COLLABORATIVE LEARNING TO GENERATE AUDIO-VIDEO JOINTLY

Vinod K Kurmi[*]   *Vipul Bajaj*[*]   *Badri N Patro*[*]   *K S Venkatesh*[*]   *Vinay P Namboodiri*[†]   *Preethi Jyothi*[‡]

[*] Indian Institute of Technology Kanpur   [†]University of Bath   [‡]Indian Institute of Technology Bombay

ABSTRACT

There have been a number of techniques that have demonstrated the generation of multimedia data for a single modality at a time using GANs such as the ability to generate images, videos, and audio. However, so far, the task of multimodal generation of data, specifically for audio and videos both, has not been explored well. Towards this problem, we propose a method that demonstrates that we are able to generate naturalistic samples of video and audio data by the joint correlated generation of audio and video modalities. The proposed method uses multiple discriminators to ensure that the audio, video, and the joint output are also indistinguishable from real-world samples. We present a dataset for this task and show that we are able to generate realistic samples. This method is validated using various standard metrics such as Inception Score, Frechet Inception Distance (FID) and through human evaluation.

Index Terms— Audio-video generation, cross-modal learning

1. INTRODUCTION

In the real world, different modalities of the same data are highly correlated. Video-audio pairs are one example of them. Recent studies have explored the generation of data such as audio, videos, and texts exhaustively. Modeling and being able to generate these realistically is always a challenge. Several works solve for generative modeling of images, audio signals, and videos (without the corresponding audio). These methods consider the single modality generation problem either in the form of unconditional generation or conditional generation. The meta information provided for conditional generation can be the class label or one of the other modalities. Cross model generation focuses on the latter form of conditional generation. The former is known as single modality generation, such as audio generation or video generation, which generate the required modality independently. One of the major drawbacks of these methods is that they do not consider that audio and video streams of the same data are highly correlated and could be generated jointly both benefitting each other. Most multimedia synchronization methods developed in the past have dealt with uni-modal streams and consider only audio data or video data. Recently, there have been a few works [1, 2] that aim to solve for a cross-modal generation. However, they generate a single image frame for an audio duration and hence fail in real-world video generation as they do not generate a sequence of image frames (i.e.
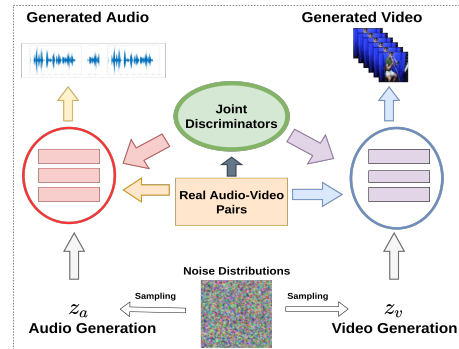

Fig. 1. Illustration of collaborative generation of audio-video pairs using joint discriminators

a real video signal) corresponding to the synchronized audio. The interaction between audio and vision, despite seeing traction as of late, is still largely unexplored. This is a particularly relevant topic to the vision community because humans routinely perform tasks which involve multiple modalities. In order to solve the problem effectively, we propose a collaborative learning framework that uses adversarial models to ensure that there is out of sample generalization. Specifically, we incorporate multiple discriminators to ensure that the samples generated individually for each modality are realistic. Figure 1 shows our proposed framework.

The exact problem that we want to tackle has not been explored precisely and therefore we did not have a dataset directly applicable on our task. Hence, we composed two novel datasets - Hand Motion Music Dataset (HMMD) and Sub-URMP-video dataset (the latter by modifying an existing dataset). These datasets are primarily related to musical instruments, and through these datasets, we explore the accurate joint-modal correspondence between audio and video signal. Concisely, we make the following main contributions through this paper: 1) We propose a joint collaborative audio-video generation module that uses noise as a latent vector. 2) We provide an audio-video multimodal dataset comprising specifically chosen videos where audio is a direct outcome of the motion (actions) happening in the video frames. 3) We formulate the joint audio-video generation training problem in an adversarial framework and achieve encouraging results.

2. RELATED WORK

Generative adversarial networks [3] is one of the unsupervised techniques that maps low dimensional latent features to

high dimensional data. Various architectures have been suggested for generating images using GANs [4, 5, 6, 7]. These image generation architectures have also been extended for video generation. MoCoGAN [8] proposes to solve the video generation problem by decomposing motion and content vector. The work in generation is not only limited on to the visual modality, but it also extended for other modalities such as music [9, 10] and speech generation [11, 12, 13]. For instance, SEGAN [14] uses and adversarial framework to generate speech signals. Van et al. [15] introduced a WaveNet model based on an autoregressive method to predict raw audio samples. Engel et al. [16] used an auto-encoder model in WaveNet to generate musical instrument sounds. Other multi-modal generation frameworks that have been proposed in the space of audio-visual modalities are [1, 2, 17, 18]. Most of the above audio generation works are based on conditional generation conditioned on an image frame of the corresponding video. Recently, several approaches that solve for the alignment of various modalities [19, 20, 21, 11, 22, 12, 23] have also been suggested. A few very recent works have also explored the multimodal generation problem. Gan et al.[24] synthesized plausible music for a silent video clip of people playing musical instruments. Another similar work [25] generated music for a given video. [26] generated visually aligned sounds from videos. In sound2sight [27], future video frames and motion dynamics are generated by conditioning on audio and a few past frames. In SA-CMGAN [28] self-attention mechanism is applied to cross-modal visual-audio generation.

3. METHOD

The proposed framework consists of three modules - audio generation module, video generation module and joint discriminators module.

Audio Generation Module: It consists of an audio generator A_G, that generates audio from noise ϵ_0 and a discriminator [29]. This generator module is parameterized through a function $G_a(\epsilon_0, W_{ga})$ where W_{ga} are the weights of the model. We work with the raw audio waveform, and the generator produces raw waveforms instead of spectrograms. The generator module uses 1-D deconvolutional layers. Since there are frequency bands present in each audio waveform, unlike the image signal where the frequency patterns are not so common, an image discriminator architecture doesn't work well for audio generation. The discriminator easily learns the frequency pattern commonly known as the checker board effect and discards those waveforms. These artifact frequencies occur for a particular phase. Therefore, our discriminator needs to learn this trivial policy to discard such generated examples. To resolve this hurdle, we employ the skill of phase shuffle operation [29] to prevent our audio discriminator from learning such artifacts. It is parameterized through a function $D_a(\epsilon_0, W_{da})$ where W_{da} are the weights of the model. Generator G_a is trained to minimize the following value function,

while D_a is trained to maximize it.
$$V(D_a, G_a) = \mathbb{E}_{\boldsymbol{a} \sim R_A}[\log D_a(\boldsymbol{a})] + \\ \mathbb{E}_{\boldsymbol{\epsilon} \sim P_\epsilon}[\log(1 - D_a(G_a(\boldsymbol{\epsilon_0})))] \tag{1}$$

$\boldsymbol{a}$ is sampled from real data(R_A) and $\boldsymbol{\epsilon}$ is random noise.
Video Generation Module: Our video generation component takes inspiration from the idea of decomposing videos into a separate motion and content part [8]. Our video clips $v_n = [I_1, \ldots, I_n]$ are composed of N frames. The temporal relationship between the video frames is obtained by modeling the latent space vector Z_n. The content of the video clips is encapsulated by introducing a context vector Z_C and is combined with every motion vector. Each input to the video generator is a combination of the latent motion vector and the latent content vector. Each Z_I is an input to the image generator to generate each frame in the video clip. We use individual image discriminator D_I and a video discriminator D_V to play the role of a judge to provide criticism to video (image) generator G_I. The learning problem is formulated as:

$$\min_{G_I} \max_{D_I, D_V} (C_v(D_I, D_V, G_I)) =$$

$$\mathbb{E}_{v \sim \mathcal{V}}[\log D_I(S_1(v))] + \mathbb{E}_{\hat{v} \sim G_V}[\log(1 - D_I(S_1(\hat{v})))] +$$

$$\mathbb{E}_{v \sim \mathcal{V}}[\log D_V(S_T(v))] + \mathbb{E}_{\hat{v} \sim G_V}[\log(1 - D_V(S_T(\hat{v})))]$$

S_1 and S_T are two sample functions for image discriminator D_I and video discriminator D_V which takes as input, one frame and T consecutive frames respectively. The sample functions S_1, and S_T are sampled from real $\mathcal{V}$ and generated video $\hat{V}$ clip. $\hat{v}$ is the generated video.
Joint Discriminators: The marginal distributions are represented for audio and video as $p(a)$ and $p(v)$. The joint distribution of audio and video is denoted by $p(a, v)$. To jointly learn the audio-video generation, we use two joint discriminators, inspired by [30]. This pair of joint discriminators is trained to distinguish real data pairs (a, v) and two kinds of generated data pairs $(a, \hat{v})$, $(\hat{a}, v)$. The main objective of this module is to match the three joint distributions: $p(a, v), p(a, \hat{v}) \& p(\hat{a}, v)$ such that the generated fake data pairs $(a, \hat{v})$ and $(\hat{a}, v)$ are indistinguishable from the real data pairs (a, v). We follow an adversarial framework to match the joint distributions. The overall-cost function for joint discriminators is defined as:

$$\min_{G_A, G_I} \max_{D_1, D_2} C_M(D_1, D_2, G_A, G_I) = \mathbb{E}_{a \sim \mathcal{A}, v \sim \mathcal{V}}[\log D_1(a, v)]$$

$$+ \mathbb{E}_{\hat{a} \sim G_A, v \sim \mathcal{V}} \Big[\log \Big((1 - D_1(\hat{a}, v) D_2(\hat{a}, v)) \Big) \Big]$$

$$+ \mathbb{E}_{a \sim \mathcal{A}, \hat{v} \sim G_I} \Big[\log \Big((1 - D_1(a, \hat{v})(1 - D_2(a, \hat{v})) \Big) \Big]$$

$$\tag{2}$$

Here, $\mathcal{A}$ and $\mathcal{V}$ are the audio and video data distributions respectively. The objective of D_1 is to distinguish whether a sample pair is from real $p(a, v)$ or not, If the sample pair is not from $p(a, v)$, we use another discriminator D_2 to distinguish if the sample pair is from $p(\hat{a}, v)$ or $p(a, \hat{v})$. Both the discriminators $D_1 \& D_2$ work collectively, and the use of both of them implicitly define a ternary discriminative function D.
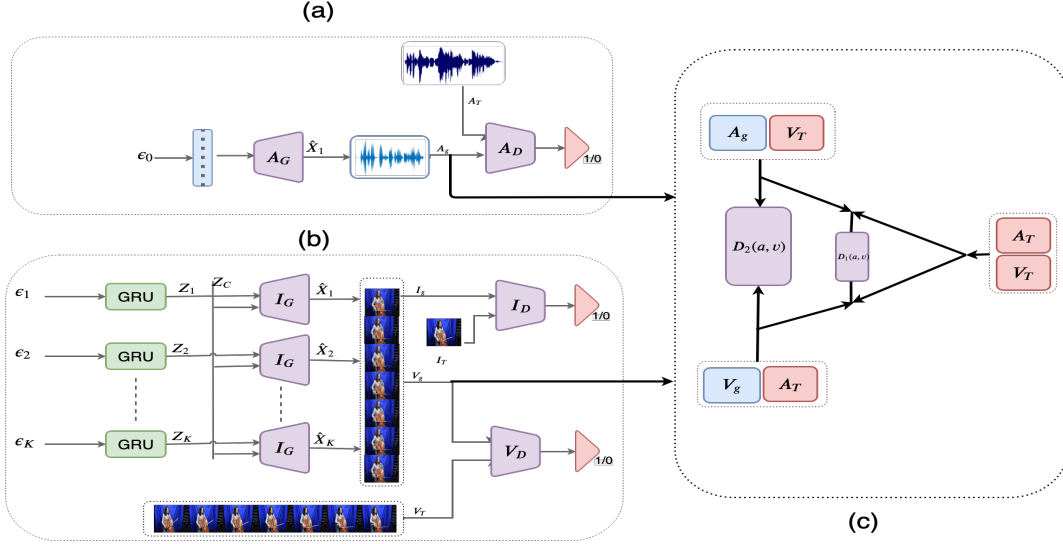
Fig. 2. In our proposed model, audio and video are generated using a latent noise vector. The joint discriminators are trained on joint distributions of audio-video pairs.

Table 1. Sub-URMP Video dataset

Cat.	Cello	D.Bass	Oboe	Sax	Trumpet	Viola	Bassoon	Clarinet	Horn	Flute	Trombone	Tuba	Violin
Train	881	123	394	696	100	648	135	716	475	511	787	323	707
Test	94	114	36	81	49	48	37	92	49	90	77	49	91

Table 2. Hand Motion Music Dataset(HMMD). Complete videos refer to raw video clips collected from Youtube for composing our HMMD dataset detailed in Section 4.1. We had an 80:20 train-test split for the composed dataset. AG:Acoustic Guitar

Category	Cello	Tabla	Ghatam	Erhu	AG	Xylophone	Violin
Videos	15	3	1	12	11	10	25
Frame	1251	317	454	519	1041	800	1847

4. EXPERIMENTS

4.1. Datasets

We compose two novel datasets to train and evaluate our models- Sub-URMP-Video Dataset, a video version of the Sub-URMP dataset [1] and Hand Motion Music Dataset (HMMD). To show the effectiveness of our method, we need videos with significant movement across frames and where audio is directly related of the motion happening in the video frames. We construct our own dataset (HMMD) comprising videos from Youtube focusing particularly on percussion instruments like Tabla, etc. which have significant hand movements across frames. Both the datasets and the code will be made publicly available.

Sub-URMP-Video Dataset: It contains 13 musical instrument categories. It contains image-audio pairs with ten images per second and audio of 0.5 seconds paired with each image. We form 10 frames per second(fps) video with 1 second of audio and match them to make it suitable for our task. We also manually remove the videos that contain different speakers. The basic information is summarized in Table 1.

HMMD: There are seven types of instruments, in total and 77 video clips. The number of clips ranges from 1 to 25 for various instruments. We preprocess the dataset to form videos of 1 sec each having a frame rate of 25 per sec. Audio streams are sampled at 16kHz. The proposed Hand Motion Music Dataset consists of 6229 video-audio pairs. The details of the dataset are present in Table 2. Some sample examples of the HMMD dataset are presented in supplementary at link [1].

4.2. Results

We compare our model with the baseline model using both quantitative and qualitative methods such as inception score, Statistically-Different Bins, pitch entropy, Frechet Inception Distance (FID) for audio and qualitative performance, along with a human evaluation for video and audio. We work with two modalities having separate architecture for generation. The video generation module is inspired from MoCoGAN [8], and for audio generation model, it is adapted from WaveGAN [29]. For video generation, we used a deconvolution layers based generator for producing 64x64 images from a 100-dim noise vector. To the best of our knowledge, there was no prior work in joint audio video generation as has been iterated in the paper and therefore the absence of a precise baseline. Consequently, we remove the joint discriminator component (c) from our model and consider the remaining model as a baseline for quantitative evaluations. and compare the effect of joint generation to isolated generations. Therefore, the baseline model we choose are the single modality generation components for audio and video respectively. The detailed algorithm and implementation details are presented in supplementary material[1].

---

[1]http://tiny.cc/gan

Table 3. Audio Evaluation Metrics for proposed models. BL: Baseline, JM:Proposed Joint Model

Dataset-Model	Human(↑)	NDB(↓)	FID(↓)	IS(↑)	PE(↓)
S-URMP(BL)	3.2	14	111.25	1.38	7.53
S-URMP(JM)	3.6	10	97.43	1.44	6.12
HMMD(BL)	3.1	36	93.00	1.34	7.33
HMMD(JM	3.4	29	79.89	1.48	5.42

Table 4. Video Evaluation Metrics for the proposed models

Dataset	Model	Human(↑)	IS(↑)
Sub-URMP	Ground Truth	4.6	1.42
Sub-URMP	Baseline	3.1	1.88
Sub-URMP	Joint Model	3.7	1.91
HMMD	Ground Truth	4.8	1.26
HMMD	Baseline	2.9	4.24
HMMD	Joint Model	3.5	4.33

4.3. Audio Performance Evaluation method

To validate the proposed model, we provide a quantitative analysis of the generated audio signals. The audio signals are generated using only noise as a latent vector, unlike the previously proposed methods [1, 2] that use a conditional vector for audio generation. We also compare the quality of the audio signal produced without the joint discriminator model. From this assessment, we can clearly say that the joint audio-video model helps to generate better audio samples.

4.4. Video Performance Evaluation

We also validate our model using the video generation task. The state-of-art video generation model [8] is considered as a baseline for our task which is a component of our joint model. We cannot compare our model with other video generation methods because they rely on conditional video generation which is a much easier task. The generated frames are shown in Figure- 3 for Sub-URMP dataset. We observe that the joint audio-video generation framework helps the model to learn better video frames. We observe that the inception score improves 0.3% for Sub-URMP and 0.9% for HMMD dataset.

Fig. 3. Results on Sub-URMP Dataset:In this figure, the first row indicates the given ground truth video. The second row indicates generated videos for our baseline model and third row shows generated video through our joint model

5. ANALYSIS

The joint model outperforms the baseline models in all our evaluations-in Table 2 as well as Table 4. We have designated alongside every evaluation metric with arrows, where up-arrow(↑) indicates that a larger value is better and similarly the down-arrow(↓) suggests that a lower value is better.

Comparison with Deep Cross model generation: In Deep cross-modal generation [1], audio and image generations are conditional in nature. But in our proposed model, audio and video are generated from a latent noise vector. The former also uses the music class label for generation while our work is completely unsupervised. Another difference is that we are generating videos (a sequence of image frames), which is a more challenging task, unlike the previous work which generates only images.

Comparison with CMC GAN: The CMC-GAN [2] uses the cycle consistency between generated audio and video samples. However, this work too relies on conditional generation of audio and images data. It faces the same drawback as the previous one as it works only on images and not on videos whereas our model generates audio and video modalities collaboratively without conditioning on each other.

Comparison with Triangle GAN: The Triangle GAN [30] model too does conditional generation by joint training of the modalities. Again, our work does not use any conditional dependence rather it jointly learns the audio-video generation.

Comparison with independent modality Generation model: The other video generation models such as MoCoGAN [8] do not consider audio modality for generating videos. Similarly audio generation models such as WaveGAN [31] also generate audio without considering the video information. Therefore, to the best of our knowledge, there are no other models available that can generate audio and video simultaneously by jointly training the generators.

Comparison with SA-CMGAN: In SA-CMGAN [28] audios and images are generated using the conditional GAN, while we generate videos and audio, which is more challenging using the joint learning. They also use the class label while training the model. In contrast to it, proposed framework is completely unsupervised.

6. CONCLUSION

This paper proposes a collaborative generation paradigm for audio and video generation without any conditioning involved. It can also be trained for an end to end fashion for better convenience. The generated videos and audio are realistic, as is manifested by numerous experiments as well as human evaluation. The results clearly depict that collaborative generation successfully takes advantage of the subtle relationship that video stream has with the audio stream to generate naturalistic images and improved quality audio. The proposed two datasets will immensely help the community in audio-visual research. Using collaborative generation of related modalities to improve the quality of individual modalities generated is a promising direction which we have initiated through this work. Hence, we believe that this paper will act as a catalyst for work in multimodal generation.

7. REFERENCES

[1] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in *ACM Multimedia Workshops*, 2017.

[2] Wangli Hao, Zhaoxiang Zhang, and He Guan, "Cmcgan: A uniform framework for cross-modal visual-audio mutual generation," in *AAAI*, 2018.

[3] I. Goodfellow, J. P.-Abadie, M. Mirza, B. Xu, S. W.-Farley, D.and Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.

[4] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," 2018.

[5] A Radford, L Metz, and S Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv:1511.06434*, 2015.

[6] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz, "Multimodal unsupervised image-to-image translation," in *(ECCV)*, 2018, pp. 172–189.

[7] C. Wan, S.-Po Chuang, and H.-Yi Lee, "Towards audio to scene image synthesis using generative adversarial network," in *ICASSP*, 2019.

[8] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz, "Mocogan: Decomposing motion and content for video generation," in *CVPR*, 2018.

[9] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck, "Music transformer: Generating music with long-term structure," 2018.

[10] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts, "Gansynth: Adversarial neural audio synthesis," 2018.

[11] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein, "Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 112, 2018.

[12] David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass, "Jointly discovering visual objects and spoken words from raw sensory input," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 649–665.

[13] Chris Donahue, Bo Li, and Rohit Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *(ICASSP)*. IEEE, 2018, pp. 5024–5028.

[14] Santiago Pascual, Antonio Bonafonte, and Joan Serrà, "Segan: Speech enhancement generative adversarial network," *Proc. Interspeech 2017*, 2017.

[15] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," in *9th ISCA Speech Synthesis Workshop*.

[16] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan, "Neural audio synthesis of musical notes with wavenet autoencoders," in *ICML*, 2017.

[17] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg, "Visual to sound: Generating natural sound for videos in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3550–3558.

[18] C. Young Lee, A. Toffy, G. J. Jung, and W.J. Han, "Conditional wavegan," *arXiv preprint arXiv:1809.10636*, 2018.

[19] Relja Arandjelovic and Andrew Zisserman, "Objects that sound," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 435–451.

[20] A. Senocak, Tae-Hyun Oh, J. Kim, Ming-Hsuan Yang, and In So Kweon, "Learning to localize sound source in visual scenes," in *CVPR*, 2018, pp. 4358–4366.

[21] Ruohan Gao, Rogerio Feris, and Kristen Grauman, "Learning to separate object sounds by watching unlabeled video," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 35–53.

[22] Andrew Owens and Alexei A Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 631–648.

[23] Tae-Hyun Oh, T. Dekel, C. Kim, I. Mosseri, W. T Freeman, M. Rubinstein, and W. Matusik, "Speech2face: Learning the face behind a voice," in *CVPR*, 2019, pp. 7539–7548.

[24] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba, "Foley music: Learning to generate music from videos," *ECCV*, 2020.

[25] Kun Su, Xiulong Liu, and Eli Shlizerman, "Audeo: Audio generation for a silent performance video," *NeurIPS*, 2020.

[26] Peihao Chen, Yang Zhang, Mingkui Tan, Hongdong Xiao, Deng Huang, and Chuang Gan, "Generating visually aligned sound from videos," *IEEE Transactions on Image Processing*, vol. 29, pp. 8292–8302, 2020.

[27] Anoop Cherian, Moitreya Chatterjee, and Narendra Ahuja, "Sound2sight: Generating visual dynamics from sound and context," *ECCV*, 2020.

[28] Huadong Tan, Guang Wu, Pengcheng Zhao, and Yanxiang Chen, "Spectrogram analysis via self-attention for realizing cross-model visual-audio generation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4392–4396.

[29] Chris Donahue, Julian McAuley, and Miller Puckette, "Adversarial audio synthesis," 2018.

[30] Zhe Gan, Liqun Chen, Weiyao Wang, Y. Pu, Y. Zhang, Hao Liu, Chunyuan Li, and Lawrence Carin, "Triangle generative adversarial networks," in *NIPS*, 2017.

[31] Chris Donahue, Julian McAuley, and Miller Puckette, "Synthesizing audio with generative adversarial networks," *arXiv preprint arXiv:1802.04208*, 2018.