Listen and Move: Improving GANs Coherency in Agnostic Sound-to-Video Generation

Rafael Redondo
Eurecat, Centre Tecnològic de Catalunya, Tecnologies Multimèdia
Barcelona, 08005, Spain
rafael.redondo@eurecat.org

Abstract

Deep generative models have demonstrated the ability to create realistic audiovisual content, sometimes driven by domains of different nature. However, smooth temporal dynamics in video generation is a challenging problem. This work focuses on generic sound-to-video generation and proposes three main features to enhance both image quality and temporal coherency in adversarial models: a triple sound routing, a multi-scale residual and dilated recurrent network for extended sound analysis, and a novel recurrent and directional convolutional layer for video prediction. Each of the proposed features improves, in both quality and coherency, the baseline neural architecture typically used in the SoTA.

1. Introduction

After the invention of Generative Adversarial Networks (GANs) [15], the emergence of conditionals GANs [35, 52, 20] have resulted in a diversity of cross-modal synthesis by mapping data between different domains. Some examples are text-to-image [52, 49, 57, 29, 51], text-to-video [32, 42, 1, 11, 24, 17], speech-to-face [44, 8], video-to-sound [28, 55, 10, 6], or video-to-video[47].

This paper focuses on generic sound-to-video translation, which entails a highly unconstrained problem. Certainly, multiple visual variations are plausible for a single sound. Video synthesis must deal also with pixel jittering and audio-visual lagging. In contrast to other modalities, such as speech-to-face, no prior distribution is assumed for either image or sound. Compared to pose-based approaches, this could tackle situations in which the human pose is difficult to infer, such as a face profile, occluding objects, or even an entire music ensemble performance.

2. Related works

In recent years, the ability of GANs to translate from audio to image has been demonstrated [9, 18, 45, 50, 12], yet frames are independently synthesized without temporal coherency. A common practice to leverage GANs from image to video is to additionally use a 3D video discriminator [36, 28, 39, 42, 1, 11, 54, 24, 39, 8, 44].

Another common approach to induce temporal coherency consists in feeding the generator with a series of noise vectors temporally encoded by a Recurrent Neural Network (RNN). In particular, MoCoGAN [42] —recently extended in [38]— proposed a novel framework for unconditional video generation disentangling motion from content. Other methods [41, 26] presented a particular feedforward scheme to map motion trajectories in a latent space pre-trained for still images, which ultimately fails to represent fine-grained temporal dynamics. Surprisingly, dilated recurrent schemes have been omitted for GANs hitherto.

In parallel, some optimizations have been proposed in speech-driven video synthesis, specially tailored for facial attributes [7, 39, 39, 34, 33, 8, 54, 44]. Most of these approaches use a reference frame, which reduces image uncertainty but ultimately produces rigid, unnatural movements. Other approaches have exploited body landmarks to assist in music-driven video performances [56]. However, these methods cannot be applied in its entirety for generic (faceless or bodyless) sound-to-video synthesis.

To induce video coherency, vid2vid [47] relies on feeding previously synthesized frames, unfeasible for sound-to-video. Fortunately, the apparition of ConvRNNs [37, 40] has enabled an efficient spatial analysis and temporal coherence [2, 30, 8, 56]. Inspired by [4], a novel directional and causal ConvGRU for video prediction is proposed here.

Contributions This work aims to improve image quality and temporal coherency of generic sound-to-video GANs through (1) a more versatile triple sound routing for motion encoding, content representation, and conditional normalization layers, (2) a residual multi-scale DilatedRNN for an extended listening range, and (3) a multi-orientation video prediction layer built upon a novel Directional ConvGRU.

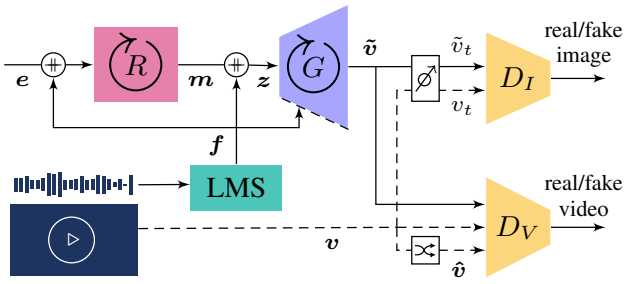

Figure 1: Main architecture with a triple sound feature routing, a residual multi-scale DilatedRNN (R), and Directional ConvRNNs in the generator (G).

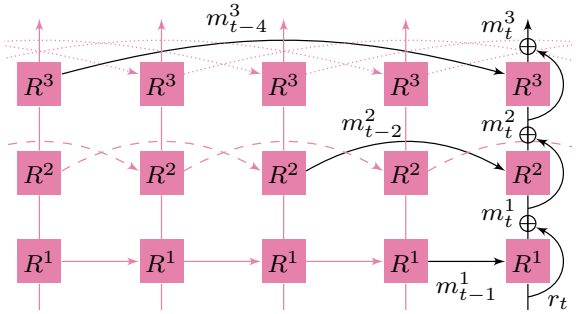

Figure 2: Motion encoding: a residual 3-layer DilatedRNN.

3. Main architecture

The proposed model, illustrated in Fig. 1, is mainly made of a recurrent neural network R, a generator G, and two discriminators for image D_I and video D_V.

3.1. Sound representation

The waveform is effectively converted into $B = 64$ log mel frequency (LMS) bins, with a 25 ms window length and 10 ms hop size. It is then split in chunks of 85 ms. The effective temporal resolution will eventually be constrained by the video frame rate, hereafter fixed to 50 ms or 20 fps. The sound features $\boldsymbol{f}$ are finally calculated by averaging chunks over time. Note that chunk overlapping —in this case 37.5%— already encourages temporal coherency.

3.2. Audio temporal coherency

Routing types in Fig. 1 are not either mutually exclusive or redundant, in fact their combination is beneficial, as shown later. Hereafter, let $\boldsymbol{m} = R(\boldsymbol{r}) \in \mathbb{R}^{T\times M}$ denote a T-enrolled and compact motion encoding of size M.

Sound as content. Let $\boldsymbol{r} := \boldsymbol{e} = \{e_t \in \mathbb{R}^E | t \in [1..T]\}$ be a T-length sequence of random vectors $\mathcal{N}(0, 1)$, where typically $M = E$. Here, sound features $\boldsymbol{f} = \{f_t \in \mathbb{R}^B | t \in [1..T]\}$ replace content noise, so that the generator's input is concatenated across the temporal dimension $\boldsymbol{z} := [\boldsymbol{f}, \boldsymbol{m}] \in \mathbb{R}^{T\times Z}$, with $Z = B+E$. In this way, the generator has direct access to the raw sound features at each time step.

Motion from sound. Sound features are recurrently encoded, sometimes concatenated with motion noise as $\boldsymbol{r} := [\boldsymbol{f}, \boldsymbol{e}]$ of size $M = B + E$, where simply $\boldsymbol{z} := \boldsymbol{m}$. In this way, R has to infer motion from sound, being this the preferred configuration [39, 44, 39, 8, 27, 1, 26].

Dilated and residual recurrency. A multi-layer RNN with dilated skip connections is proposed to deal with more complex temporal dynamics of sound, which sometimes unfolds at different resolution speeds. Also, residual connections facilitate the propagation of audio features through the network, see Fig. 2. Let's reformulate it as:

$$m_t^l = \varphi(R^l(m_t^{l-1}, m_{t-d}^l) + m_t^{l-1}) \quad (1)$$

where m_t^l is the output at time step t of the layer $l \in [1..L_R]$ with dilation factor $d = 2^l$. φ stands for LeakyReLU. For instance, by stacking $L_R = 3$ layers with the Sec. 3.1 settings, audio hops are 50, 100, and 200 ms.

3.3. The adversarial networks

The generator and both discriminators are built upon a series of feedforward layers, made of convolutional, normalization, and activation layers. While D_I and D_V use batch normalization, G additionally uses noise injection [13, 22] and audio conditional instance normalization.

Audio conditional generator. To reinforce access to conditional data traveling through the generator, class vectors are usually embedded to modulate normalization layers [22, 3], also exploited in conditional video generation [41, 38, 26]. Here, sound features are encoded into class feature vectors, ultimately modulating an instance normalization layer.

Image and video discriminators. D_V is built on 3D kernels, one temporal and two spatial dimensions equally sampled at each layer. Note the temporal dimension can be downsampled $\lfloor log_2(T) \rfloor$ times. D_I receives a random frame $\tilde{v}_t$ from the synthetic (fake) video sequence $\tilde{\boldsymbol{v}}$. The same random index is used to pick up a real frame v_t from the source (real) sequence $\boldsymbol{v}$. D_V receives real $\boldsymbol{v}$ and fake $\tilde{\boldsymbol{v}}$ video sequences, and shuffled versions $\hat{\boldsymbol{v}}$ [1].

3.4. Video temporal coherency

A novel recurrent multi-directional convolutional layout is proposed for causal video prediction. The activation distribution at time step t can be expressed as $x_t^d \sim p(x_t | x_{t-1}^d)$, where x are the hallucinated activations of a generative layer and x^d directional predictions. As illustrated in Fig. 3, the video prediction layer comprises

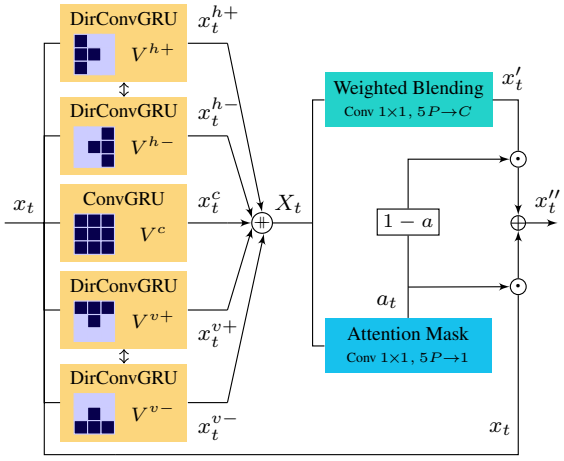

Figure 3: Video prediction layer (kernel size 3). Spatial predictions are channel-wise concatenated X_t and merged x'_t. Previous hallucinated activations x_t contribute to the output x''_t according to an auto-regressive mask a_t.

4-directional ConvGRUs (DirConvGRU), predicting positive and negative motion in horizontal and vertical directions, and a squared-centered ConvGRU dealing with motion along the camera axis.

The video prediction activations, after the outermost generative layer, are finally concatenated as $X_t = [x_t^{h+}, x_t^{h-}, x_t^{c}, x_t^{v+}, x_t^{v-}]$, where $x_t^d = V^d(x_t, x_{t-1}^d) \in \mathbb{R}^{P\times N\times N}$ is the t-th hidden state, also output, of a K-size directional or centered ConvGRU for $d \in \{h+, h-, c, v+, v-\}$. Next, a weighted blending layer merges all directional predictions [4] and predicts an auto-regressed attention mask [34, 8, 47, 5] to merge both the hallucinated x_t and predicted x'_t pixels as follows:

$$\begin{aligned} x'_t &= \varphi(X_t^\intercal \cdot W_x + b_x) \\ a_t &= \sigma(X_t^\intercal \cdot W_a + b_a) \\ x''_t &= a_t \odot x_t + (1 - a_t) \odot x'_t \end{aligned} \tag{2}$$

where the blending $W_x \in \mathbb{R}^{(5\times P)\times C\times N\times N}$ and masking weights $W_a \in \mathbb{R}^{(5\times P)\times 1\times N\times N}$ are implemented as P-channel 1×1 convolutions ($P = C$ for convenience). The activation functions σ and φ are sigmoid and the identity, and $\odot$ denotes channel-wise Hadamard product.

4. Experiments

Successful setups needed about a training day (50k iterations) for 128×128 and about 3 days (100k iterations) for 256×256 to reach a decent image quality[1]. Full-model higher resolutions required far beyond 24GB GPU-memory and too long training runs that were finally discarded.

[1]Intel(R) Core(TM) i9-10900X CPU @ 3.70GHz and NVIDIA GeForce RTX 3090.

4.1. Hyper-parameters setup

Batches have $L = 32$ frames forwarded one at a time. Random noise vectors are $E = 2$ (just a guess). All recurrent layers have size $M = 66$, so eventually G ingests vectors of size $Z = 130$. Image resolution 256×256 is achieved with 6 generative layers. The outermost one has 16 channels, incremented in powers of two up to 512 maximum. Images are $[-1, 1]$ normalized. R uses normal initialization and zeroed init states (no prior). Convolutional weights receive random initialization from $\mathcal{N}(0, 0.02)$, while DirConvGRUs use orthogonal initialization. All LeakyReLUs have 0.2 negative slope. TTUR learning rates are heuristically set to 10^{-4} for G and R, while D_I and D_V updates at $4\cdot 10^{-4}$, without scheduling. The Adam optimizers [25] have momentums of 0.3 and 0.999. Here, the Wasserstein loss with gradient penalty [16] and the perceptual loss [21] were distinctive to provide training stability and image quality.

4.2. Model assessment

Text- or pose-based sound-to-video methods are not directly comparable within our aim. Furthermore, this study entails a particular one-shoot training on a single video. Therefore, a baseline implementation, with a common MfS scheme and the SOTA advances herein described, is used to compare with each of the features proposed in this work.

Figure 4: Artifacts produced by a 512×512 vanilla GAN with (left) skip-connections, (middle) residual connections, and (right) a residual G and skip-connected Ds.

Skip vs. residual connections. In contrast to observations in [23], layer connection has a strong impact on the synthesis quality. As illustrated in Fig. 4, skip connections in the generator tend to produce large colored patches, increasing their presence and intensity as the input audio deviates from the original audio. Discriminators built upon residual connections often produce severe blurring and erase some details. Instead, a residual generator and skip-connected discriminators usually produce sharper images.

Ablation study. A series of experiments were conducted on diverse audiovisual content of various minutes in length,

[2]Cello: https://www.youtube.com/watch?v=zah9B0toTBQ
[3]Quintet: https://www.youtube.com/watch?v=R-Tk7-Ytes4
[4]Drums: https://www.youtube.com/watch?v=7H4hrsb_0tE&t=59s
[5]Talking head: MEAD M31 [46].

		Cello			*Quintet*			*Drums*			*Talking Head*		
		SSIM (↑)	LPIPS (↓)	FVD (↓)	SSIM (↑)	LPIPS (↓)	FVD (↓)	SSIM (↑)	LPIPS (↓)	FVD (↓)	SSIM (↑)	LPIPS (↓)	FVD (↓)
	Baseline (MfS)	0.82 ±0.06	0.16 ±0.06	2219 ±1095	0.69 ±0.11	0.24 ±0.10	6480 ±2543	0.68 ±0.02	0.16 ±0.02	2387 ±976	0.43 ±0.03	0.62 ±0.02	1082 ±349
Baseline +	R+D-RNN	0.87 ±0.03	0.09 ±0.02	1184 ±905	0.89 ±0.02	0.05 ±0.02	1706 ±2537	0.75 ±0.03	0.12 ±0.02	2230 ±1276	0.81 ±0.03	0.07 ±0.02	520 ±555
	SaC	0.87 ±0.03	0.09 ±0.02	1354 ±980	0.89 ±0.02	0.05 ±0.02	1775 ±2701	0.75 ±0.03	0.12 ±0.02	2127 ±1977	0.77 ±0.03	0.07 ±0.01	366 ±470
	acIN	0.87 ±0.03	0.08 ±0.02	1426 ±1195	0.88 ±0.02	0.05 ±0.01	1903 ±1866	0.76 ±0.03	0.11 ±0.03	2122 ±2857	0.69 ±0.02	0.11 ±0.01	486 ±362
All Abv. +	ConvGRU	0.87 ±0.03	0.08 ±0.02	1101 ±1011	0.90 ±0.02	0.05 ±0.01	1502 ±2090	0.75 ±0.03	0.11 ±0.02	1700 ±1603	0.77 ±0.03	0.06 ±0.01	327 ±435
	DirConvGRU	0.87 ±0.03	0.08 ±0.02	1085 ±991	0.88 ±0.02	0.06 ±0.02	1455 ±1999	0.75 ±0.03	0.11 ±0.02	1614 ±1390	0.80 ±0.03	0.05 ±0.01	280 ±414

Table 1: Ablation study at 128×128 resolution on diverse audiovisual content: a cello[2](melodic), a classic quintet[3](harmonic), drums[4](percussive), and a side-view talking head[5](speech). With motion from sound (MfS) as a baseline, the sound routing features are activated one at a time: Residual+Dilated-RNN, sound as content (SaC), and audio conditional instance normalization (acIN). All routing features are activated in the video prediction group to compare between a basic ConvGRU and the proposed DirConvGRUs ($K=3$). Averages calculated over batches of the 20%-split source video.

FID (↓)	*Cello*	*Quintet*	*Drums*	*Talking Head*
Baseline (MfS)	320 ±14	423 ±20	385 ±14	528 ±16
Full Model	279 ±24	405 ±29	378 ±28	535 ±42

Table 2: Baseline (MfS) and full model sound robustness comparison for the videos in Tab. 1. FID averaged over 500 audio clips of 1-5 s randomly selected from FSD50K [14].

as summarized in Tab. 1. Two perceptual objective image quality metrics, SSIM [48] and LPIPS [53], and one video quality metric FVD[6] [43] were used. We observed a clear tendency to objectively improve not only image but also video quality by adding independently each proposed feature. Moreover, video prediction layers, and in particular DirConvGRU more than a basic ConvGRU, provided an extra boost of temporal enhancement.

Robustness to sound replacement. In order to evaluate the generalization capacity, when the distribution of feed-forward sound deviates from the training source, random audio clips were selected from FSD50K [14]. Since large variations in distribution are plausibly expected, FID [19] can measure distortions by comparing how far real and synthetic images are at high-level representation. From the results in Tab. 2, the full model proves more robustness on in-the-wild sound replacement when trained on music, while apparently it is reduced for speech.

Qualitative results. Our model is able to synthesize expressive long video sequences by synchronously responding to audio without apparent degradation over time. Some illustrative examples are shown in Fig. 5. The combined sound routing significantly reduces artifacts, especially when the input sound deviates from the original audio distribution, in accordance with Tab. 1 and Tab. 2. Also, the video prediction layer tends to generate smoother motions and reduced flickering in general. Nevertheless, the persistence of temporal artifacts certainly affects the overall realism, specially for complex or sound-uncorrelated visual dynamics.

[6]From this implementation [38].

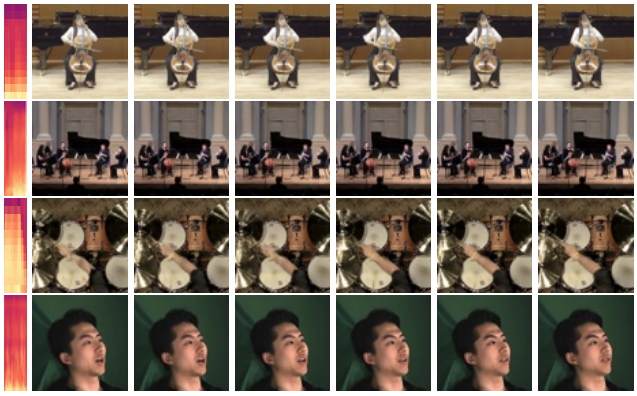

Figure 5: Examples of consecutive frames synthesized by our full-model at 256×256 resolution and 20fps fed with validation audio samples of the videos in Tab. 1.

5. Discussion

In this work various strategies to improve temporal coherence of sound-to-video GANs have been proposed, which can be straightforwardly applied to other domains. Many other strategies were tested, including sound encoding [31, 45], conditioning augmentation [52], spectral and pixel normalization, joint audio-image discriminators, or feature matching loss [47, 5].

Limitations. While our method struggles to reproduce large and complex motion dynamics, specially those unrelated to sound, it is particularly demanding on both computation and memory despite our amenable DirConvRNN.

Broader impact. Generative models like this can certainly leverage smart tools for audiovisual content creation. On the other hand, there is still a significant gap to achieve high-quality realistic videos. Even so, it may already raise important ethical concerns on inappropriate usage.

Acknowledgements

This work was financially supported by the Catalan Government through the funding grant ACCIÓ-Eurecat (Project PRIV - DeepArts).

References

[1] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*, volume 1, page 2, 2019. 1, 2

[2] Nicolas Ballas, Li Yao, Chris Pal, and Aaron C. Courville. Delving deeper into convolutional networks for learning video representations. In *4th International Conference on Learning Representations*, 2016. 1

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *7th International Conference on Learning Representations*, 2019. 2

[4] Wonmin Byeon, Qin Wang, Rupesh Kumar Srivastava, and Petros Koumoutsakos. Contextvp: Fully context-aware video prediction. In *ECCV*, 2018. 1, 3

[5] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. Everybody dance now. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5932–5941, 2019. 3, 4

[6] Kan Chen, Chuanxi Zhang, Chen Fang, Zhaowen Wang, Trung Bui, and Ram Nevatia. Visually indicated sound generation by perceptually optimized classification. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1

[7] 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

[8] Lele Chen, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7824–7833, 2019. 1, 2, 3

[9] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 349–357, 2017. 1

[10] Peihao Chen, Yang Zhang, Mingkui Tan, Hongdong Xiao, Deng Huang, and Chuang Gan. Generating visually aligned sound from videos. *IEEE Transactions on Image Processing*, 29:8292–8302, 2020. 1

[11] Kangle Deng, Tianyi Fei, Xin Huang, and Yuxin Peng. Irc-gan: Introspective recurrent convolutional gan for text-to-video generation. In *IJCAI*, pages 2216–2222, 2019. 1

[12] Bin Duan, Wei Wang, Hao Tang, Hugo Latapie, and Yan Yan. Cascade attention guided residue learning gan for cross-modal translation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 1336–1343. IEEE, 2021. 1

[13] Ruili Feng, Deli Zhao, and Zheng-Jun Zha. Understanding noise injection in gans. In *International Conference on Machine Learning*, pages 3284–3293. PMLR, 2021. 2

[14] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2022. 4

[15] 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. 1

[16] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017. 3

[17] Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin Minaee, Dimitris Metaxas, and Sergey Tulyakov. Show me what and tell me how: Video synthesis via multimodal conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3615–3625, 2022. 1

[18] Wangli Hao, Zhaoxiang Zhang, and He Guan. Cmcgan: A uniform framework for cross-modal visual-audio mutual generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 1

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017. 4

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 3

[22] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[23] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 3

[24] Doyeon Kim, Donggyu Joo, and Junmo Kim. Tivgan: Text to image to video generation with step-by-step evolutionary generator. *IEEE Access*, 8:153113–153122, 2020. 1

[25] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, 2015. 3

[26] Seung Hyun Lee, Gyeongrok Oh, Wonmin Byeon, Jihyun Bae, ChanYoung Kim, Wonjae Ryoo, Sang Ho Yoon, Jinkyu Kim, and Sangpil Kim. Sound-guided semantic video generation. In *ECCV*, 2022. 1, 2

[27] Jiguo Li, Xinfeng Zhang, Chuanmin Jia, Jizheng Xu, Li Zhang, Y. Wang, Siwei Ma, and Wen Gao. Direct speech-to-image translation. *IEEE Journal of Selected Topics in Signal Processing*, 14:517–529, 2020. 2

[28] Yitong Li, Martin Renqiang Min, Dinghan Shen, David Edwin Carlson, and Lawrence Carin. Video generation from text. In *AAAI*, 2018. 1

[29] Jiadong Liang, Wenjie Pei, and Feng Lu. Cpgan: Content-parsing generative adversarial networks for text-to-image synthesis. In *ECCV*, 2020. 1

[30] Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P. Xing. Dual motion gan for future-flow embedded video prediction. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1762–1770, 2017. 1

[31] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 4

[32] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. To create what you tell: Generating videos from captions. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1789–1798, 2017. 1

[33] Hai Xuan Pham, Samuel Cheung, and Vladimir Pavlovic. Speech-driven 3d facial animation with implicit emotional awareness: A deep learning approach. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2328–2336, 2017. 1

[34] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 818–833, 2018. 1, 3

[35] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016. 1

[36] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *Proceedings of the IEEE international conference on computer vision*, pages 2830–2839, 2017. 1

[37] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. 1

[38] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3626–3636, 2022. 1, 2, 4

[39] Yang Song, Jingwen Zhu, Dawei Li, Andy Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 919–925. AAAI Press, 2019. 1, 2

[40] Marijn F Stollenga, Wonmin Byeon, Marcus Liwicki, and Juergen Schmidhuber. Parallel multi-dimensional lstm, with application to fast biomedical volumetric image segmentation. *Advances in neural information processing systems*, 28, 2015. 1

[41] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N. Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. In *International Conference on Learning Representations*, 2021. 1, 2

[42] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. 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

[43] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *ArXiv*, abs/1812.01717, 2018. 4

[44] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 128:1398–1413, 2019. 1, 2

[45] Chia-Hung Wan, Shun-Po Chuang, and Hung yi Lee. Towards audio to scene image synthesis using generative adversarial network. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 496–500, 2019. 1, 4

[46] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI*, pages 700–717. Springer, 2020. 3

[47] 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. 1, 3, 4

[48] Zhou Wang, Alan Conrad Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004. 4

[49] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 1

[50] Pei-Tse Yang, Feng-Guang Su, and Yu-Chiang Frank Wang. Diverse audio-to-image generation via semantics and feature consistency. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1188–1192. IEEE, 2020. 1

[51] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 833–842, 2021. 1

[52] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stack-

gan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 1, 4

[53] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 4

[54] 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

[55] 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*, pages 3550–3558, 2018. 1

[56] Hao Zhu, Yi Li, Feixia Zhu, Aihua Zheng, and Ran He. Let's play music: Audio-driven performance video generation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3574–3581. IEEE, 2021. 1

[57] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5795–5803, 2019. 1