3886 hours of training data. The model follows a similar architecture to LipNet [2] but with an increased number of convolutional layers and LSTM layers rather than GRUs. It should be commented that due to the size of the model and dataset dictated the use of 64 GPUs for training to allow a batch size of 128. Unlike previous models, the V2P model predicts phonemes as opposed to characters, these phonemes are then processed by a language model for word prediction as with previous models.

To the best of the author's knowledge, currently there do not exist any papers which have explored the use of 3D temporal datasets for the use with lip reading models. This is likely due to the shortage of 3D temporal data on which such models could be trained on.

3 3D Facial Modelling

Unlike 2D video where the subject is captured from a single angle without any depth information, 3D models cannot be captured with a single camera. In most instances, 3D head scans are captured with the use of a multi-camera rig with multiple cameras capturing video simultaneously. The number and type of camera can vary from system to system, but the principle remains consistent. All cameras must be synchronised to capture footage at the same time, with the same frame rate. This is inherently a more complex

system than 2D video capture as is therefore more expensive, limiting the accessibility of such systems. Once the images have been captured, the footage must be processed in order to produce a facial mesh of the subject [8].

3.1 Blendshapes

A facial mesh is made up of a point cloud of vertices and vertex connections. Depending on the mesh capture pipeline, the number of vertices may vary, but even a low resolution mesh may contain several thousand vertices. Capturing realistic facial motion by modelling the displacements of all of these vertices is impractical when manipulating a model by hand. Blendshapes attempt to model aspects of realistic facial motion by finding the relations between these vertices and manipulating them simultaneously [14], by reducing the number of parameters, the model becomes more manageable to control. One method of producing model blendshapes is with principle component analysis (PCA). By applying PCA on a set of facial meshes, the resulting principle components represent the changes in facial motion which capture the largest amount of variation in the set while reducing the number of model parameters substantial.

4 Data Generation

In order to construct a deep learning lip reading model capable of being trained on 3D temporal data, appropriate datasets must be established. Current datasets have been captured directly [5, 6] with the use of multi-camera capture rigs under controlled situations. The total duration of both of these datasets is very short in comparison to the video datasets such as LRW, LRS and LSVSR, which is a limiting factor as to the models which could be trained using them. As the models also use different mesh models to represent the data that has been captured this also prevents the two datasets being joined directly. Unlike video data, there currently lacks a large body of 3D video data which is publicly available, limiting the construction of 3D datasets to directly capturing more 3D scans with multi-camera capture rigs, similar to those used in [5, 6] and generating synthetic training data.

4.1 Audio Driven Data Generation

Karras et al proposed a method for generating 3D facial animation from audio with the use of a CNN architecture [15]. The model is actor specific, but is only trained on 3-5 minutes of data of the actor. Short range temporal features are first extracted from the audio by a formant analysis network. This representation is then analysed by an articulation network which also accepts a learned emotional state of the speaker. The output layer drives displacements from a neutral 3d mesh of the actor.

However, as the model by Karras et al. is not independent of the actor it cannot generalise to new subjects. Tzirakis et al. propose a model which is independent of speaker and capture rig [5]. As discussed in section 2.2, a dataset was constructed of 3D speaking faces using the LRW dataset [1] for the corpus. This allowed Tzirakis et al. to create a model which can synthesise facial motion from audio from the LRW dataset. The model used is similar to that used in [15], firstly extracting short term temporal features from

the input audio with a convolutional network followed by another convolutional network to analyse the extracted features. Unlike the model used in [15], the model used in [5] is trained to drive learnt blendshapes. This reduces the number of output parameters of the model substantially in comparison to that used by Karras et al.

Similar to [5], the VOCA model [6] synthesises video sequences of 3D models speaking given an audio input. The VOCASET dataset discussed in section 2.2, was captured with the intention of training this model to be independent of the speaker, hence a large range of speakers are used within the dataset. The model is comprised of three sections: audio feature extraction, a feature encoder and a decoder to drive a template facial mesh from the FLAME model [8]. The audio feature extraction makes use of the pre-trained Mozilla implementation of the DeepSpeech model, based on the paper by Hannun et al. [12]. The DeepSpeech model takes audio as an input and returns the unnormalised log-probabilities for an alphabet of the 26 standard characters, a space, apostrophe and blank character for time slices in the audio input. The encoder is a convolutional network which is conditioned on the speakers identity, such that the latent space of speaker styles can later be explored on new audio inputs. Finally, the decoder is made up of a fully connected layer with a linear activation function is used to output the displacements of the 5023 vertices in the template face.

4.2 Adversarial Trained Generative Methods

A recent development in generating synthetic data samples is the use of generative adversarial networks [16]. The concept behind generative adversarial networks (GANs) is two have two machine learning models; a generator and a discriminator. The task of the generator is to produce samples from an unknown high order probability distribution which correctly resemble samples from the distribution defined by training data. The generator achieves this by transforming a random sample from a known probability distribution, such as a Gaussian distribution as used in [16], to a sample from the unknown distribution. This is achieved by finding the function which maps between the two distributions. The discriminator however, attempts to correctly learn to discriminate between the real and the fake generated samples.

$$\min_G \max_D V(G, D) = \mathbb{E}_{\boldsymbol{x} \sim p_{data}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p_z(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))] \tag{1}$$

The original loss function (1) proposed by Ian Goodfellow forms a min-max game, where the loss of the generator is attempting to be minimised by having the discriminator label all the generated samples as real. While the loss of the discriminator is maximised by correctly classifying real and fake samples. The two networks are trained in an alternating fashion until the discriminator achieves an accuracy of 50%, effectively making binary guesses between real and generated samples.

GANs however, are difficult to train for two main reasons. Firstly, the equation (1) is challenging as it provides small gradients while generated samples are poor as discussed in [16], making training difficult. Progress in developing new loss functions is discussed in section 4.2.1 Secondly, early networks must also be balanced with a similar model capacity to prevent one from getting too much better than the other, preventing the other from

improving. Various architectural changes have improved this issue [17, 18], although it seems to be closely tied to the loss function being used [19].

4.2.1 Stability Improvements of GANs Networks

There have been a large number of papers presenting new techniques with differing levels of success and training stability, a small handful of key papers which provided large advances in the generative adversarial training model shall be discussed here. A primary research focus around GANs has been in finding new loss functions on which to train the model to improve stability and performance. The original loss function proposed in the original paper [16] identifies an issue with equation (1) in that the gradient back propagated to the generator when the generated samples are poor, is very low as shown in figure 1.

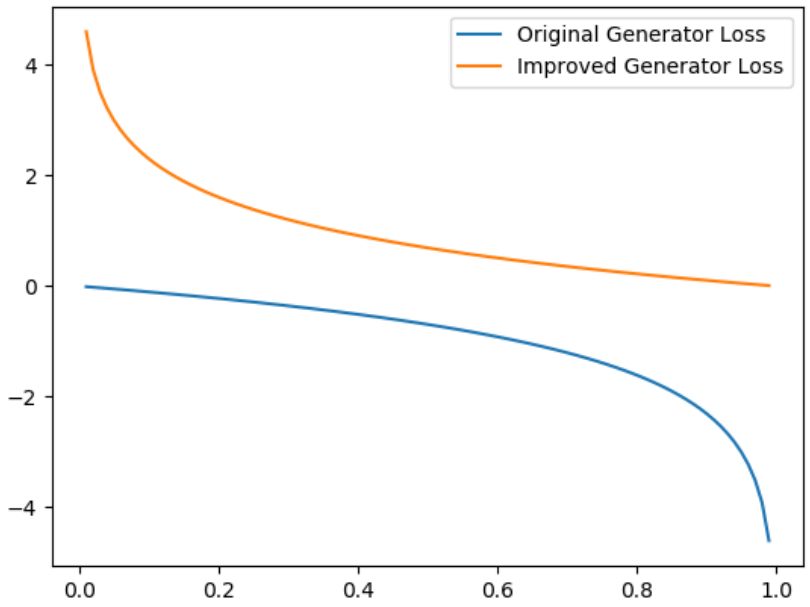

Figure 1: Goodfellow's Generator Loss Plots [16]

This in turn makes it very challenging to improve the performance of the generator. The suggested improvement made in [16] is rather than to maximise the the number of generated samples which the discriminator incorrectly classifies, but to minimise the number of generator samples the discriminator correctly classifies, described in equation (2). Figure 1 shows how this results in the gradient propagated back to the generator is far larger when the generated samples are poor.

$$\min_G \max_D V(G, D) = \mathbb{E}_{\boldsymbol{x} \sim p_{data}(\boldsymbol{x})}[\log D(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{z} \sim p_z(\boldsymbol{z})}[\log(D(G(\boldsymbol{z})))] \tag{2}$$

Further stability improvements were proposed in [17] which allowed deep convolutional generative adversarial networks (DCGANs) to be successfully trained for the first time. Radford et al. proposed three main contributions. Replace deterministic pooling layers with strided convolutions in both the generator and discriminator networks, allowing the networks to learn their own spatial upsampling and downsampling. Remove all fully connected layers used on top of convolutional layers, resulting in a fully convolutional model. Apply batch normalisation [20] before the input of each layer. This normalises the input to each unit to zero mean and unit variance. This assists with training problems due to poor weight initialisation and allows gradients to flow through deeper networks more easily. Radford et al. state this to be a critical improvement to allow generator networks to begin learning by preventing all samples from collapsing to a single point.

In an attempt to stabilize the training of GANs models, the Wasserstein or 'Earth Mover Distance' loss function was proposed by Arjovsky et al. [21] shown in equation (3) where $\mathcal{D}$ is a set of 1-Lipschtiz functions.

$$\min_G \max_{D \in \mathcal{D}} W(\mathbb{P}_r, \mathbb{P}_g) = \mathbb{E}_{\boldsymbol{x} \sim \mathbb{P}_r}[D(\boldsymbol{x})] - \mathbb{E}_{\hat{\boldsymbol{x}} \sim \mathbb{P}_g}[D(\hat{\boldsymbol{x}})] \tag{3}$$

The WGAN model uses a critic as opposed to a discriminator, this is due to the fact that the discriminator is no longer a binary classifier, but being used to critique the real and generated samples. As the Wasserstein function is continuous and differentiable, the critic can be trained until optimality, and [21] argues that it should be. As the critic is trained to optimality it does not saturate, but converges to a linear function. This provides the generator with a gradient which is more reliable, resulting in the generator learning consistently. The Wasserstein loss function has been shown to greatly improve training stability by providing consistent gradients throughout training and no longer requiring that the two networks have a balanced model capacity.

In order to use the Wasserstein distance as the loss for WGAN, the Lipschtiz condition must be enforced. In the WGAN the model weights are clipped to enforce this constraint [21], however this is stated to be a non-ideal method of achieving this constraint and calls for future work to be investigate more effective methods. The use of gradient penalty is proposed in [19] in order to satisfy this condition more elegantly as described in equation (4).

$$\min_G \max_{D \in \mathcal{D}} W(\mathbb{P}_r, \mathbb{P}_g) = \mathbb{E}_{\boldsymbol{x} \sim \mathbb{P}_r}[D(\boldsymbol{x})] - \mathbb{E}_{\tilde{\boldsymbol{x}} \sim \mathbb{P}_g}[D(\tilde{\boldsymbol{x}})] + \lambda \mathbb{E}_{\hat{\boldsymbol{x}} \sim \mathbb{P}_{\hat{\boldsymbol{x}}}}[(\|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\| - 1)^2] \tag{4}$$

Using the Wasserstein loss function with gradient penalty enforces the Lipschtiz condition and allows highly complex architectures to be trained successfully, including those with residual units [19].

4.2.2 Architectural Developments

Mirza et al. showed that GANs could be conditioned on additional inputs to the generator network as the noise component [22]. This has allowed facilitated other uses of GANs such as image to image translation for style transfer [23]. Vougioukas et al. used a temporal model to generate a sequence of video frames given an image of a subject

and an audio sequence of spoken text to synthesise the subject speaking [24]. The model uses two discriminators, one to determine if individual frames are realistic images of the subject's face and a second which evaluates the sequence of video frames to determine if it is realistic. The model uses temporal components to examine if the frames of video are consistent in time, preventing sudden jumps in facial position.

Other novel architectures include the progressively growing GAN model [25] which is able to generate highly realistic images of faces to a high resolution. This is achieved by initially training a shallow model to produce 4x4 pixel images before increasing the depth of the network and training further at a higher resolution. By forcing the model to firstly produce and examine low resolution images the model firstly has to be able to synthesise simple low level features effectively, such as facial shape which are common to all samples. Once the model can produce these features a higher resolution is used, allowing it to learn more complex features, such as hair and eyes.

The attention mechanism [26] has been shown to be usefully when applied to image data [27] in order to capture relationships between spatially distant points. As such points are further apart in the image, previously deeper convolutional models were required to allow a large enough receptive field to capture information on these two points. Zhang et al. point out that this is a common issue with DCGANs [17], where generated samples fail to produce structural patterns, while they do exceedingly well at local textural patterns. This is often seen in generated images of animals with realistic fur, but oddly shaped or an incorrect number of limbs. The attention mechanism is applied to GANs [18] to allow for long range dependencies in the image to be modelled by convolutional models more effectively.

To the best of the author's knowledge, there currently exists no generative adversarial networks which aim to generate 3D facial models with speech audio as a conditional input.

5 Current Work and Future Work

At the current stage of the project, the current goal is to construct a new dataset of 3D facial scans of subjects speaking selected phrases by collecting the relevant data, this dataset can be combined with LRW-3D and VOCASET. In order to achieve this, the facial meshes from the individual datasets must be brought into correspondence, creating a larger dataset. The current issue with models which are to be trained on 3D temporal facial scans is the lack of such datasets due to the complexity in creating such datasets. By combining new and existing datasets, this aids in the solution of this issue.

Future work will include training a GANs model which is capable to driving a template 3D mesh given an conditional audio input. This will allow for synthetic data augmentation to be applied to the current 3D facial datasets. This augmented dataset can be used to train further models which accept 3D facial meshes as input for tasks such as lip reading in order to see if model performance can be enhanced without the need for additional data capture.

References

[1] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In Shang-Hong Lai, Vincent Lepetit, Ko Nishino, and Yoichi Sato, editors, *Computer Vision - ACCV 2016 - 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II*, volume 10112 of *Lecture Notes in Computer Science*, pages 87–103, 2016. pages 1, 2, 3, 4

[2] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. Lipnet: Sentence-level lipreading. *CoRR*, abs/1611.01599, 2016. pages 1, 3

[3] Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3444–3453. IEEE Computer Society, 2017. pages 1, 2, 3

[4] Brendan Shillingford, Yannis M. Assael, Matthew W. Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, Marie Mulville, Ben Coppin, Ben Laurie, Andrew W. Senior, and Nando de Freitas. Large-scale visual speech recognition. *CoRR*, abs/1807.05162, 2018. pages 1, 2, 3

[5] Panagiotis Tzirakis, Athanasios Papaioannou, Alexander Lattas, Michail Tarasiou, Björn W. Schuller, and Stefanos Zafeiriou. Synthesising 3d facial motion from "in-the-wild" speech. *CoRR*, abs/1904.07002, 2019. pages 1, 2, 4, 5

[6] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael Black. Capture, learning, and synthesis of 3D speaking styles. *Computer Vision and Pattern Recognition (CVPR)*, 2019. pages 1, 2, 4, 5

[7] 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.*, 120(5):2421–2424, 2006. pages 1, 3

[8] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. pages 1, 2, 4, 5

[9] Philip C. Woodland, C. J. Leggetter, Julian Odell, Valtcho Valtchev, and Steve J. Young. The 1994 HTK large vocabulary speech recognition system. In *1995 International Conference on Acoustics, Speech, and Signal Processing, ICASSP '95, Detroit, Michigan, USA, May 08-12, 1995*, pages 73–76. IEEE Computer Society, 1995. pages 2

[10] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 07 2009. pages 2

[11] Hank Liao, Erik McDermott, and Andrew W. Senior. Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12, 2013*, pages 368–373. IEEE, 2013. pages 2

[12] Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. Deep speech: Scaling up end-to-end speech recognition. *CoRR*, abs/1412.5567, 2014. pages 2, 5

[13] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014. pages 3

[14] J. P. Lewis and K. Anjyo. Direct manipulation blendshapes. *IEEE Computer Graphics and Applications*, 30(4):42–50, July 2010. pages 4

[15] 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 Trans. Graph.*, 36(4):94:1–94:12, 2017. pages 4, 5

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. pages 5, 6

[17] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. pages 6, 7, 8

[18] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *CoRR*, abs/1805.08318, 2018. pages 6, 8

[19] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. *CoRR*, abs/1704.00028, 2017. pages 6, 7

[20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 448–456. JMLR.org, 2015. pages 7

[21] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *CoRR*, abs/1701.07875, 2017. pages 7

[22] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. pages 7

[23] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017. pages 7

[24] K. Vougioukas, S. Petridis, and M. Pantic. End-to-End Speech-Driven Facial Animation with Temporal GANs. *arXiv e-prints*, May 2018. pages 8

[25] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017. pages 8

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6000–6010, 2017. pages 8

[27] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2048–2057. JMLR.org, 2015. pages 8