Cascaded Siamese Self-supervised Audio to Video GAN

Nuha Aldausari, Arcot Sowmya, Nadine Marcus, Gelareh Mohammadi
School of Computer Science and Engineering, University of New South Wales
Sydney, Australia
{n.aldausari, a.sowmya, nadinem, g.mohammadi}@unsw.edu.au

Abstract

Generating meaningful videos that are synchronised to audio signals is a complex synthesis task that requires generation of not only realistic videos but also coherent video motions that conform to the provided audio signals. While tremendous effort has been expended on audio-to-video generative models, these models rely heavily on supervised signals such as face/body key points or 3D meshes. However, key point annotation requires time and effort. Besides, some dataset domains do not have predictable structure, which makes the extraction of points of interest infeasible. Our proposed model consists of a cascaded generator-discriminator architecture that works at the pixel level to generate videos according to the associated soundtracks. It adopts a new self-supervised temporal augmentation technique to optimise the correlation between the audio signal and the generated video instead of relying on supervised signals. The proposed architecture has proven its effectiveness in extensive experiments that compared different models across two datasets.

1. Introduction

While a myriad of image synthesis applications are available, more progress is needed in video generative models to achieve higher user satisfaction. The synthesis of realistic videos is much more complex than its counterpart in the image domain for several reasons. First, a video is a sequence of images, and these images need to be realistic. Moreover, not only should the flow of the frames be coherent, but it is also important to maintain synchronisation between the motion and the audio signal, such as music or speech. Early attempts to generate video content were limited to unconditional video generation such as MoCoGAN [42] and VGAN [44]. Later, additional conditional signals were used to include a broader range of applications in the generation process .

In generating videos based on audio signals, current state-of-the-art models may be divided into two applications: dance and speech synchronisation. Models that generate dance movements according to music are trained on the key points of the joints of a human body [4, 15, 27, 33, 38, 39, 53, 57], or 3D body meshes [28]. In speech-to-video frameworks, most models are built after preprocessing the datasets by cropping or aligning the face [24, 32, 45] or mouth [9] and centring them around the same pixel in all the samples. Other speech-to-video models use 3D meshes [8, 22, 25, 36] or facial landmarks [10, 13, 31, 55, 56] to facilitate the generation process. The model proposed in this paper generates videos at the pixel level, unlike other audio-to-video models that are based on intermediate supervised signals.

There are multiple motivations for avoiding the use of additional intermediate signals. First, manual annotation requires effort and time, and deep learning models that rely on annotations are limited to specific tasks. For example, OpenPose [7] extracts the spatial locations of the joints of human skeletons, which can then be used as key points. Also, some tasks require expert data annotations, such as medical datasets. Due to structure variability, some tasks cannot be annotated with key points or 3D meshes, and the generation models need to deal with such tasks at the pixel level. For example, ocean waves datasets and fireworks datasets have irregular structures.

In this work, we first created an audio-video dataset as the available video datasets usually lack audio signals [12, 37], provide only key points [3, 40], or do not have coherent audio and video signals [37]. For example, in the action recognition dataset UCF101 [37], soundtracks are unavailable in some categories, such as Playing Piano, Playing Tabla and Playing Violin. Moreover, sounds in some other categories do not reflect the changes in motion because of strong background noise. Also, individual category samples alone are insufficient to train a generative model because of the limited number of samples per class. We choose the domain of our dataset to be phonic songs because it has a relatively higher number of samples on YouTube. In addition, we want our dataset to have both audio and video signals, and the audio signals to correlate

with the motion. In our dataset, the audio signal represents a song about letters and objects that start with these letters. The videos contain illustrations of the letters and objects, as well as animations of these components. Our dataset has multiple categories, and to the best of our knowledge there is no automated method to annotate it. We, therefore, challenge our model to use the audio as a self-supervised signal to learn the motion.

In this work, the aim is to generate videos without intermediate signals such as key points or 3D meshes, therefore, we use audio signals as a self-supervised signal to guide the motion in the generated videos. Inspired by Hyun et al. [20] who adopted a shuffling technique in an unconditional video GAN as an auxiliary component to optimise the GAN architecture and generate smooth random motion trajectories, we propose a novel self-supervised GAN architecture conditioned on the audio signal which uses the shuffling technique to generate motion in the video according to the shuffled audio signal.

Our Contributions

1. We build a novel GAN architecture that generates videos at the pixel level according to the audio signals.
2. We propose a novel optimisation and augmentation method that focuses on maintaining the coherence between the audio segments and the motions in the generated clips.
3. We build a new dataset that contains audio and video signals, where the audio signal correlates with the motion in the video.
4. We perform extensive analyses and comparisons between the proposed architecture and state-of-the-art ones in the video realm and with different datasets, which demonstrates that our model surpasses other state-of-the-art models in terms of the quality of the generated images, smoothness of the motion and the synchronisation between the audio and motion signals.

2. Related Work

The introduction of GAN [17] has led to an explosion of content creation. It has proven its effectiveness in the image, audio, text and video domains. While several video generation applications have been successful, video generation is still considered more complex than image generation because of the multimodal nature of videos [6]. Early attempts at video GAN models focussed on generating videos unconditionally. VGAN [44] utilises unsupervised disentanglement for two aspects of video generation, namely background and foreground. The model combines the generation of foreground and background streams to generate the final video. MoCoGAN [42] employs multiple noise vectors that represent changes in the motion and a shared content vector in a recurrent architecture. G3an [49] uses two streams for decomposing the motion and content. In addition, a third stream generates videos, with help from the content and motion streams. While these models [42, 44, 49] synthesise random videos, our proposed model is conditioned on audio and identity frames.

The unconditional video generation models [42, 42, 44] share a strategy, namely decomposing the video representations into multiple sub representations such as content and motion. Such disentanglement facilitates the learning of complex representations to generate videos at the pixel level. However, in conditional video generation models, supervised signals are used to render the final videos. For example, in the case of video-to-video generation, the models learn to generate videos based on segmentation maps [47, 48]. These models can create, for example, a video of portraits given a segmentation map of the face. Image-to-video models usually employ key points [21, 46], 3D body poses [54] or movement directions [14], depending on the application. ImaGINator [50] synthesises videos based on images without a supervised signal. Similar to unconditional models, ImaGINator decomposes videos into a motion vector and a conditional image to reflect the identity in the generated video. The main difference between our work and ImaGINator is that ImaGINator is conditioned only on an image, while our model is conditioned on an image and audio signal to maintain the coherence between audio and video signals. While ImaGINator uses an encoder-decoder architecture for the generator, our model uses multiple streams of the same architecture to perform self-supervision, as explained in section 3.3.

Audio-to-video generative models are the closest to our proposed model. These models may be categorised based on their applications. One such category is dance retargeting, where the training phase uses key points of the skeleton [4, 15, 19, 27, 33, 38, 39, 53, 57] or 3D body poses [28]. Another related application is speech synchronisation, where the input is an audio speech signal along with the target identity image, and the output is a video of the person uttering the conditioned speech. These models use a 3D mesh of the face [8, 22, 25, 36] or some key points [10, 13, 31, 55, 56]. Deviating from these models, our proposed model can generate coherent motion from the audio signal in a self-supervised manner in scenarios where key points or 3D meshes are not readily available. There are a few audio-to-video models that synthesise videos at the pixel level [41]. Tsuchiya et al. [41] trained their model on a dataset from a single category. In contrast, we develop a model that can generate videos from multiple categories. Unlike Tsuchiya et al. [41], we perform extensive experiments, presented in section 4, to compare our models with others and across multiple datasets.

Self-supervised learning is a machine learning technique

that uses the provided data to learn a representation, without the need for manual annotation or automatic labelling of the dataset. Self-supervised learning has been applied in the video domain to achieve a specific goal, such as predicting or verifying the natural order of the video frames. For example, Misra et al. [30] built a triple Siamese network to classify whether video frames are in the correct order without predicting the actual order. Lee et al. [26] solved a more complex problem using a Siamese architecture to predict the order of shuffled frames, where the number of all possible orders is a factorial of the number of input frames. Xu et al. [52] dealt with videos as small clips and showed that using clips instead of frames can help to better learn the motion. The videos are divided into clips first; then the clips are shuffled. After that, these clips can be used to train a network to predict the order of the shuffled clips. The trained network could be used in downstream tasks that require temporal features. Hyun et al. [20] adopted clip shuffling in unconditional video generative models. The shuffled and non-shuffled clips were fed into a network with shared weights as the discriminator to predict the order of the input clips. The shuffling technique optimised their model to generate random smooth changes in the motion of the synthesised videos. However, while the introduction of the shuffling technique [20] into an audio-to-video GAN results in smooth motion, it is still unsynchronised with audio, as explained in section 4.5. Therefore, in order to generate motion trajectories in accordance with the audio, we introduce audio as an input in the shuffling technique. Our proposed model initiates shuffling from the generator side, and the encoded shuffled audio segments are fed into a cascaded Siamese GAN model to generate shuffled videos. In addition to image and video discriminators, there is also an order predictor network with shared weights as a video discriminator to estimate the order of the clips. Providing the shuffled audio segments and comparing the generated clips with the ground truth shuffled clips helps as an augmentation technique to learn the correlation between different orders of the audio and the motion signals more effectively.

3. Methodology

In this paper, we propose a novel cascaded Siamese self-supervised audio-to-video GAN. The overall architecture is divided into multiple GAN models based on the number of clips in each video. Each GAN model has a generator and discriminator, and the generator has an encoder and a decoder with skip connections. In this paper, a video refers to the entire video sample that we aim to generate, whereas a clip means a temporal segment of a video. A visualisation of the architecture is provided in Fig. 1, with only two streams because of space limitations. Each stream generates one clip given a corresponding audio segment and an identity image. The architecture may be extended to any number of streams based on the number of clips. The following subsections describe the functionality of different components in this architecture.

3.1. Generator

When loading the data during training, a video sample and the corresponding soundtrack are divided into equal segments called clips. These clips are then shuffled, and the order of the shuffled clips is recorded, to be used later in the order prediction model as explained in section 3.3. The flow of the generation process starts with the encoding of the audio signal, which is the corresponding audio chunk for the video segment that is used in the video discriminator side. The audio wave signal is converted to log Mel-spectrogram since this format better represents the audio signal [16, 23, 38, 51]. The resulting log Mel-spectrogram is divided into overlapping equal segments of the same number as the number of clips in a video. These segments are encoded using GRU units [11]. The initial frame is encoded using a CNN architecture. These two encoded signals, along with the encoded class, are concatenated and then input into the decoder to generate the first clip of the video. There are skip connections between the image encoder and the decoder, in order to focus on generating a video with the same identity as the encoded image. The same procedure is repeated in the second stream to generate another segment of the video, and all streams share the trainable weights. However, in the second stream, the last frame of the generated clip from the previous stream is used to generate the following clip of the generated video. It is important to note that every generated clip is forced to have coherent motion, while the clip order is shuffled during training. The shuffling effect can be removed at test time to produce coherent overall motion, as discussed in section 3.3.

3.2. Discriminators

There are two levels of discriminators: video level and image level. The image level discriminator compares the real and fake images in terms of the spatial aspect. The video level discriminator consists of multiple clip-discriminators. The architecture is shared among the clip discriminators and consists of a 3D convolutional neural network. The final decision for the video level is made using a majority voting ensemble of clip discriminators. The video discriminator not only evaluates the spatial aspect of the video, but also considers the motion in the clips.

3.3. Shuffling

Inspired by the shuffling technique in Self-supervised Video GAN (SVGAN) [20], we propose a novel shuffling process for audio-to-video GAN. Unlike SVGAN however, where the shuffling is performed in the discriminator to op-

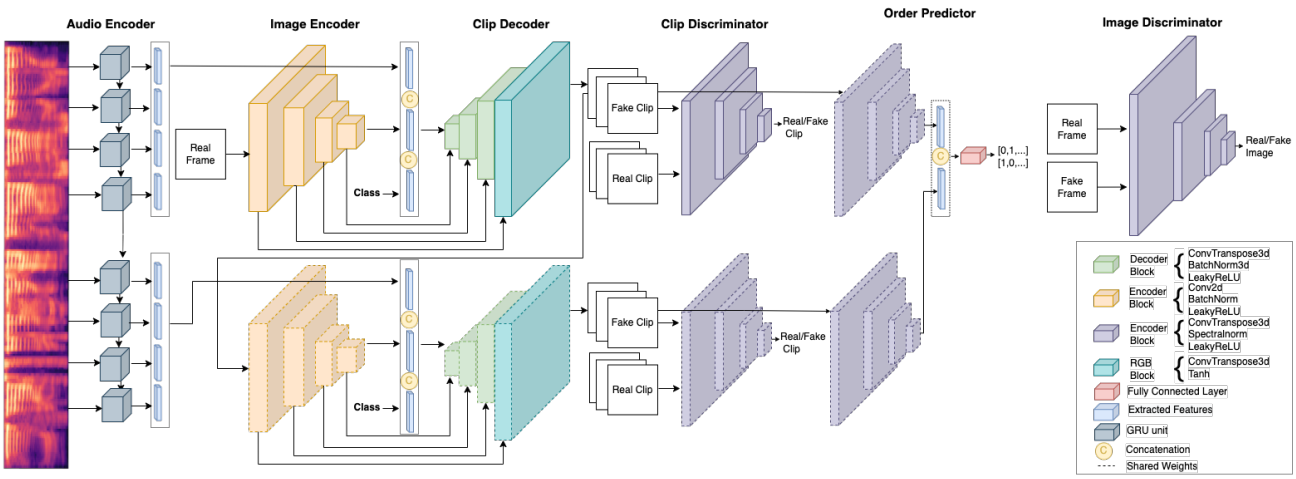

Figure 1. Cascaded Audio-to-Video GAN architecture includes multiple streams. Each stream has encoders, decoder, clip discriminator and order predictor model. There are two encoders, one each for the audio signal and the initial image.

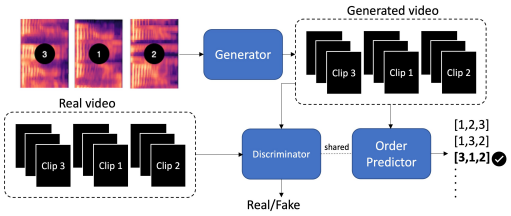

Figure 2. Order predictor model in the training stage. The audio segments may be in any order.

timise the GAN architecture, in our model we perform shuffling in both the generator and discriminator. We introduce shuffling in the audio segments, which are used as input to generate shuffled clips. We train our model on audio-video datasets that have a correlation between audio and video motion. During training, our model learns the coherent motion within the clips. Having multiple generators and discriminators maintains coherent motion within the clips, while the cascade architecture helps maintain coherent motion in the entire video. This approach may be considered as a method of temporal augmentation of the dataset by providing clips and soundtracks in all possible shuffled permutations during the training phase. However, only chronologically ordered audio segments are used to generate the ordered video during test time. Having control over the order of the audio signal stops leakage of any unnatural motion trajectory into the generated videos. The order predictor model shown in Fig. 2 shares weights with the clip discriminators. Similar to the video discriminator, the main structure of the order predictor model consists of multiple encoders to downscale the generated clips. The outputs of these encoders are concatenated and input to fully connected layers to predict the order.

3.4. Loss function

Three main loss functions are used: adversarial loss, self-supervised loss and reconstruction loss. Adversarial loss is applied using image discriminator DI for real image i and fake image x, as in equation (1). In addition, video discriminator DV evaluates the spatio-temporal aspects of real video v and the fake video. The fake samples are generated using $G(i, c, a)$, where i is the initial image, c is the category, and a is the audio segment, as in equation (2).

$$\mathcal{L}_I(DI, G) = E_{i\sim p_{data}}[logDI(i)] + E_{i,c,a\sim p_{data}}[log(1 - DI(x)] \quad (1)$$

$$\mathcal{L}_V(DV, G) = E_{v\sim p_{data}}[logDV(v)] + E_{i,c,a\sim p_{data}}[log(1 - DV(G(i, c, a))] \quad (2)$$

Self supervised loss is used to evaluate the correctness of the resulting order prediction. To compare the prediction p_i with the ground truth permutation y_i, we use cross-entropy as in equation (3):

$$\mathcal{L}_T = -\sum_{n=1}^{j} y_i log(p_i) \quad (3)$$

For reconstruction loss, we use $\mathcal{L}_1$, as in (4), to sharpen the generated images and make them closer to real ones.

$$L_{reconstuction} = E[\|v - G(i, c, a)\|_1] \quad (4)$$

Figure 3. five clips from the proposed dataset samples at 5Hz. Each clip belongs to a category. For example, the first clip is within "S" category.

3.5. Dataset

We constructed a dataset from alphabetic videos from YouTube[1] illustrating letters and objects along with phonic songs. The choice was motivated by the accessibility and variability of such videos. In these videos, the object and the letter are moving according to the music. Candidate videos were chosen manually, downloaded and segmented into short videos, where each short video illustrates only one letter. The segmentation into short videos was performed based on the transcript of the videos, if available. If the transcript was not available on YouTube, speech-to-text Google API [2] was used to generate the transcript. The final videos have varying lengths, between 14 frames and 100 frames. Metadata for each video, such as the class (e.g., letter) and the number of frames, was collected. The dataset has 26 categories to represent all the alphabetic letters, and Fig. 3 illustrates 5 examples by showing the first 10 frames sampled at 5Hz. The current version of the dataset used in this work has 1570 clips, with 30 to 90 short videos per letter. This dataset has been made publicly accessible through this link[2].

For evaluation purposes, we also trained our model on the VidTIMIT Audio-Video dataset [35]. This dataset contains 43 people uttering 10 short sentences; the videos were captured for the head region. The total number of videos is 430, with varying lengths between 56 to 240 frames. The videos were extracted and saved as images, and the audio was stored in a mono WAV file. The dataset was recorded in a lab setting.

4. Experiments

Due to the different design of our proposed audio-to-video GAN model, we could not directly compare our model with other baseline audio-to-video models [4, 8, 10, 13, 15, 19, 22, 25, 27, 28, 31, 33, 36, 38, 39, 53, 55–57], as these models use supervised signals such as body/face landmarks or 3D meshes to learn changes in motion, as mentioned in section 2. In contrast, our model utilises a temporal self-supervised technique to learn the relationship between audio and frame motions. Therefore, we compared our model to two unconditional video-GAN models, namely MoCoGAN [42] and G3an [49] after adjusting these models to use the same input signals for fair comparison. We applied the same audio and image encoders that our model uses to these models. For MoCoGAN [42], we encoded the log Mel-spectrogram using GRU units to represent the motion vectors in the original implementation of MoCoGAN. Also, instead of the content vector, we used the encoded initial image to act as the identity of the generated video. In G3an [49], the motion and appearance random vectors were replaced by the encoded audio and image, respectively. We also compared our model with ImaGINator [50], an image-to-video model that shares a similar architecture as one of our streams. We adjusted the model to have an audio encoder to ensure that the same conditional signals are being input to the model. In addition, we compared our model with the PhonicsGAN [5] audio-to-video model that generates videos at the pixel level without any initial image. To make PhonicsGAN comparable in terms of input conditions, we also added the initial image as an additional condition signal.

4.1. Implementation Details

The number of frames in the generated videos is 32. The spatial dimension of the frames is 64x64, due to limited computational resources. Similarly, the number of audio segments is 32, and every audio segment corresponds to one frame. The dimension of the audio segments is 64x20. The class of the video (e.g., alphabet letter) was encoded using One-Hot Encoding. In our dataset, we used 26 bits to encode the letter illustrated in the clip, with each bit corresponding to one alphabet letter. In the VidTIMIT dataset, we used 43 bits to encode the category, which is the face of the person uttering the sentences. During training, the batch size was 32. Adam optimizer was used, and the learning rate was 0.0002. To enable temporal shuffling for our model, we divided the aligned audio signal of the 32-frames into four segments. In addition, we divided the 32-frame video into four 8-frame clips. Thus the implementation contains four Siamese streams. The number of permutations is 4!. Therefore for each sample video, 4! shuffled orders were provided during training.

4.2. Qualitative Evaluation

We compared the results of our model with other state-of-the-art models such as MoCoGAN [42], G3an [49], ImaGINator [50] and PhonicsGAN [5] qualitatively. Our

[1] https://www.youtube.com/

[2] https://github.com/NuhaAldausari/Cascaded-Siamese-Self-supervised-Audio-to-Video-GAN

model surpasses other models in terms of the quality of the generated images, smoothness of the motion and better correlation with the audio signal, as shown in Fig. 4. More videos from our model and more comparison videos are provided in this link[3]. Visual inspection of the results clearly shows that MoCoGAN, PhonicsGAN and G3an tend to have unclear objects and artefacts because the dimensions of the encoded content vectors of these models are smaller than our proposed model, which results in lower quality images. Our model and ImaGINator have better quality in the generated images because both models have skip connections between the encoder and decoder. Feeding the initial image at multiple scales helps to reflect better content in the generated image. In addition, our generated videos have smooth and correlated motion compared to other models for several reasons. First, other models have a limited receptive field for the input audio signal as their original implementations have limited motion vectors. In addition, our shuffling technique facilitates the augmentation of the temporal aspects of the video, resulting in better learning of the correlation between the audio and changes in the video.

We also analysed the latent representations that correspond to the content and the motion. For the temporal dimension, we kept the audio signal constant and changed the initial image. We observe that our model can successfully generate similar motion trajectories for several initial images given the same audio, as shown in Fig. 5. The object starts to appear around the same highlighted frames in all the generated videos. In addition, we conducted an experiment by fixing the initial image and changing the audio signals. As shown in Fig. 6, our model can generate different motion trajectories based on the audio signals. A user study complements these results is described in section 4.4.

4.3. Quantitative Evaluation

In addition to qualitative evaluation, we deploy three quantitative evaluation metrics: Inception Score (IS) [34], Frechet Inception Distance (FID) [18] and Frechet Video Distance (FVD) [43]. All these metrics are calculated after extracting the features using pre-trained networks. IS provides the confidence that the generated images belong to a given class. It also checks the diversity of the generated images by evaluating how wide a range of classes are included in the generated samples. While IS [34] evaluates the synthesised samples exclusively, FID and FVD compare the generated samples with real ones. Both FID and FVD use statistics to compare the distributions of real and fake videos. The main difference is that FID uses a pre-trained image network while FVD is calculated based on extracted features from a temporal pre-trained network. The mean and covariance for real data and fake distribution are used to calculate the final score. Overall, our model shows much better IS, FID and FVD scores on both datasets as reported in Tab. 1. While our model has a comparable spatial score (IS, and FID) to ImaGINator (because both models use skip connections and sufficient dimensions for the encoded content), it outperforms other models in terms of FVD significantly. The FVD score captures the video correlation with the audio signal and the smoothness of the motion.

[3]https://github.com/NuhaAldausari/Cascaded-Siamese-Self-supervised-Audio-to-Video-GAN

Model	Phonics Dataset			VidTIMIT		
	IS↑	FID↓	FVD↓	IS↑	FID↓	FVD↓
MoCoGAN	0.0001	59.81	1248.73	0.0001	160.82	2041.02
PhonicsGAN	0.0002	46.82	961.85	0.0002	113.97	613.46
G3an	0.0002	36.93	936.99	0.0003	47.02	826.29
ImaGINator	0.0005	35.23	377.25	0.0006	45.07	171.29
Our model	**0.0008**	**27.56**	**95.13**	**0.0009**	**26.05**	**81.49**

Table 1. Quantitative Evaluation using IS, FID, and FVD scores. Our model shows better performance in terms of quality of the generated images, and synchronisation of the generated motion.

4.4. User Study

The available GAN metrics are limited and do not provide information about specific aspects of the generated samples, such as the naturalness of the motion and realism of the generated examples in comparison with real ones. We, therefore, conducted a user study to evaluate our model in terms of perceived quality and motion coordination with the input music. We conducted a subjective analysis on Amazon Mechanical Turk [1], where 40 participants completed two surveys in one session. In the first survey, we asked the participants to rate 26 videos, one per letter, generated by the proposed model in terms of motion synchrony between the audio and video. We asked, "how well is the object movement coordinated with the song's rhythm?" Participants could answer using a 9-point Likert scale, where 1 corresponds to "Not well at all" and 9 corresponds to "Extremely well". We observed that 83% of human raters chose a rating above 5, with an average value of 7.13. The second part of the survey included comparisons between the generated videos from our model and other models such as MoCoGAN [42], G3an [49], ImaGINator [50] and PhonicsGAN [5]. The results, reported in Tab. 2, show that raters preferred our generated examples over other models using a comparative 7-point Likert scale in terms of quality of the frames, association with the song and naturalness of motion.

4.5. Ablation Study

We evaluated the effectiveness of the different components of our model by incrementally adding components. First, we trained a basic version of our model consisting of one stream that accepts the entire audio and generates a

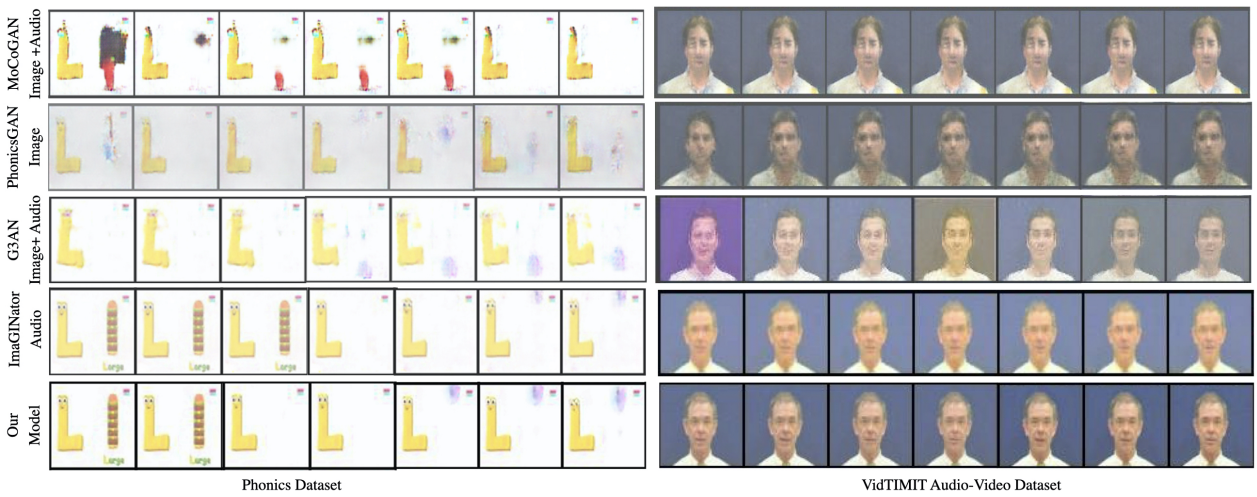

Figure 4. Qualitative comparison of generated videos using multiple models, namely MoCoGAN, PhonicsGAN, G3an, ImaGINator and our model. Frames are sampled with a time step = 3.

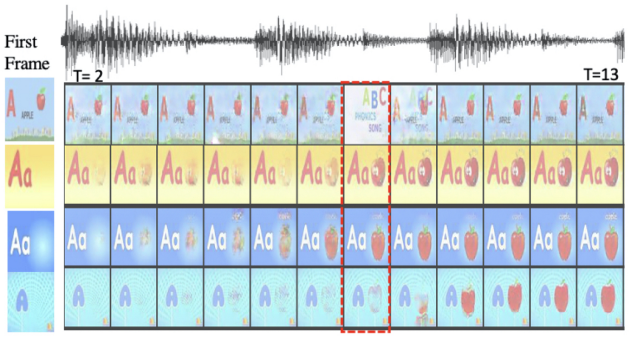

Figure 5. Generated videos conditioned on the same audio and different initial images. The same motion is observed in all of the generated videos. The transaction starts to occur around the red dotted frames (T=8) in all videos.

Method	Quality(%)	Assoc. w/ music(%)	Motion(%)
Ours/MoCoGAN	67 / 20	52 / 31	60 / 35
Ours/PhonicsGAN	63 / 4	64 / 14	70 / 14
Ours/G3an	54 / 6	43 / 25	50 / 31
Ours/ImaGINator	54 / 22	56 / 27	56 / 22

Table 2. User preferences in terms of quality, association with the song and natural motion when compared with our generated examples given the same input. In each cell, the number on the left represents the raters' preference (greater than 4 on the Likert scale) for the proposed framework over the model in the right. The number on the right represents the opposite. The remaining percentage (not shown) is when raters chose 4 on the Likert scale.

32-frame video. The model tends to generate artefacts, as shown in the first example of Fig. 7, and one reason could be an inability to reflect the input signals in the generated sample using limited computational parameters. Then, we added the shuffling technique as introduced in [20] to examine its effectiveness with the methodology of our baseline model. We observed that the shuffling technique helps in learning stable and smooth motion trajectories, resulting in an improved FVD score (Tab. 3). However, the generated motion does not correlate well with the audio signal. The song transcript was "T is for Telephone T T T Telephone." We expected to have an entrance motion of an object at the end of the frames. However, the same frame is generated from frame 5 onward, as in the second row of Fig. 7, even though there is a noticeable change in the audio. This is because the shuffling technique does not consider the shuffling of the audio segments. Adding multiple generators that are conditioned on the shuffled audio segments helps in generating correlated motion as in the third video in Fig. 7. However, we started to see artefacts around the start/end of clips. Then, we added multiple discriminators along with multiple generators. The FVD score Tab. 3 is enhanced, however we still noticed artifacts, as shown in Fig. 7. Then, we added the cascade architecture along with the Siamese GAN, which resulted in a better FVD score. The cascade architecture helps in diminishing the artefacts between the clips and generates better motion smoothness and audio-video correlation.

4.6. Limitations and Ethics

The main limitation of the generated examples of our proposed model when trained on the phonics dataset is that the generated object might not necessarily reflect the corresponding object in the song if the object does not appear in the input frame. We use the letter class as a conditional signal to always generate the correct letter. However, the

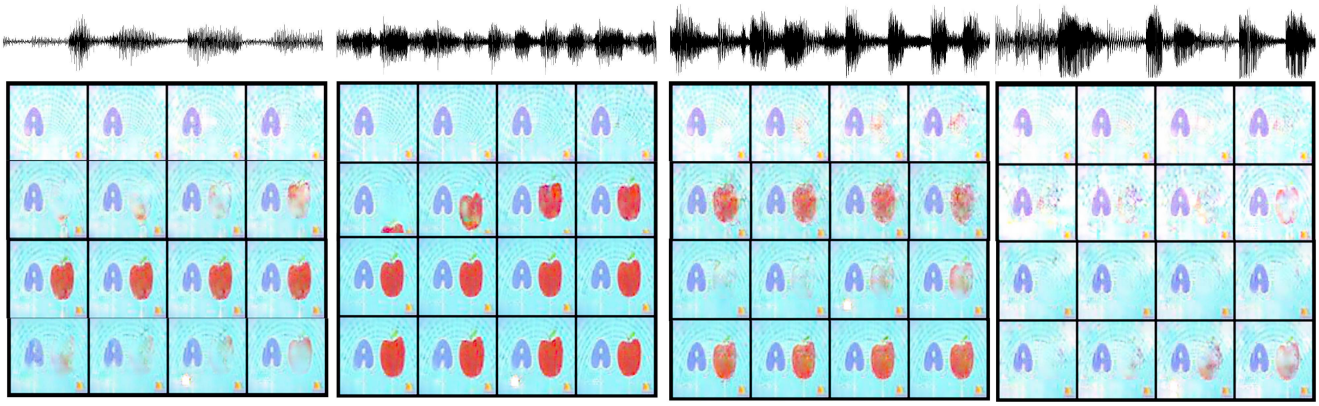

Figure 6. Generated videos conditioned on the same initial image and different audio signals. From left to right, motions are generated based on the input songs and may be interpreted as: disappear-appear-disappear, disappear-appear, disappear-appear-disappear-appear, and disappear. Frames are sampled with a time step = 2.

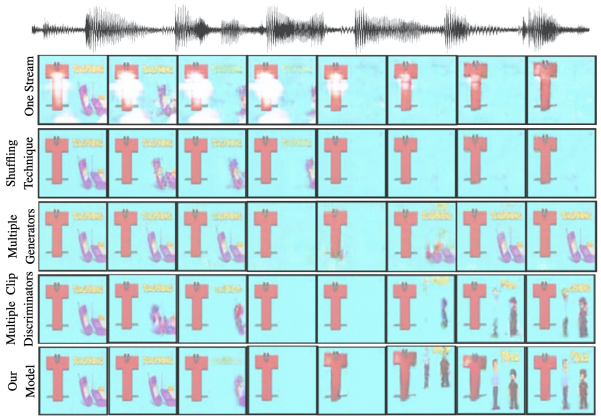

Figure 7. Generated videos using multiple versions of our model. We incrementally added a component and tested our generated examples. The corresponding song's transcript for the video was "T is for Telephone T T T Telephone." Frames are sampled with a time step = 3.

Architecture	FVD
Baseline (one stream)	305.17
+ With shuffling technique	95.80
+ With multiple generators	480.53
+ With multiple discriminators	366.01
+ **With cascade architecture**	93.13

Table 3. Ablation study on the proposed model. We added components of our model incrementally to validate the importance of each structure.

utterance of the object name in the song, even if present, is insufficient to build such a correlation because we have around 20 objects on average per letter, and the average number of samples is only 3 per object. Therefore it is a complex task to find the relationship between the object in the audio signal and the object in the generated frames. One solution could be to manually collect images of the object class and use it as an additional conditional signal. Another solution could be to collect more samples, but this might be a more time-consuming solution and that is only marginally more effective. Another limitation is related to the generated motion. Because some samples in the phonics dataset have static-image videos, the generated motion trajectories sometimes may be static even though there are conditional audio signals. Similar to speech synchronisation models, the proposed model could be used to generate fake videos of persons uttering any speech. The regulations for such applications have been discussed elsewhere [29].

5. Conclusion

The main goal of this paper was to maximise the correlation between synthesised video and a corresponding audio signal without using any intermediate supervised signal. We constructed a dataset that has both audio and video signals to cover the limitations in other audio-video datasets, such as lacking audio signals, or video frames (e.g., key points instead of actual frames). Our dataset can be used to train models that work at the pixel level. Audio is a separate signal that might help in drawing the content and moving objects. Therefore, we have proposed a self-supervised approach based on a novel shuffling mechanism to better utilise this signal. The proposed model outperforms other base models in terms of the quality and motion flow as measured quantitatively, qualitatively and perceptually.

6. Acknowledgements

The first author is supported by a scholarship from Princess Nourah bint Abdulrahman University, KSA.

References

[1] Amazon mechanical turk. 6

[2] Google speech to text. 5

[3] Sfu motion capture database. 1

[4] Hyemin Ahn, Jaehun Kim, et al. Generative autoregressive networks for 3d dancing move synthesis from music. *IEEE Robotics and Automation Letters*, 5(2):3500–3507, 2020. 1, 2, 5

[5] Nuha Aldausari, Arcot Sowmya, Nadine Marcus, and Gelareh Mohammadi. Phonicsgan: Synthesizing graphical videos from phonics songs. In *International Conference on Artificial Neural Networks*, pages 599–610. Springer, 2021. 5, 6

[6] Nuha Aldausari, Arcot Sowmya, Nadine Marcus, and Gelareh Mohammadi. Video generative adversarial networks: a review. *ACM Computing Surveys (CSUR)*, 55(2):1–25, 2022. 2

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 1

[8] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *European Conference on Computer Vision*, pages 35–51. Springer, 2020. 1, 2, 5

[9] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 520–535, 2018. 1

[10] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7832–7841, 2019. 1, 2, 5

[11] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 3

[12] Patsorn Sangkloy Bhavishya Mittal Sean Dai James Hays Daniel Castro, Steven Hickson and Irfan Essa. Let's dance: Learning from online dance videos. In *eprint arXiv:2139179*, 2018. 1

[13] Dipanjan Das, Sandika Biswas, Sanjana Sinha, and Brojeshwar Bhowmick. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *European Conference on Computer Vision*, pages 408–424. Springer, 2020. 1, 2, 5

[14] Michael Dorkenwald, Timo Milbich, Andreas Blattmann, Robin Rombach, Konstantinos G Derpanis, and Bjorn Ommer. Stochastic image-to-video synthesis using cinns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3742–3753, 2021. 2

[15] Yinglin Duan, Tianyang Shi, et al. Semi-supervised learning for in-game expert-level music-to-dance translation. *arXiv preprint arXiv:2009.12763*, 2020. 1, 2, 5

[16] Yinglin Duan, Tianyang Shi, et al. Semi-supervised learning for in-game expert-level music-to-dance translation. *arXiv preprint arXiv:2009.12763*, 2020. 3

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[18] Martin Heusel, Hubert Ramsauer, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637. 6

[19] Ruozi Huang, Huang Hu, et al. Dance revolution: Long sequence dance generation with music via curriculum learning. *arXiv preprint arXiv:2006.06119*, 2020. 2, 5

[20] Sangeek Hyun, Jihwan Kim, and Jae-Pil Heo. Self-supervised video gans: Learning for appearance consistency and motion coherency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10826–10835, 2021. 2, 3, 7

[21] Yunseok Jang, Gunhee Kim, and Yale Song. Video prediction with appearance and motion conditions. In *International Conference on Machine Learning*, pages 2225–2234. PMLR, 2018. 2

[22] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14080–14089, 2021. 1, 2, 5

[23] Takuhiro Kaneko, Shinji Takaki, et al. Generative adversarial network-based postfilter for stft spectrograms. In *Interspeech*, pages 3389–3393. 3

[24] Neeraj Kumar, Srishti Goel, Ankur Narang, and Mujtaba Hasan. Robust one shot audio to video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 770–771, 2020. 1

[25] Avisek Lahiri, Vivek Kwatra, Christian Frueh, John Lewis, and Chris Bregler. Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2755–2764, 2021. 1, 2, 5

[26] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 667–676, 2017. 3

[27] Hsin-Ying Lee, Xiaodong Yang, et al. Dancing to music. *arXiv preprint arXiv:1911.02001*, 2019. 1, 2, 5

[28] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 1, 2, 5

[29] Edvinas Meskys, Julija Kalpokiene, Paulius Jurcys, and Aidas Liaudanskas. Regulating deep fakes: legal and ethical considerations. *Available at SSRN 3497144*, 2019. 8

[30] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order

verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016. 3

[31] Gaurav Mittal and Baoyuan Wang. Animating face using disentangled audio representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3290–3298, 2020. 1, 2, 5

[32] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020. 1

[33] Xuanchi Ren, Haoran Li, et al. Self-supervised dance video synthesis conditioned on music. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 46–54. 1, 2, 5

[34] Tim Salimans, Ian Goodfellow, and other. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*, 2016. 6

[35] Conrad Sanderson and Brian C Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In *International conference on biometrics*, pages 199–208. Springer, 2009. 5

[36] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want. *arXiv preprint arXiv:2001.05201*, 2020. 1, 2, 5

[37] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1

[38] Guofei Sun, Yongkang Wong, et al. Deepdance: music-to-dance motion choreography with adversarial learning. *IEEE Transactions on Multimedia*, 23:497–509, 2020. 1, 2, 3, 5

[39] Taoran Tang, Jia Jia, et al. Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1598–1606. 1, 2, 5

[40] Taoran Tang, Jia Jia, and Hanyang Mao. Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1598–1606, 2018. 1

[41] Yukitaka Tsuchiya, Takahiro Itazuri, et al. Generating video from single image and sound. In *CVPR Workshops*, pages 17–20. 2

[42] Sergey Tulyakov, Ming-Yu Liu, et al. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018. 1, 2, 5, 6

[43] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 6

[44] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29:613–621, 2016. 1, 2

[45] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven facial animation with temporal gans. *arXiv preprint arXiv:1805.09313*, 2018. 1

[46] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *Proceedings of the IEEE international conference on computer vision*, pages 3332–3341, 2017. 2

[47] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. *arXiv preprint arXiv:1910.12713*, 2019. 2

[48] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018. 2

[49] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. G3an: disentangling appearance and motion for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5264–5273, 2020. 2, 5, 6

[50] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. Imaginator: Conditional spatio-temporal gan for video generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1160–1169, 2020. 2, 5, 6

[51] Yuxuan Wang, RJ Skerry-Ryan, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017. 3

[52] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019. 3

[53] Nelson Yalta, Shinji Watanabe, et al. Weakly-supervised deep recurrent neural networks for basic dance step generation. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE. 1, 2, 5

[54] Jae Shin Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, and Christian Theobalt. Pose-guided human animation from a single image in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15039–15048, 2021. 2

[55] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9299–9306, 2019. 1, 2, 5

[56] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020. 1, 2, 5

[57] Wenlin Zhuang, Congyi Wang, et al. Music2dance: Music-driven dance generation using wavenet. *arXiv preprint arXiv:2002.03761*, 2020. 1, 2, 5