Sound2Sight: Generating Visual Dynamics from Sound and Context

Moitreya Chatterjee*[1] Anoop Cherian[2]

[1] University of Illinois at Urbana-Champaign, Urbana IL 61801, USA
[2] Mitsubishi Electric Research Laboratories, Cambridge MA 02139, USA
metro.smiles@gmail.com cherian@merl.com

Abstract. Learning associations across modalities is critical for robust multimodal reasoning, especially when a modality may be missing during inference. In this paper, we study this problem in the context of audio-conditioned visual synthesis – a task that is important, for example, in occlusion reasoning. Specifically, our goal is to generate future video frames and their motion dynamics conditioned on audio and a few past frames. To tackle this problem, we present *Sound2Sight*, a deep variational encoder-decoder framework, that is trained to learn a per frame stochastic prior conditioned on a joint embedding of audio and past frames. This embedding is learned via a multi-head attention-based audio-visual transformer encoder. The learned prior is then sampled to further condition a video forecasting module to generate future frames. The stochastic prior allows the model to sample multiple plausible futures that are consistent with the provided audio and the past context. Moreover, to improve the quality and coherence of the generated frames, we propose a multimodal discriminator that differentiates between a synthesized and a real audio-visual clip. We empirically evaluate our approach, vis-á-vis closely-related prior methods, on two new datasets viz. (i) Multimodal Stochastic Moving MNIST with a Surprise Obstacle, (ii) Youtube Paintings; as well as on the existing Audio-Set Drums dataset. Our extensive experiments demonstrate that Sound2Sight significantly outperforms the state of the art in the generated video quality, while also producing diverse video content.

1 Introduction

Evolution has equipped the intelligent species with the ability to create mental representations of sensory inputs and make associations across them to generate world models [9]. Perception is the outcome of an inference process over this world model, when provided with new sensory inputs. Consider the following situation. You see a kid going into a room which is occluded from your viewpoint, however after sometime you hear the sound of a vessel falling down, and soon enough, a heavy falling sound. In the blink of an eye, your mind simulates a large number of potential possibilities that could have happened in that room; each

* Work done as an intern at MERL.

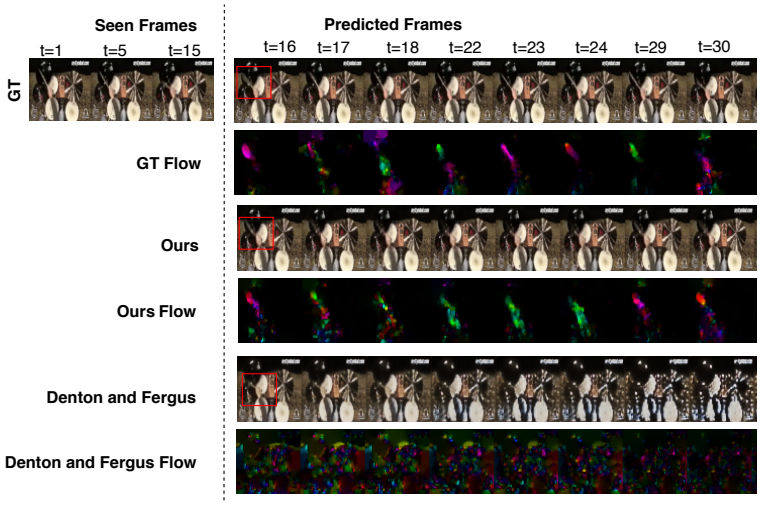


Fig. 1: Video generation using our Sound2Sight against Denton and Fergus [10] on AudioSet-Drums [14]. We also show the optical flow between consecutive generated frames. The red square indicates the region of dominant motion.

simulation considered for its coherence with the sound heard, and its urgency or risk. From these simulations, the most likely possibility is selected to be acted upon. Such a framework that can synthesize modalities from other cues is perhaps fundamental to any intelligent system. Efforts to understand such mental associations between modalities dates back to the pioneering work of Pavlov [43] (on his drooling dogs) who proposed the idea of *conditioning* on sensory inputs.

In this paper, we explore this multimodal association problem in the context of generating plausible visual imagery given the accompanying sound. Specifically, our goal is to build a world model that learns associations between audio and video dynamics in such a way as to infer visual dynamics when only the audio modality (and the visual context set by a few initial frames) is presented to the system. As alluded to above, such a problem is fundamental to occlusion reasoning. Apart from this, it could help develop assistive technologies for the hearing-impaired, could enable a synergy between video and audio inpainting technologies [28,66], or could even compliment the current "seeing through corners" methods [35,65] using the audio modality.

From a technical standpoint, the task of generating the pixel-wise video stream from only the audio modality is severely ill-posed. For instance, a drummer playing a drum to a certain beat would sound the same irrespective of the color of his/her attire. To circumvent this challenge, we condition our video generator using a few initial frames. This workaround not only permits the generation of videos that are pertinent to the situation, but also allows the model

to focus on learning the dynamics and interactions of the visual cues assisted by audio. There are several recent works in a similar vein [7,56,6] that explore speech-to-video synthesis to generate talking heads, however they do not use the past visual context or assume very restricted motion dynamics and audio priors. On the other hand, methods that seek to predict future video frames [10,55,13] given only the past frames, assume a continuity of the motion pattern and are unable to adapt to drastic changes in motion that might arise in the future (e.g., the sudden movements of the drummer in Figure 1). We note that there also exist several recent works in the audio-visual synthesis realm, such as generating audio from video [27,63,64] that looks at a complementary problem and multimodal generative adversarial networks (GAN) that generates a single image rather than forecasting the video dynamics [8,18,58].

To tackle this novel task, we present a stochastic deep neural network: *Sound2 Sight*, which is trained end-to-end. Our main backbone is a conditional variational autoencoder (VAE) [30] that captures the distribution of the *future video frames* in a latent space. This distribution is used as a prior to subsequently condition a video generation framework. A key question that arises then, is how to incorporate the audio stream and its correlations with the video content? We propose to capture this synergy within the prior distribution - through a joint embedding of the audio features and the video frames. The variance of this prior distribution, permits diversity in the video generation model, thereby synthesizing disparate plausible futures.

An important component in our setup is the audio-visual latent embedding that controls the generation process. Inspired by the recent success of transformer networks [53], we propose an adaptation of multi-head transformers to effectively learn a multimodal latent space through self-attention. As is generally known, pixel generations produced using variational models often lack sharpness, which could be attributed to the Euclidean loss typically used [32]. To this end, in order to improve the generated video quality, we further propose a novel *multimodal discriminator*, that is trained to differentiate between real audio-visual samples and generated video frames coupled with the input audio. This discriminator incorporates explicit sub-modules to verify if the generated frames are realistic, consistent, and synchronized with the audio.

We conduct experiments on three datasets, two new multimodal datasets: (i) Multimodal Stochastic Moving MNIST with a Surprise Obstacle (M3SO) and (ii) Youtube-Painting, alongside a third dataset – AudioSet-Drums – which is an adaptation of the well-known AudioSet datset [14]. The M3SO dataset is an extension of stochastic moving MNIST [10], however incorporates audio based on the location and identity of the digits in the video, while also including a surprise component that requires learning audio-visual synchronization and stochastic reasoning. The Youtube-Painting dataset is created by crawling Youtube for painting videos and provides a challenging setting for Sound2Sight to associate painting motions of an artist and the subtle sounds of brush strokes. Our experiments on these datasets show that Sound2Sight leads to state-of-the-art performances in quality, diversity, and consistency of the generated videos.

Before moving on, we summarize below the key contributions of this paper.

- We study the novel task of future frame generation consistent with the given audio and a set of initial frames.
- We present *Sound2Sight*, a novel deep variational multimodal encoder-decoder for this task, that combines the power of VAEs, GANs, and multimodal transformers in a coherent learning framework.
- We introduce three datasets for evaluating this task. Extensive experiments are provided, demonstrating state-of-the-art performances, besides portraying diversity in the generation process.

2 Related Works

In this section, we review prior works that are closely related to our approach.
Audio-Visual Joint Representations: The natural co-occurrence of audio-and-visual cues is used for better representation learning in several recent works [1,3,20,39,40,41]. We too draw upon this observation, however, our end-goal of future frame generation from audio is notably different and manifests in our proposed architecture. For example, while both [1] and [39] propose a common multimodal embedding layer for video representation, our multimodal embedding module is only used for capturing the prior and posterior distributions of the stochastic components in the generated frames.
Video Generation: The success of GANs has resulted in a myriad of image generation algorithms [11,15,16,30,31,36,61]. Inspired from these techniques, methods for video generation have also been proposed [46,52,55]. These algorithms usually directly map a noise vector sampled from a known or a learned distribution into a realistic-looking video and as such are known as *unconditional video generation* methods. Instead, our proposed generative model uses additional audio inputs, alongside encoding of the past frames. Models like ours are therefore, typically referred to as *conditional video generation* techniques. Prior works [17,34,19,42,59] have shown the success of conditional generative methods when information, such as the video categories, captions, etc., are available, using which constraints the plausible generations, improving their quality. Our proposed architecture differs in the modalities we use to constrain the generations and the associated technical innovations required to accommodate them.
Video Prediction/Forecasting: This is the task of predicting future frames, given a few frames from the past. Prior works in this area typically fall under: (i) *Deterministic*, and (ii) *Diversity-based* methods. Deterministic methods often use an encoder-decoder model to generate video frames autoregressively. The inherent stochasticity within the video data (due to multiple plausible futures or encoding noise) is thus difficult to be incorporated in such models [44,54,12,25,37,48,22]. Our approach circumvents these issues via a stochastic module. There have been prior efforts to capture this stochasticity from unimodal cues, such as [57,10,62,4], by learning a parametric prior distribution. Different from these approaches, we model the stochasticity using multimodal inputs.

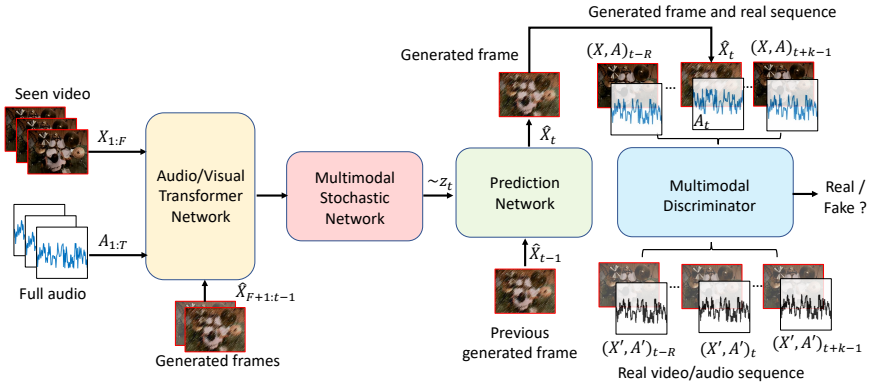


Fig. 2: Overview of the architecture of Sound2Sight. Our model takes F "seen" video frames (during inference) and all T audio samples, producing $T-F$ video frames (each denoted by $\hat{X}_t$). During training, the multimodal discriminator predicts if an input video is real or fake. We construct the fake video by replacing the t-th frame of the ground truth by $\hat{X}_t$. Note that during training, the generated frames ($\hat{X}_{F+1:t-1}$) which are input to the audio/visual transformer, are replaced by their real counterparts ($X_{F+1:t-1}$), while also using the current frame X_t to train the stochastic network.

We also note that there are several works in the area of generating human face animations conditioned on speech [24,26,47,50,51], however these techniques often make use of additional details, such as the identity of the person or leverage strong facial cues such as landmarks, textures, etc. - hindering their applicability to generic videos. There are methods free of such constraints, such as [38], however they synthesize images and not videos. A work similar to ours is Vougioukas *et al.* [56] that synthesizes face motions directly from speech and an initial frame, however it operates in the restricted domain of generating facial motions only.

3 Proposed Method

Given a dataset $\mathcal{D} = \{V_1, V_2, \cdots, V_N\}$ consisting of N video sequences, where each V is characterized by a pair $(X_{1:T}, A_{1:T})$ of T video frames and its time-aligned audio samples, i.e., $X_{1:T} = \langle X_1, X_2, ..., X_F, X_{F+1}, ..., X_T \rangle$ and $A_{1:T} = \langle A_1, A_2, ..., A_T \rangle$. We assume that the audio and the video are synchronized in such a way that A_t corresponds to the sound associated with the frame X_t in the duration $(t, t+1)$. Now, given as input a sequence of F frames $X_{1:F}$, $(F < T)$ and the audio $A_{1:T}$, our task is to generate frames $\hat{X}_{F+1:T}$ that is as realistic as possible compared to the true frames $X_{F+1:T}$. Given the under-constrained nature of the audio to video generation problem, we empirically show that it is essential to provide the past frames $X_{1:F}$ to set the visual context besides providing the audio input.

Sound2Sight Architecture: In this section, we first present an overview of the proposed model, before discussing the details. Figure 2 illustrates the key components in our model and the input-output data flow. In broad strokes, our model follows an encoder-decoder auto-regressive generator architecture, generating the video sequentially one frame at a time. This generator module has two components, viz. the *Prediction Network* and the *Multimodal Stochastic Network*. The former module takes the previous frame X_{t-1} as input,[3] encodes it into a latent space, concatenates it with a prior latent sample z_t obtained from the stochastic network, and decodes it to generate a frame $\hat{X}_t$, which approximates the target frame X_t. Sans the sample z_t, the prediction network is purely deterministic and unimodal, and hence can fail to capture the stochasticity in the motion dynamics. This challenge is mitigated by the multimodal stochastic network, which uses transformer encoders [53] on the audio and visual input streams to produce (the parameters of) a prior distribution from which z_t is sampled. The generator can thus be thought of as a non-linear heteroskedastic dynamical system (whose variance is decided by an underlying neural network), which generates $\hat{X}_t$ from the pair $(\hat{X}_{t-1}, z_t)$, and implicitly conditioned on the (latent) history of previous samples and the given audio.

During training, two additional data flows happen. (i) The transformer and the stochastic network take as input the true video sequence $X_{1:t}$ as well. This is used to estimate a posterior distribution which is in turn used to train the stochastic prior so that it effectively captures the distribution of real video samples. (ii) Further, the generated frames are evaluated for their realism, motion synchrony, and audio-visual alignment using a multimodal adversarial discriminator [15] (Figure 2). This discriminator uses $\hat{X}_t$ – the synthetic frame, inserted at the t-th index of the original sequence, and $X_{t-R:t+(k-1)}$ the set of R past, and $(k-1)$ future frames, along with the corresponding audio, and compares it with real (arbitrary) audio-visual clips of length $R+k$ from the dataset. Since discriminators match distributions, rather than matching individual samples, this ensures that incorporating the generated frame $\hat{X}_t$ results in a coherent video that is consistent with the input audio, while permitting diversity. We now elaborate on each of the above modules and layout our training strategy.

Prediction Network: Broadly speaking, the prediction network (PN) is a standard sequence-to-sequence encoder-decoder network. It starts off by embedding the previous frame X_{t-1} into a latent space. We denote this embedding by $f(X_{t-1})$, where $f(\cdot)$ abstracts a convolutional neural network (CNN) [33]. Each layer of this CNN consists of a set of convolution kernels, followed by 2D-Batch Normalization [23] layers, Leaky ReLU activations, and has skip-connections to the decoder part of the network. These skip connections facilitate reconstruction of static parts of the video [45]. The embedding of the frame $f(X_{t-1})$ is then concatenated with a sample $z_t \sim \mathcal{N}(\mu_\phi, \Sigma_\phi)$, a Gaussian prior provided by the stochastic module (described next) where μ_ϕ and Σ_ϕ denote the mean and a diagonal covariance matrix of this Gaussian prior. Our key idea is to

[3] X_{t-1} is the *real* frame during training, however during inference, it is the generated frame $\hat{X}_{t-1}$ if $t-1 > F$.

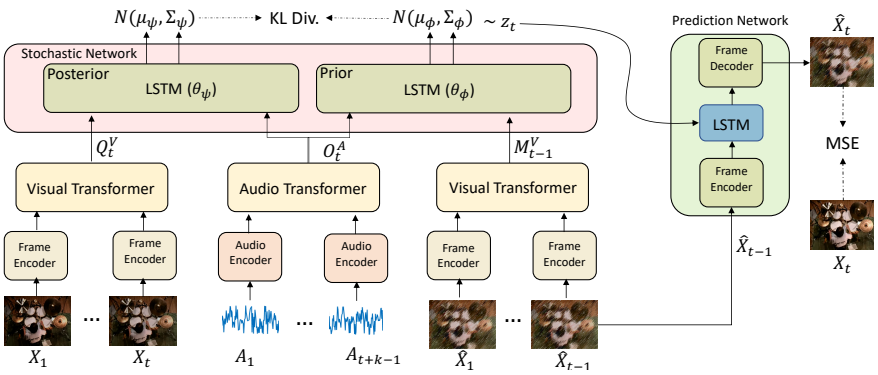


Fig. 3: Details of our Multimodal Stochastic Network and Prediction Network.

have z_t capture the cues about the future as provided by the available audio, as well as the randomness in producing the next frame. We then feed the pair $(f(X_{t-1}), z_t)$ to a Long-Short Term Memory (LSTM) [21], parametrized by θ_L within the PN; this LSTM keeps track of the previously generated frames via its internal states. Specifically, if h_{t-1} denotes the hidden state of this LSTM, then we define its output η_t as: $\eta_t = \text{LSTM}_{\theta_L}((f(X_{t-1}), z_t), h_{t-1})$. The LSTM output η_t is then passed to the decoder network $g(\cdot)$, to generate the next frame, i.e., $\hat{X}_t = g(\eta_t)$. The decoder consists of a set of deconvolution layers with Leaky ReLU activations, coupled with 2D-Batch Normalization layers.

Multimodal Stochastic Network: Several prior works have underscored the importance of modeling the stochasticity in video generation [4,10,57,62], albeit using a single modality. Inspired by these works, we introduce the multimodal stochastic network (MSN) that takes both the audio and video streams as inputs to model the stochastic elements in producing the target frame X_t. As alluded to earlier, such a stochastic element allows for capturing the randomness in the generated frame, while also permitting the sampling of multiple plausible futures conditioned on the available inputs. As shown in Figure 3, the stochastic network is effectuated by computing a prior and a posterior distribution in the embedding space (from which z_t is sampled) and training the model to minimize their mutual discrepancy. The prior distribution is jointly conditioned on an embedding of the audio sub-clip $A_{1:t+(k-1)}$ and an embedding of the video frames $X_{1:t-1}$, both obtained via transformer encoders. We denote the t-th audio encoding by O_t^A, while the $(t-1)$-th video encoding is denoted by M_{t-1}^V. Let the prior distribution be $p_\phi(z_t|O_t^A, M_{t-1}^V)$, parametrized as a Gaussian, with mean μ_ϕ and diagonal covariance Σ_ϕ. Likewise, the posterior distribution $p_\psi(z_t|O_t^A, Q_t^V)$, which is also assumed to be a Gaussian $\mathcal{N}(\mu_\psi, \Sigma_\psi)$, is jointly conditioned on audio clips $A_{1:t+(k-1)}$ and visual frames $X_{1:t}$. Its audio embedding is shared with the prior distribution and its visual input is obtained from the t-th transformer encoding is denoted Q_t^V. Here, it is worth noting that the visual conditioning of the prior distribution, unlike the posterior, is *only upto frame* $t-1$, i.e. the

past visual frames. Since the posterior network has access to the t-th frame in its input, it may attempt to directly encode this frame to be decoded by the prediction network decoder to produce the next frame. However, due to the KL-divergence loss between the prior and the posterior distributions, such a direct decoding cannot happen; unless the prior is trained well such that the KL-loss is minimized; which essentially implies the prior $p_\phi(z_t|O_t^A, M_{t-1}^V)$ will be able to predict the latent distribution of the future samples (as if from the posterior $p_\psi(z_t|O_t^A, Q_t^V)$), which is essentially what we require during inference.

To generate the prior distribution, we concatenate the embedded features M_{t-1}^V and O_t^A as input to an LSTM_ϕ. Different from standard LSTMs, this LSTM predicts the parameters of the prior distribution directly, i.e., $\mu_\phi, \log \Sigma_\phi = \text{LSTM}_\phi(O_t^A, M_{t-1}^V)$. The posterior distribution parameters are estimated similarly, using a second LSTM, denoted LSTM_ψ that takes as input the embedded and concatenated audio-video features O_t^A and Q_t^V to produce: $\mu_\psi, \log \Sigma_\psi = \text{LSTM}_\psi(O_t^A, Q_t^V)$.

Audio-Visual Transformer Encoder: Next, we describe the process of producing the prior and posterior distributions from audio-visual joint embeddings. As we want these embeddings to be "temporally-conscious" while computable efficiently, we bank on the very successful *Transformer Encoder Networks* [53], which are armed with self-attention modules that are well-known to produce powerful multimodal representations. Re-using the encoder CNN f from the prediction network, our visual transformer encoder takes as input the matrix $\mathcal{F} = \langle f'(X_1), f'(X_2), \cdots, f'(X_{t-1}) \rangle$ with $f'(X_i)$ in its i-th column, where $f'(X_i)$ denotes the feature encoding $f(X_i)$ augmented with the respective temporal position encoding of the frame in the sequence, as suggested in [53]. We then apply ℓ-head self-attention to $\mathcal{F}$ by designing *Query* ($\mathcal{Q}$), *Key* ($\mathcal{K}$), and *Value* ($\mathcal{V}$) triplets via linear projections of our frame embeddings $\mathcal{F}$; i.e., $\mathcal{Q}_j = W_q^j \mathcal{F}$, $\mathcal{K}_j = W_k^j \mathcal{F}$, and $\mathcal{V}_j = W_v^j \mathcal{F}$, where W_q^j, W_k^j, W_v^j are matrices of sizes $d_k \times d$, d is the size of the feature f', and $j = 1, 2, \cdots, \ell$. Using ℓ, $d_k \times d$ weight matrix W_h, our self-attended feature $\hat{M}_{t-1}^V$ from this transformer layer is thus:

$$\hat{M}_{t-1}^V = \text{concat}_{j=1}^{\ell} \left(\text{softmax} \left(\frac{\mathcal{Q}_j \mathcal{K}_j^\top}{\sqrt{d_k}} \right) \mathcal{V}_j \right) W_h, \tag{1}$$

where concat denotes the concatenation operator. We use four consecutive self-attention layers within every transformer encoder, which are then combined via feed-forward layers to obtain the final encoding [53] M_{t-1}^V, which is subsequently used in the MSN module. Likewise, the re-purposed visual features for the posterior distribution, Q_t^V, can also be computed by employing a separate transformer encoder module, which ensures a separation of the visual components of the prior and the posterior networks. To produce the audio embeddings O_t^A, we first compute STFT (Short-Time Fourier Transform) features $(S_1, S_2, ..., S_{t+(k-1)})$ from the raw audio by choosing appropriate STFT filter sizes and strides, where each $S_i \in \mathbb{R}^{d_{H_A} \times d_{W_A}}$ and encode them using an audio transformer.

Generator Loss: To train our generator model, we directly maximize the variational *Empirical Lower BOund* (ELBO) [30] by optimizing the objective:

$$\mathcal{L}_V = \sum_{t=F+1}^{T} \mathop{\mathbb{E}}_{z_t \sim p_\phi} \log p_\phi(\hat{X}_t | M_{t-1}^V, z_t) - \beta \, \mathrm{KL}\left(p_\psi\left(z_t | O_t^A, Q_t^V\right) \parallel p_\phi\left(z_t | O_t^A, M_{t-1}^V\right)\right),$$

where the KL-divergence matches the closeness of the posterior distribution and the prior, while β is a constant. Casting the above as a minimization and approximating the first term by the pixel-wise ℓ_2 error, reduces the objective to:

$$\mathcal{L}_V \approx \sum_{t=F+1}^{T} \|X_t - \hat{X}_t\|_2^2 \; + \; \beta KL(p_\psi || p_\phi). \tag{2}$$

Multimodal Discriminator Network: Computing the training loss, as in (2), is entirely based on the supplied ground truth (which is only one of many possibilities) and thus might restrict generative diversity. We rectify this shortcoming using a multimodal discriminator (see Fig. 2), which is designed to match the distribution of synthesized frames p_G against the ground truth distribution $p_{\mathcal{D}}$. In contrast to conventional image-based GAN discriminators [5,15], our variant couples a classifier, denoted D_{std}, and an LSTM D, to produce binary labels indicating if the t-th frame is drawn from $p_{\mathcal{D}}$ or from p_G. This is done via using a set of ground-truth audio-visual frames from the neighborhood of the generated frame, where this neighborhood spans the previous R and future $(k-1)$ frames. When judging its inputs, the discriminator, besides looking into whether the t-th frame appears real or fake, also looks at how well the regularities of object motions are preserved with respect to the neighborhood via a motion dynamics (MD) loss, and if the frames are synchronized with the audio via an audio alignment (AA) loss. With these additional terms, our discriminator loss is:

$$\begin{aligned}
\mathcal{L}_{adv} = &- \sum_{t=F+1}^{T} \mathop{\mathbb{E}}_{X'_t \sim p_{\mathcal{D}}} \log D_{std}(X'_t) + \mathop{\mathbb{E}}_{\hat{X}_t \sim p_G} \log(1 - D_{std}(\hat{X}_t)) \\
&+ \mathop{\mathbb{E}}_{X'_t \sim p_{\mathcal{D}}} \log \underbrace{D(X'_t | A'_t, B'_{t+(k-1)}, \cdots, B'_{t+1}, B'_{t-1}, \cdots, B'_{t-R})}_{\text{Real Data - Motion Dynamics (MD)}} \\
&+ \mathop{\mathbb{E}}_{X'_t \sim p_{\mathcal{D}}} \log \underbrace{(1 - D(X'_t | A'_{t'}, C'_{t+(k-1),t'+(k-1)}, \cdots, C'_{t+1,t'+1}, C'_{t-1,t'-1}, \cdots, C'_{t-R,t'-R}))}_{\text{Real Data - Audio Alignment (AA)}} \\
&+ \mathop{\mathbb{E}}_{\hat{X}_t \sim p_G} \log \underbrace{(1 - D(\hat{X}_t | A_t, B_{t+(k-1)}, \cdots, B_{t+1}, B_{t-1}, ..., B_{t-R}))}_{\text{Synthetic Frame - Motion Dynamics (MD)}} \\
&+ \mathop{\mathbb{E}}_{\hat{X} \sim p_G} \log \underbrace{(1 - D(\hat{X}_t | A_{t'}, C_{t+(k-1),t'+(k-1)}, \cdots, C_{t+1,t'+1}, C_{t-1,t'-1}, \cdots, C_{t-R,t'-R}))}_{\text{Synthetic Frame - Audio Alignment (AA)}}
\end{aligned} \tag{3}$$

where (X'_t, A'_t) denotes a visual frame X'_t and its associated audio A'_t from a clip $B' = (X'_{1:T}, A'_{1:T})$ arbitrarily sampled from the training set. Similarly, we

define $C_{t,t'} = (X_t, A_{t'})$, $t' \neq t$, $B_t = (X_t, A_t)$, $C'_{t,t'} = (X'_t, A'_{t'})$, $B'_t = (X'_t, A'_t)$, $X_t \neq X'_t$, $A_t \neq A'_t$. The first term in (3) defines a standard image-based GAN loss, while D in the other terms denotes a convolutional LSTM. The motion dynamics term captures the consistency of the generated frame against other frames in the sequence (i.e., X'_t against B' on the real, and $\hat{X}_t$ against B on the generated), while the audio alignment of the generated frame $\hat{X}_t$ against arbitrary audio samples A' is captured by the AA term. We optimize for the discriminator parameters by minimizing this loss above.

Combining the adversarial losses above with (2), our final objective for optimizing the generator is: $\mathcal{L} = \mathcal{L}_V - \gamma \mathcal{L}_{adv}$, where γ is a constant. We minimize this loss using ADAM [29], while employing the reparameterization trick [30] to ensure differentiability of the stochastic sampler.

4 Experiments

To benchmark the performance of our model, we present empirical experiments on a synthetic and two real world datasets, which will be made publicly available.
Multimodal MovingMNIST with a Surprise Obstacle (M3SO): is a novel extension of the stochastic MovingMNIST dataset [10] adapted to our multimodal setting, and consists of MNIST digits moving along rectilinear paths in a fixed size box (48×48) which bounce in random directions upon colliding with the box boundaries. In addition: (i) we equip each digit with a unique tone, (ii) the amplitude of this tone is inversely proportional to the digit's distance from the origin, and (iii) the tone changes momentarily when the digit bounces off the box edge. We make this task even more challenging by introducing an obstacle (square block of fixed size) at a random location within the unseen part of the video. When the digit bounces against the block, a unique audio frequency is emitted. The task on this dataset is not only to generate the frames, but also to predict the location of the block by listening to the tone changes. See supplementary materials for details. We also construct a version of the dataset, where no block is introduced, called M3SO-NB. We produced 8,000 training, 1,000 validation, and 1,000 test samples for both M3SO and M3SO-NB.
AudioSet-Drums: includes videos from the *Drums* class of AudioSet [14]. We clipped and retained only those video segments from this dataset for which the drum player is visible when the drum beat is heard. This yielded a dataset consisting of 8K clips which we split as 6K for training, 1K for validation, and 1K for test. Each video is of 64×64 resolution, 30fps, and is 3 seconds long.
YouTube Painting: To analyze Sound2Sight in a subtle, yet real world setting, we introduce the *Youtube Painting* dataset. The videos in this dataset are manually collected via crawling painting videos on Youtube [2]. We selected only those videos that contain a painter painting on a canvas in an indoor environment, and which have a clear audio of the brush strokes. These videos provide a wide assortment of brush strokes and painting colors. The painter's motions and the camera viewpoints are often arbitrary which adds to the complexity and diversity, making it a very challenging dataset. Here the task is to generate frames

showing the dynamics of the painter's arms, while preserving the static components in the scene. We collected 4.8K videos for training, 500 for validation and 500 for test. Each video is of 64×64 resolution, 30fps, and 3s long.

Evaluation Setup: On the M3SO dataset, we conduct experiments in two settings: (i) in M3SO-NB, all methods are shown 5 frames and the full audio, with the task being to predict the next 15 frames at training and 20 frames at test time, and (ii) using M3SO in which blocks are presented, we show 30 frames at training and 30 frames are predicted, however the block appears at the 42-nd frame. We predict 40 frames at test time. For the real-world datasets, we train all algorithms on 15 seen frames and predict the next 15, while has to predict 30 at test time. We use the standard structural similarity (SSIM) [60] and the Peak Signal to Noise Ratio (PSNR) scores for quantitative evaluation of the quality of the generated frames against the ground-truth.

Baselines: As our task is novel, we compare our algorithm against the following closely-related baselines: (i) *Audio-Only*: using a sequence-to-sequence model [49] taking only the audio as input and generate the frames using an LSTM (thus, the past context is missing), (ii) *Video-Only*, using three baselines: (ii-a) Denton and Fergus [10], (ii-b) Hsieh et al. [22], and (ii-c) an ablated variant of our model without audio (Ours - No audio), and (iii) *Multimodal*: with further three baselines: (iii-a) Vougioukas et al. [56], that predicts the video from audio and the first frame, (iii-b) [56] modified to use a set of seen frames (Multiframe [56]), (iii-c) ablated variants of our model without the AA loss term in the discriminator (Ours - No AA) and without the AA and MD loss terms (Ours - No AA, MD).

Implementation Details: The PN module uses an LSTM with two layers and produces 128-D frame embeddings. We use 10-D stochastic samples (z_t). The prior and posterior LSTMs are both single-layered, each with 256-D inputs from audio-frame embeddings (which are each 128-D). All LSTMs have 256-D hidden states. Each transformer module has one layer and four heads with 128-D feedforward layer. The discriminator uses an LSTM with a hidden layer of 256-D, a frame-history $R = 2$, and look-head $k = 1$. We train the generator and discriminator jointly with a learning rate of 2e-3 using ADAM [29]. We set both β and γ as 0.0001, and increased γ by a factor of 10 every 300 epochs. All hyper-parameters are chosen using the validation set. During inference, we sample 100 futures per time step, and use sequences that best matches the ground-truth, for our method and the baselines.

4.1 Experimental Results

M3SO Results: Table 1 shows the performance of our model versus competing baselines on the M3SO dataset in two settings: (i) without block (M3SO-NB) and (ii) with block (M3SO). For M3SO-NB, we observe that our method attains significant improvements over prior works, even on long-range generation. In M3SO, when the block is introduced at the 42-nd frame, the generated frame quality drops across all methods. Nevertheless, our method continues to demonstrate better performance. Figure 4(b) presents a visualization of the generated frames by our method vis-á-vis prior works on the M3SO dataset. Contrasting

Table 1: SSIM, PSNR for M3SO-NB and M3SO. **Highest**, Second highest scores. Notation: Multimodal (M), Unimodal-Video (V), Unimodal-Audio (A)

Experiments with M3SO-NB with 5 seen frames							
Method	**Type**	**SSIM**			**PSNR**		
		Fr 6	Fr 15	Fr 25	Fr 6	Fr 15	Fr 25
Our Method	M	**0.9575**	**0.8943**	**0.8697**	21.69	**17.62**	16.84
Ours - No AA	M	0.9547	0.8584	0.8296	21.80	17.36	**16.97**
Ours - No AA, MD	M	0.9477	0.8546	0.8251	21.16	16.16	15.49
Ours - No audio	V	0.9556	0.8351	0.6920	**22.66**	15.59	12.40
Multiple Frames - [56]	M	0.9012	0.8690	0.8693	18.09	15.23	15.33
Vougioukas *et al.* [56]	M	0.8600	0.8571	0.8573	15.17	14.99	15.01
Denton and Fergus [10]	V	0.9265	0.8300	0.7999	18.59	14.65	13.98
Audio Only	A	0.8499	0.8659	0.8662	13.71	13.16	12.94
Experiments on M3SO with 30 seen frames (Block appears: 42^{nd} frame)							
		Fr 31	Fr 42	Fr 70	Fr 31	Fr 42	Fr 70
Our Method	M	**0.8780**	**0.6256**	**0.6170**	**19.50**	**9.39**	**9.41**
Multiple Frames - [56]	M	0.8701	0.6073	0.6050	15.41	8.53	8.53
Vougioukas *et al.* [56]	M	0.8681	0.6009	0.6007	15.17	8.48	8.48
Denton and Fergus [10]	V	0.7353	0.5115	0.4991	12.25	7.13	7.00
Audio Only	A	0.6474	0.5397	0.5315	12.39	9.25	8.84

the output by our method against prior works clearly reveals the superior generation quality of our method, which closely resembles the ground truth. We find that the method of [10] fares well under uncertainty, however our task demands reasoning over audio - an element missing in their setup. Further note that our model localizes the block in time (i.e. after the 42-th frame) better than other methods. This is quantitatively analyzed in Table 3 by comparing the mean IoU of the predicted block location in the final generated frame against the ground truth. Our scheme outperforms the closest baseline [10] by ∼30%.

Comparisons on Real-world Datasets: As with M3SO, we see from Table 2 that our approach outperforms the baselines, even at long-range generation. Due to the similarity in visual content (e.g., background) of the unseen frames to the seen frames, prior methods (e.g., [56] and [10]) are seen to copy the seen frames as *predicted* ones, yielding relatively high SSIM/PSNR early on (Figures 1 and 4(a) that show that drummer's and painter's arms remain fixed); however their performances drop in the long-range. Instead, our method captures the hand motions. Further, our generations are free from artifacts, as corroborated by the fooling rate on the fully-trained discriminator, that achieves 79.26% for AudioSet Drums and 65.99% for YouTube Painting.

Human Preference Scores: To subjectively assess the video generation quality, we conducted a human preference evaluation between a randomly selected subset of our generated videos and those produced by the closest competitor-Vougioukas *et al.* [56] on both the real-world datasets. The results in Table 4 evince that humans preferred our method for more than 80-90% of the videos against those from [56].

Table 2: SSIM, PSNR for AudioSet, YouTube Painting. **Highest**, Second highest scores. Notation: Multimodal (M), Unimodal-Video (V), Unimodal-Audio (A)

Experiments on the AudioSet Dataset [14], with 15 seen frames							
Method	**Type**	**SSIM**			**PSNR**		
		Fr 16	Fr 30	Fr 45	Fr 16	Fr 30	Fr 45
Our Method	M	**0.9843**	**0.9544**	**0.9466**	**33.24**	**27.94**	**26.99**
Multiple Frames - [56]	M	0.9398	0.9037	0.8959	26.21	23.78	23.29
Vougioukas *et al.* [56]	M	0.8986	0.8905	0.8866	23.62	23.14	22.91
Denton and Fergus [10]	V	0.9706	0.6606	0.5097	30.01	16.57	13.49
Hsieh *et al.* [22]	V	0.1547	0.1476	0.1475	9.42	9.54	9.53
Audio Only	A	0.6485	0.6954	0.7277	18.81	19.79	20.50
Experiments on the YouTube Painting Dataset, with 15 seen frames							
Our Method	M	0.9716	**0.9291**	**0.9110**	**32.73**	**27.27**	**25.57**
Multiple Frames - [56]	M	0.9657	0.9147	0.8954	30.09	25.40	24.08
Vougioukas *et al.* [56]	M	0.9281	0.9126	0.9027	26.97	25.58	24.78
Denton and Fergus [10]	V	**0.9779**	0.6654	0.4193	32.52	16.05	11.84
Hsieh *et al.* [22]	V	0.1670	0.1613	0.1618	9.11	9.57	9.72
Audio Only	A	0.5997	0.6462	0.6743	16.75	17.53	18.04

Table 3: Block IoU on M3SO.

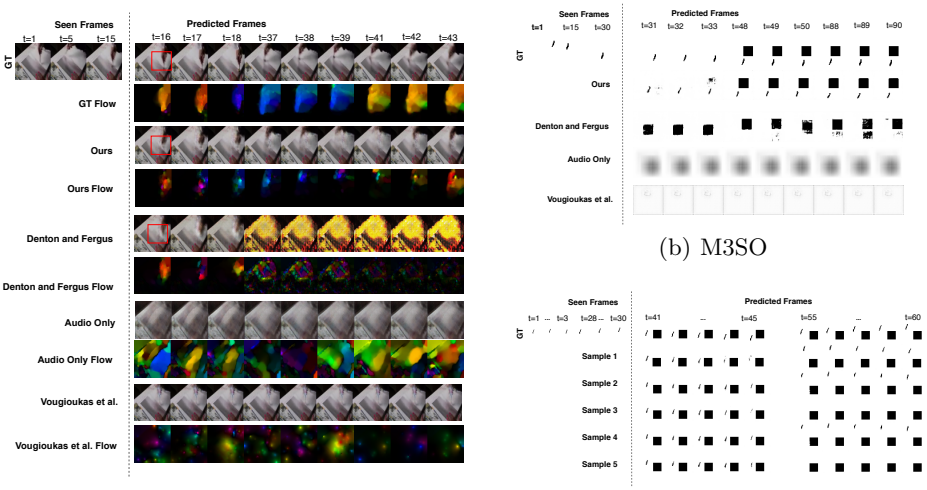
Method	**Localization IoU**
Ours	**0.5801**
[10]	0.2577
[56]	0.1289

Table 4: Human preference score on samples from our method vs. [56]

Datasets	**Prefer ours**
AudioSet	**83%**
YouTube Painting	**92%**

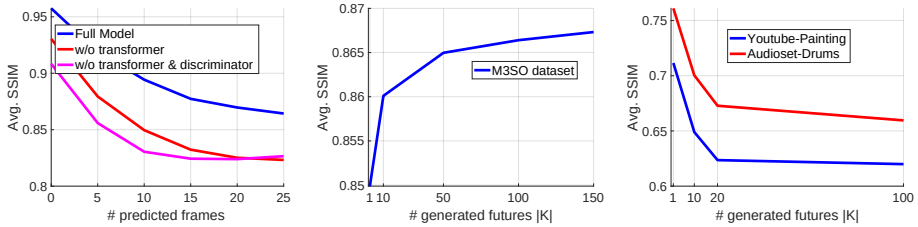
Sample Diversity: In Figure 4(c), we show the diversity in the samples generated on the M3SO dataset. Figure 5(b) shows quantitative evaluations of diversity. Specifically, we generated a set of $\mathcal{K}$ futures at every time step (for $|\mathcal{K}|$ ranging from $1-100$), and plotted the SSIM of the samples which matched maximally with the ground truth. As is clear, this plot shows an increasing trend suggesting that samples closer to the ground-truth are obtainable by increasing $\mathcal{K}$; i.e., generative diversity. We further analyze this over SSIMs on optical flows computed from the Youtube Painting and Drums datasets. In Figure 5(c), we plot the intra-sample diversity, i.e., the average pairwise SSIMs for sequences in $\mathcal{K}$; showing a downward trend, suggesting these sequences are self-dissimilar.

Ablation Results: To study the influence of the transformer network, we contrast our model by substituting the transformer by an LSTM with 128-D hidden states. Figure 5(a) shows the result, clearly suggesting the benefits of using transformers. From this plot, we also find that having our discriminator is important. Tables 1 and 2 show that removing the AA and MD loss terms from the discriminator hurts performance.



(a) Youtube Painting (b) M3SO (c) M3SO-diversity

Fig. 4: (a,b) show qualitative comparisons of generated frames and optical flow images, (c) shows generative diversity on M3SO.



(a) Ablation Study (b) Inter-SSIM (c) Intra-SSIM

Fig. 5: Ablation and diversity studies (see text for details).

5 Conclusions

In this work, we explored the novel task of video generation from audio and the visual context for generic videos. We proposed a novel deep variational encoder-decoder model for this task, that also characterizes the underlying stochasticity in real-world videos. We combined our video generator with a multimodal discriminator to improve its quality and diversity. Empirical evaluations on three datasets demonstrated the superiority of our method over competing baselines.
Acknowledgements: MC thanks the support from the Joan and Lalit Bahl Fellowship and inputs from Prof. Narendra Ahuja and the annotators.

References

1. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 609–617 (2017)
2. ASMR, T.: Painting ASMR (2019 (accessed November 5, 2019)), `https://www.youtube.com/playlist?list=PL5Y0dQ2DJHj47sK5jsbVkVpTQ9r7T090X`
3. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: Proceedings of Advances in neural information processing systems. pp. 892–900 (2016)
4. Babaeizadeh, M., Finn, C., Erhan, D., Campbell, R.H., Levine, S.: Stochastic variational video prediction. arXiv preprint arXiv:1710.11252 (2017)
5. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
6. Cardoso Duarte, A., Roldan, F., Tubau, M., Escur, J., Pascual de la Puente, S., Salvador Aguilera, A., Mohedano, E., McGuinness, K., Torres Viñals, J., Giró Nieto, X.: Wav2pix: speech-conditioned face generation using generative adversarial networks. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing: proceedings: May 12-17, 2019: Brighton Conference Centre, Brighton, United Kingdom. pp. 8633–8637. Institute of Electrical and Electronics Engineers (IEEE) (2019)
7. Chen, L., Maddox, R.K., Duan, Z., Xu, C.: Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7832–7841 (2019)
8. Chen, L., Srivastava, S., Duan, Z., Xu, C.: Deep cross-modal audio-visual generation. In: Proceedings of the on Thematic Workshops of ACM Multimedia 2017. ACM (2017)
9. Corlett, P.R., Powers, A.R.: Conditioned hallucinations: historic insights and future directions. World Psychiatry **17**(3), 361 (2018)
10. Denton, E., Fergus, R.: Stochastic video generation with a learned prior. In: Proceedings of International Conference on Machine Learning. pp. 1182–1191 (2018)
11. Deshpande, I., Zhang, Z., Schwing, A.G.: Generative modeling using the sliced wasserstein distance. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3483–3491 (2018)
12. Finn, C., Goodfellow, I., Levine, S.: Unsupervised learning for physical interaction through video prediction. In: Proceedings of Advances in neural information processing systems. pp. 64–72 (2016)
13. Fragkiadaki, K., Agrawal, P., Levine, S., Malik, J.: Learning visual predictive models of physics for playing billiards. arXiv preprint arXiv:1511.07404 (2015)
14. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 776–780. IEEE (2017)
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of Advances in neural information processing systems. pp. 2672–2680 (2014)
16. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Proceedings of Advances in neural information processing systems. pp. 5767–5777 (2017)
17. Gupta, T., Schwenk, D., Farhadi, A., Hoiem, D., Kembhavi, A.: Imagine this! scripts to compositions to videos. In: Proceedings of the European Conference on Computer Vision. pp. 598–613 (2018)

18. Hao, W., Zhang, Z., Guan, H.: Cmcgan: A uniform framework for cross-modal visual-audio mutual generation. In: Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence (2018)
19. Hao, Z., Huang, X., Belongie, S.: Controllable video generation with sparse trajectories. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7854–7863 (2018)
20. Harwath, D., Torralba, A., Glass, J.: Unsupervised learning of spoken language with visual context. In: Proceedings of Advances in Neural Information Processing Systems. pp. 1858–1866 (2016)
21. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
22. Hsieh, J.T., Liu, B., Huang, D.A., Fei-Fei, L.F., Niebles, J.C.: Learning to decompose and disentangle representations for video prediction. In: Proceedings of Advances in Neural Information Processing Systems. pp. 517–526 (2018)
23. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of International Conference on Machine Learning. pp. 448–456 (2015)
24. Jamaludin, A., Chung, J.S., Zisserman, A.: You said that?: Synthesising talking faces from audio. International Journal of Computer Vision pp. 1–13 (2019)
25. Jia, X., De Brabandere, B., Tuytelaars, T., Gool, L.V.: Dynamic filter networks. In: Proceedings of Advances in Neural Information Processing Systems. pp. 667–675 (2016)
26. Karras, T., Aila, T., Laine, S., Herva, A., Lehtinen, J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics **36**(4), 94 (2017)
27. Kidron, E., Schechner, Y.Y., Elad, M.: Pixels that sound. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 1, pp. 88–95. IEEE (2005)
28. Kim, D., Woo, S., Lee, J.Y., Kweon, I.S.: Deep video inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5792–5801 (2019)
29. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
30. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
31. Kolouri, S., Pope, P.E., Martin, C.E., Rohde, G.K.: Sliced-wasserstein autoencoder: an embarrassingly simple generative model. arXiv preprint arXiv:1804.01947 (2018)
32. Lamb, A., Dumoulin, V., Courville, A.: Discriminative regularization for generative models. arXiv preprint arXiv:1602.03220 (2016)
33. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
34. Li, Y., Min, M.R., Shen, D., Carlson, D., Carin, L.: Video generation from text. In: Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence (2018)
35. Lindell, D.B., Wetzstein, G., Koltun, V.: Acoustic non-line-of-sight imaging. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6780–6789 (2019)
36. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Proceedings of Advances in neural information processing systems. pp. 700–708 (2017)

37. Luo, Z., Peng, B., Huang, D.A., Alahi, A., Fei-Fei, L.: Unsupervised learning of long-term motion dynamics for videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2203–2212 (2017)
38. Oh, T.H., Dekel, T., Kim, C., Mosseri, I., Freeman, W.T., Rubinstein, M., Matusik, W.: Speech2face: Learning the face behind a voice. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7539–7548 (2019)
39. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: Proceedings of the European Conference on Computer Vision. pp. 631–648 (2018)
40. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2405–2413 (2016)
41. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Ambient sound provides supervision for visual learning. In: Proceedings of European conference on computer vision. pp. 801–816. Springer (2016)
42. Pan, J., Wang, C., Jia, X., Shao, J., Sheng, L., Yan, J., Wang, X.: Video generation from single semantic label map. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3733–3742 (2019)
43. Pavlov, I.P.: The work of the digestive glands. Charles Griffin, Limited; Exeter Street, Strand (1910)
44. Ranzato, M., Szlam, A., Bruna, J., Mathieu, M., Collobert, R., Chopra, S.: Video (language) modeling: a baseline for generative models of natural videos. arXiv preprint arXiv:1412.6604 (2014)
45. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
46. Saito, M., Matsumoto, E., Saito, S.: Temporal generative adversarial nets with singular value clipping. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2830–2839 (2017)
47. Shlizerman, E., Dery, L., Schoen, H., Kemelmacher-Shlizerman, I.: Audio to body dynamics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7574–7583 (2018)
48. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using lstms. In: Proceedings of International conference on machine learning. pp. 843–852 (2015)
49. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Proceedings of Advances in neural information processing systems. pp. 3104–3112 (2014)
50. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics **36**(4), 95 (2017)
51. Taylor, S., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A.G., Hodgins, J., Matthews, I.: A deep learning approach for generalized speech animation. ACM Transactions on Graphics **36**(4), 93 (2017)
52. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1526–1535 (2018)
53. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proceedings of Advances in neural information processing systems. pp. 5998–6008 (2017)
54. Villegas, R., Yang, J., Hong, S., Lin, X., Lee, H.: Decomposing motion and content for natural video sequence prediction. arXiv preprint arXiv:1706.08033 (2017)

55. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: Proceedings of Advances in neural information processing systems. pp. 613–621 (2016)
56. Vougioukas, K., Petridis, S., Pantic, M.: End-to-end speech-driven facial animation with temporal gans. arXiv preprint arXiv:1805.09313 (2018)
57. Walker, J., Marino, K., Gupta, A., Hebert, M.: The pose knows: Video forecasting by generating pose futures. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3332–3341 (2017)
58. Wan, C.H., Chuang, S.P., Lee, H.Y.: Towards audio to scene image synthesis using generative adversarial network. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 496–500. IEEE (2019)
59. Wang, T.C., Liu, M.Y., Zhu, J.Y., Liu, G., Tao, A., Kautz, J., Catanzaro, B.: Video-to-video synthesis. arXiv preprint arXiv:1808.06601 (2018)
60. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., et al.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
61. Wu, J., Huang, Z., Acharya, D., Li, W., Thoma, J., Paudel, D.P., Gool, L.V.: Sliced wasserstein generative models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3713–3722 (2019)
62. Xue, T., Wu, J., Bouman, K., Freeman, B.: Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In: Proceedings of Advances in neural information processing systems. pp. 91–99 (2016)
63. Zhao, H., Gan, C., Ma, W., Torralba, A.: The sound of motions. CoRR **abs/1904.05979** (2019)
64. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: Proceedings of the European Conference on Computer Vision. pp. 570–586 (2018)
65. Zhao, M., Li, T., Abu Alsheikh, M., Tian, Y., Zhao, H., Torralba, A., Katabi, D.: Through-wall human pose estimation using radio signals. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7356–7365 (2018)
66. Zhou, H., Liu, Z., Xu, X., Luo, P., Wang, X.: Vision-infused deep audio inpainting. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 283–292 (2019)