Vision-Infused Deep Audio Inpainting

Hang Zhou[1] Ziwei Liu[1] Xudong Xu[1] Ping Luo[2] Xiaogang Wang[1]
[1]CUHK - SenseTime Joint Lab, The Chinese University of Hong Kong
[2]The University of Hong Kong
{zhouhang@link,xx018@ie,xgwang@ee}.cuhk.edu.hk zwliu.hust@gmail.com pluo@cs.hku.hk

Abstract

Multi-modality perception is essential to develop interactive intelligence. In this work, we consider a new task of visual information-infused audio inpainting, i.e. *synthesizing missing audio segments that correspond to their accompanying videos. We identify two key aspects for a successful inpainter: (1) It is desirable to operate on spectrograms instead of raw audios. Recent advances in deep semantic image inpainting could be leveraged to go beyond the limitations of traditional audio inpainting. (2) To synthesize visually indicated audio, a visual-audio joint feature space needs to be learned with synchronization of audio and video. To facilitate a large-scale study, we collect a new multi-modality instrument-playing dataset called MUSIC-Extra-Solo (MUSICES) by enriching MUSIC dataset [52]. Extensive experiments demonstrate that our framework is capable of inpainting realistic and varying audio segments with or without visual contexts. More importantly, our synthesized audio segments are coherent with their video counterparts, showing the effectiveness of our proposed Vision-Infused Audio Inpainter (VIAI). Code, models, dataset and video results are available at https://github.com/Hangz-nju-cuhk/Vision-Infused-Audio-Inpainter-VIAI.*

1. Introduction

Audio-visual analysis provides valuable and complementary information that is crucial for comprehensively modeling sequential data. Substantial progress has been achieved in recent years. For example, it has been shown that the two modalities of audio and video can be transformed from one to the other [10, 23], that is, from video to audio [11, 9] and from audio to video [18, 53, 54].

This work focuses on a new task of audio inpainting, by using both video and audio as constraints. The inpainted audio segment is required to have the semantic concepts of the constraints, meaning that it has to be not only auditory reasonable but also visually coherent with the video. The setting of the problem is illustrated in Fig. 1.

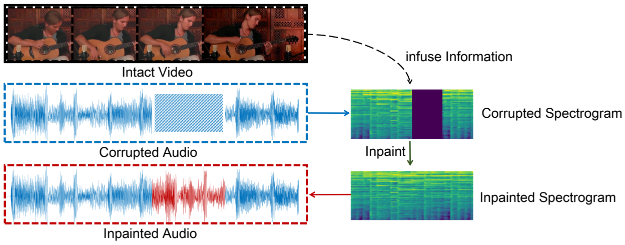


Figure 1. Problem description. We study the problem of inpainting a clip of missing audio data, particularly with its corresponding video given. It is formulated into deep spectrogram inpainting, and video information is infused for generating coherent audio.

In real life, audio signals often suffer from local distortions where the intervals are corrupted by impulsive noise and clicks. Even more, a clip of audio might be wiped out due to accident or transmission failure loss. To deal with such cases, a feasible operation is to fill the corrupted parts with newly generated samples, which can be referred to as audio inpainting [1].

While directly predicting a missing piece of audio is difficult, concrete information about audio signals could be provided by intact visual information accompanying the audio data. The visual cue can be regarded as both a constraint and self-supervision to guide audio generation. In this paper, we present a vision-infused method that can deal with both audio-only and audio-visual associated inpainting.

Audio inpainting is significantly challenging as a consequence of audio's property of high sampling rate and long-range dependency. Traditional methods normally exploit the sparse representation of audio [1, 7, 8, 39, 44], and seek to find similar signal structures. However, similar structures do not always exist in the given inputs, especially when the inputs are short. Moreover, most of the previous work cannot handle missing lengths longer than 0.25 seconds [7]. And neither are these methods able to associate with given videos.

Another idea is to apply recent advances in audio generative tasks by using deep learning. A recent work that closely related to ours is [54], which uses videos as conditions to directly generate audio signals. However, previous methods have not explored the smoothness constraint on both sides of the to-be-inpainted audio.

To tackle these problems, our key insight is that *we can effectively exploit the context information in audio by viewing the compact audio representation of spectrogram as a continuous signal*. Inspired by recent deep models of image inpainting [38, 26, 50], we formulate the problem in the same way, regarding spectrogram as a special kind of "image", by treating time and frequency as height and width. Researchers have shown that spectrogram can be effectively processed by convolutional neural networks (CNNs) [12, 34, 52]. We believe a convolutional encoder-decoder network is able to recover high-level timbre and low-level frequency of the missing audio parts. This requires the spectrogram to contain enough yet simple information. With this motivation, we use the representation of Mel-spectrogram and design a spectrogram inpainting pipeline with generative adversarial networks (GAN) [22].

We then incorporate visual information into this pipeline. We propose the core of extracting desired information is to find a joint feature space where audio and video are synchronized so that the shared rhythm information could be provided to the network. Finally, a WaveNet [45] decoder with mixture logistic loss is trained to recover high-quality audio from the spectrogram for the target source (instruments for music). The WaveNet decoder also benefits us at utilizing previous clean data. Since our spectrogram inpainting pipeline is inspired by the computer vision community, and the model itself is designed to be able to extend to audio-visual version, the proposed audio inpainting system is, in principle, infused by visual signal. Therefore we formally term our framework as Vision-Infused Audio Inpainter (VIAI).

Our **contributions** are summarized as follows. (1) We propose a novel framework for audio inpainting inspired by image inpainting to perform on spectrograms. An inpainted spectrogram is then converted into coherent audio with a WaveNet decoder. (2) We incorporate visual cues into this framework and, to the best of our knowledge, design the first system targeting video-associated audio inpainting. (3) Along with our model, we also introduce novel training strategies for effective learning. Extensive experiments show that our framework can successfully handle missing music clips at lengths around 0.8 seconds with only 4 seconds inputs. Such lengths cannot be handled by most of the existing audio inpainting methods. (4) We extend the original MUSIC dataset [52] to a richer version, named MUSICES, to benefit the entire audio-visual research community.

2. Related Work

Audio Inpainting. Previous research mainly resolves audio inpainting from a signal processing point of view. Sparse approximation in the time-frequency domain has been explored in [1, 43], but silence will be introduced when gap exceeds 50ms. Self-similarity has been employed to inpaint gaps up to 0.25 seconds using time-evolving features [7]. Recently, using similarity graphs, [39] proposes to inpaint long music segments, but it cannot handle segments shorter than 3 seconds. More importantly, similar frames do not exist for certain in the given intact input areas. This kind of method would fail when such cases are presented. Only very recently, some contemporary works exploit CNNs for audio inpainting [31].

Audio Synthesis. By applying deep learning, generative models such as SampleRNN [32], WaveNet [45] and their variants [16, 35] have successfully generated high fidelity raw audio samples. One of the most important developments is to use them as decoders for conditional audio generation tasks such as Text-to-Speech Synthesis (TTS). For example, acoustic features designed by domain expertise have been used as inputs for audio synthesis based on SampleRNN [2] and WaveNet [5, 21]. Latter in Deep Voice 3 [40] and Tacotron 2 [42], Mel-spectrogram has been successfully used to train WaveNets. Inspired by their works, we adopt a similar structure to generate raw audios in the proposed task.

Audio-Visual Joint Analysis. Recent years witness the rapid growth in audio-visual joint learning tasks such as audio-visual speech recognition [13, 12], learning audio-visual correspondence [3, 4, 6], localization [52, 41], synchronization [14, 36, 29], audio to visual generation [11, 53, 23], visual to audio generation [18, 54, 37], visually aided source separation [36, 17, 19], and spatial audio generation [20, 33].

Among them, works that map visual to sound *i.e.* source separation and sound generation are more related to ours. Source separation works perform more often on spectrograms. Zhao *et al.* [52] use Short-Time-Fourier-Transforms (STFT) to realize source localization and separation. Similarly, Ephrat *et al.* [17] use talking face videos to form masks on STFTs of speech signals to achieve speech separation. Owens and Efros [36], on the other hand, concatenate visual features in the bottleneck of a spectrogram U-net. Unlike source separation that all audio information to be recovered has already existed, generating new audio would be much more difficult. In [37], hitting sound is predicted specifically. And [54] directly generated sound for in the wild videos using SampleRNN. But our work has a different setting from all of them.

Image Inpainting. Image inpainting [38, 49] is a well-studied topic in computer vision and graphics. Deep learning methods have been successfully applied to this

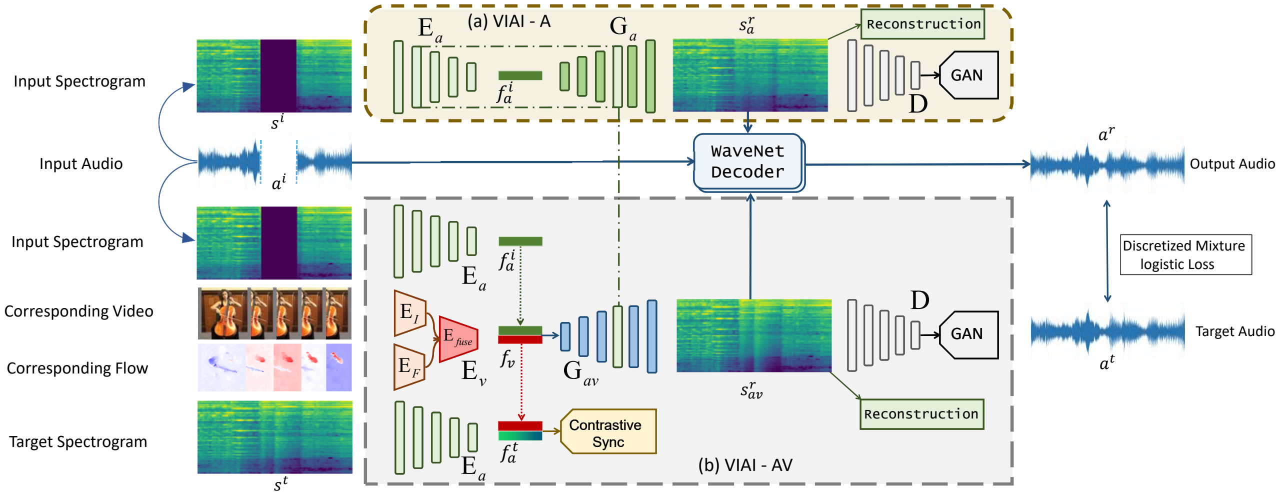


Figure 2. The whole Vision-Infused Audio Inpainter system pipeline. In the above bracket (a) is the VIAI-A inpainting schedule. First the input corrupted audio is processed into Mel-spectrogram with a missing piece. An encoder-decoder pair $\{\mathrm{E}_a, \mathrm{G}_a\}$ with one skip connection at the second layer restores the spectrogram to a complete one s_a^r. Below (b) is the VIAI-AV pipeline. Bottleneck features f_a^t, f_v are extracted from audio and visual encoder E_a and E_v. They are trained to be synchronized with each other. At the same time, concatenating f_v with the distorted audio feature f_a^i from E_a, the decoder G_{av} reconstruct the spectrogram s_{av}^r base on both information. The reconstruction output results s_a^r and s_{av}^r are constrained with reconstruction loss and GAN loss with the target s^t. Finally the results are sent into the pretrained WaveNet decoder to generate raw audio.

field with GANs. Context Encoders [38] firstly trains deep encoder-decoder networks for inpainting with large holes. [26] extends it with global and local discriminators as adversarial losses. Recently, researchers dig into the combination of deep learning methods and exemplar-based approaches [48, 46]. Same practice could be applied in our framework, but for simplification of our proposed method, we just borrow the encoder-decoder baseline.

3. Our Approach

We introduce our Vision-Infused Audio Inpainter (VIAI) in this section. VIAI consists of two parts, a pure audio module "VIAI-Audio (VIAI-A)", and an audio-visual joint inpainting module "VIAI-Audio-Visual (VIAI-AV)". They all share a modified WaveNet decoder. Fig 2 depicts the entire pipelines. The main idea is to turn audio inpainting into spectrogram inpainting in an image inpainting style. We first borrow undistorted audio around the missing part to form an input audio segment a^i. Then it is transformed into its Mel-spectrogram representation s^i given the missing data length and position. Our goal is to reconstruct a spectrogram s^r, which is as similar to the target one s^t as possible.

3.1. Audio Inpainting as Spectrogram Inpainting

Pipeline. The yellow bracket at the top of Fig 2 (a) shows the whole procedure of VIAI-A. We adopt an encoder-decoder architecture $\mathrm{Net}_a = \{\mathrm{E}_a; \mathrm{G}_a\}$ with one skip connection. The bottleneck feature f_a^i is a 1-d feature map (size of 1 × time × channel), which gives the network the ability to deal with different input lengths. The output of the network is the reconstructed spectrogram $s_a^r = \mathrm{Net}_a(s^i)$.

Reconstruction. While the skip connection benefits the network to directly take advantage of low-level information of the clean spectrogram by simple up-sampling operations, we design a weight adjusting training scheme to construct the missing part rely on high-level information from the bottleneck. Let $s_{\{m\}}^t$ be the target of the originally missing spectrogram parts and $s_{a\{m\}}^r$ be the predicted corresponding parts where m denotes "missing". When applying the reconstruction L_1 loss, the weights between the originally clean and missing areas on the prediction and the target varies according to training time. The L_1 reconstruction loss can be written as:

$$\mathcal{L}_{re}^a = \eta_1(t)\|s^t - s_a^r\|_1 + \|s_{\{m\}}^t - s_{a\{m\}}^r\|_1, \qquad (1)$$

where $\eta_1(t)$ is a parameter which decays with the training steps, and set to a very small value after certain time. We find that if $\eta_1(t)$ is fixed to 1, the network will learn mainly up-sampling. But if it is set to be very small at the first place, the network cannot restore the clean spectrogram clearly thus audio smoothness could be hurt.

Besides, a discriminator D is trained with Patch-GAN [27] objectives to maintain the local coherence and global similarity:

$$\mathcal{L}_{\mathrm{GAN}}^a(\mathrm{Net}_a, D) = \mathbb{E}_{s^t}[\log \mathrm{D}(s^t)] + \mathbb{E}_{s^i}[\log(1 - \mathrm{D}(s_a^r)] \qquad (2)$$

The total generation loss for VIAI-A is written as $\mathcal{L}^{a}_{Gen}$. β is a hyper-parameter that leverages the two losses.

$$\mathcal{L}^{a}_{total} = \mathcal{L}^{a}_{Gen} = \mathcal{L}^{a}_{\text{GAN}} + \beta\mathcal{L}^{a}_{re}. \tag{3}$$

3.2. Joint Visual-Audio Spectrogram Inpainting

Pipeline. The pipeline for VIAI-AV is illustrated in the lower part (b) of Fig 2. It evolves into a conditional inpainting problem by introducing the video encoder E_v along with a synchronization module. The structure of audio encoder E_a is kept unchanged. With the feature extracted by E_v to be f_v, we aim to generate $s^r_{av} = \text{G}_{av}(\text{E}_a(s^i), f_v)$.

Infusing Visual Cues. The video corresponding to the target audio is provided. We believe motion information [30] is strongly associated with the change of audio melody, *i.e.*, intense movements with rapid rhythms, so optical flows are extracted. Besides, [54] shows that using both image and flow data can help improve direct audio generation results from videos. Each image and flow within this video are sent into encoder E_v, which contains ResNet encoders E_I, E_F and down-sampling convolution layers E_{fuse}. Note that we control the down-sampling rate of E_{fuse} to let f_v match the size of f_a.

Audio-Visual Synchronization. Redundant information is contained in videos for audio reconstruction such as person appearances, the position of the instruments, and changing of background settings. To capture the association between videos and audios, we propose to find a joint audio-visual space with synchronized rhythm information. In this joint space, visual feature f_v is expected to be close to its corresponding intact target audio feature $f^t_a = \text{E}_a(s^t)$. We choose to use the contrastive loss as performed in [14, 29] that maps features into the same space. The training objective is to minimize the distance between synchronized audio and video features and force the distance between unpaired data to be larger than a certain margin γ :

$$\begin{aligned}\mathcal{L}_{Sync} = &\sum_{n=1}^{N} \|f^t_{a(n)} - f_{v(n)}\|_2^2 + \\ &\sum_{n\neq m}^{N,N} \max(\gamma - \|f^t_{a(n)} - f_{v(m)}\|_2, 0)^2, \end{aligned} \tag{4}$$

where N is the number of data in one batch, and γ is set to 1. All the features are normalized first before implementation. The negative samples are drawn in a similar way as [29].

Reconstruction with Probe Loss. The video feature f_v is then concatenated with the distorted bottleneck audio feature f^i_a to form f_{av}, and sent to the new audio-visual decoder G_{av} for spectrogram reconstruction s^r_{av}. The training objective is the same as section 3.1, only substituting the subscript from a to av and get the generation loss: $\mathcal{L}^{av}_{Gen} = \mathcal{L}^{av}_{\text{GAN}} + \beta\mathcal{L}^{av}_{re}$.

In this video-associated scenario, crucial information about the missing piece is expected to be extracted from the condition feature f_v by the decoder G_{av}. As f^t_a is the compression of the clean spectrogram, information recovery from f^t_a is easier and more obvious. So we reconstruct $s^r_{aa'} = \text{G}_{av}(\text{E}_a(s^i), f^t_a)$ using a similar $\mathcal{L}^{aa'}_{Gen}$ as a *probe loss* to guide the learning of the networks. The idea is that while we restrict $f_v \approx f^t_a$ by applying the synchronization loss, we can suppose $\text{G}_{av}(\text{E}_a(s^i), f_v) \approx \text{G}_{av}(\text{E}_a(s^i), f^t_a)$. The process of this additional clean-audio-based inpainting module can be specifically named as VIAI-AA'. The success of generating terrific results with VIAI-AA' also proofs the ability of passing information from the bottleneck to the output.

The overall objective of VIAI-AV can be written as:

$$\mathcal{L}^{av}_{total} = \eta_2(t)\mathcal{L}^{aa'}_{Gen} + \mathcal{L}^{av}_{Gen} + \mathcal{L}_{Sync}. \tag{5}$$

$\eta_2(t)$ is a decay parameter that is similar with $\eta_1(t)$.

3.3. Spectrogram to Audio

WaveNet Decoder. At the end of VIAI, a WaveNet decoder is attached for both the branches. Our choice of Mel-spectrogram is a way of data compression. With less information to recover for spectrogram inpainting, it is more complicate to transform it back into raw audio signals. So we utilize a modified version of the WaveNet architecture [45] to decode spectrogram into raw audio samples. WaveNet is an autoregressive model that is composed of dilated convolutions and non-linear activations. During training, it can take raw audio data as input and Mel-spectrogram as temporal conditions to predict the next-time-step audio in a teacher-forcing way. The Mel-spectrogram is first processed using up-sampling convolutions to match the sampling rate of raw audio data. During inference, WaveNet takes in one raw audio and upsampled spectrogram data at each time step, and generates the next time step's raw audio data. It models the conditional distribution between audio data and spectrogram $p(\mathbf{a}|\mathbf{s})$:

$$p(\mathbf{a}|\mathbf{s}) = \prod_{t=1}^{T} p(a_{(t)}|a_{(1)}, \cdots, a_{(t-1)}, s_{(t)}) \tag{6}$$

We follow Parallel WaveNet [35] and Tacotran 2 [42] to use the discretized mixture logistic loss for training. One WaveNet model is pretrained for each class using clean audio samples and Mel-spectrograms in the dataset. A uniform WaveNet can also be trained in the same manner of multi-speaker TTS.

Conditioning on Past Audio. In audio inpainting task, instead of simply modeling $p(\mathbf{a}^r|\mathbf{s}^r)$, we take advantage of the WaveNet to rely the generation on both spectrogram and previous clean samples to model $p(\mathbf{a}^r|\mathbf{s}^r, \mathbf{a}^i)$. Suppose the audio data is missing from time step t_0 to T, the distribution

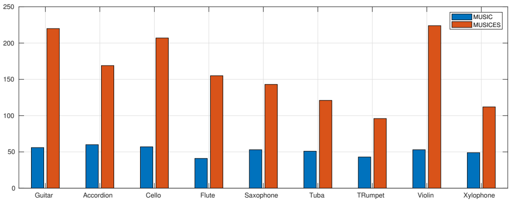


Figure 3. Data statistic comparing with the original MUSIC dataset. The x-axis is the class name and the y-axis is the number of videos per class.

we model for the reconstructed audio a^r given existing input audio a^i and reconstructed spectrogram s^r at time step t ($t > t_0$) can be written as:

$$p(\mathbf{a}^r|\mathbf{s}^r,\mathbf{a}^i) = \prod_{t=t_0}^{T} p(a^r_{(t)}|a^i_{(1)},\cdots,a^i_{(t_0-1)}, a^r_{(t_0)},\cdots,a^r_{(t-1)},s^r_{(t)}) \tag{7}$$

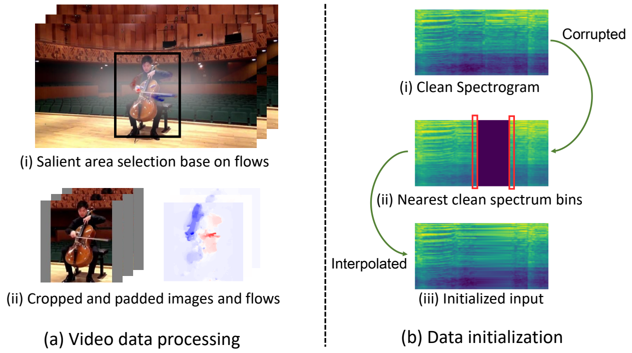
Finally, this customize WaveNet decoder can be integrated into our framework to constitute an end-to-end raw audio inference and training system.

4. MUSIC-Extra-Solo Dataset

Selection and Organization. The strong association between audio and video can usually be found in videos of instrument playing. For example, the positions of hands and movements of the bow on strings can cast certain audio notes. But normally people cannot analysis the music notes according to only visual information. This provides our task with a suitable and challenging data option, so we turn to the recently proposed MUSIC dataset [52]. However, the released version has only around 50 videos for each class, which is not sufficient. Therefore, we extend the MUSIC dataset to approximately triple its original size on 9 of its major instruments. The additional videos are all solos, thus our extension is called the MUSIC-Extra-Solo (MUSICES) dataset. The statistics of the new dataset compared to the original one are summarized in Fig. 3.

Realistic Recorded Data. Note that different from artificial music data generated with digital inference software such as MIDI, and videos recorded in a controlled lab environment, music data in MUSICES are mostly home camera recorded with minor background noise, which cast great difficulty for audio generation. The data are selected to be stable with good quality. In the original MUSIC dataset, important movements in certain videos could be invisible. This kind of video is kept out in our dataset. Different acoustic recording environments lead to domain differences of audios even in one single class, leading to a great challenge to our task.

Detecting Video Shots. We also detect the shots changing within the dataset and provide the begin and end time of each shot. We observe that videos may contain black transition frames and clips that are silent before the player starts playing. So we split the videos according to our detected shots and abandon those non-auditory ones. Besides, the first 6 seconds of each video is cut out for data cleaning. Note that the train/test sets are divided first before cutting the videos in shots.


(i) Salient area selection base on flows
(ii) Cropped and padded images and flows
Corrupted
(i) Clean Spectrogram
(ii) Nearest clean spectrum bins
Interpolated
(iii) Initialized input

(a) Video data processing

(b) Data initialization

Figure 4. Data pre-processing. (a) illustrate the procedure of flow-based salient area cropping for videos. (b) shows the results of spectrogram interpolating initialization

Set-Splitting Protocol. The train/test split is performed at the video level. Specifically, we split 10% of the videos as a fixed testing set and randomly sampled 5% as a held-out validation set.

5. Experiments

Data Processing. Data processing is important in the realization of our approach, so we elaborate in this part. All audio samples are preprocessed to 16kHz sampling rate, then all raw audio amplitudes are normalized to between -1 and 1. Our Mel-spectrograms can be computed by firstly performing STFT using a frame length of 1280 points (corresponding to 80ms) and a hop size of 320 points (20ms). The STFT magnitude is transformed to Mel scale using an 80 channel Mel filterbank with a frequency span from 125Hz to 7.6kHz, followed by log dynamic range compression. The spectrograms are normalized to between 0 and 1.

The spectrogram frame length and hop size are designed to map the 12.5 frame rate of its corresponding video. So temporally one video frame can be mapped to 4 spectrum bins. Optical flows are extracted by using TV-L-1 algorithm [51] and bounded to be maximum 20 pixels. Salient areas inside a video are approximated by setting a threshold according to the average of all optical flow values over the video. Images and flows in one video are all cropped to this rectangular area with motion detected, and padded to be square. Fig. 4 (a) depicts the procedure. Finally, the pixel values of images and flows are normalized to between -1 and 1.

Score \ Approach	SampleRNN [32]	Visual2Sound [54]	bi-SampleRNN	bi-Visual2Sound	VIAI-A	**VIAI-AV**	**VIAI-AA’**
PSNR	9.1	10.2	12.8	13.6	22.2	**23.2**	**26.6**
SSIM	0.33	0.35	0.38	0.41	0.61	**0.64**	**0.75**
SDR	4.89	3.70	4.20	4.72	6.54	**6.63**	**6.89**
OPS	51.1	51.3	51.2	52.2	52.4	**56.3**	**56.7**

Table 1. Quantitative results. The upper half are the evaluations of spectrograms and the lower half are the evaluation of audios. The maximum of OPS is 100. Larger values are better among these metrics.

Model Configurations. The audio encoder E$_a$ consists of 5 stride-2 convolution layers with 3×3 kernels. The original 80 frequency bins are compressed to 1 by a final pooling layer. Both the image and flow encoders E$_I$ and E$_F$ adopt the ResNet-18 [24] architecture. One 256-length feature vector can be obtained from each image and flow. Then the features from one video clip are concatenated along the time axis. The following E$_{fuse}$ has two stride-2 1d convolutions.

The decoder G$_a$ has 15 convolution layers with 6 bilinear upsample layers. The skip connection is at after the last upsample layer. Decoder G$_{av}$ different from G$_a$ only at the first convolution layer. It takes in twice the original feature length. As for the WaveNet decoder, we use 24 dilated convolution layers grouped into 3 dilation cycles instead of the original 30 layers for computational efficiency. One set of encoder-decoder and WaveNet model is trained for each class in the dataset.

Experimental Settings. Throughout our experiments, we only consider missing lengths longer than traditional settings. When the missing length is short, differences between methods become difficult to be discriminated, and this problem is more challenging and realistic when the missing length is longer so this is the focus of our paper.

Our choice of training input data is 4s. The distortion is shorter than 1s but longer than 0.4s. The 4-second raw audio corresponds to an 80×200 size spectrogram and maps to 50 video frames. The bottle-neck feature map is extracted to be 256×13 with 13 to be the compressed time dimension.

Implementation Details. During training, we manually crop a clip randomly within the clean spectrogram to create distortion. Different from image inpainting, the distorted part will be along the time axis (see Fig 4 (b) for visualization of input spectrogram). Based on the continuity of audio data, we initialize it to be the interpolation of the nearest clean spectrum bins as shown in Fig 4 (b), instead of averaging “pixel” value like done in image inpainting. This interpolation-based initialization can directly lead to reasonable results under certain circumstances where the missing part is a stable music note but would fail in most cases.

Our implementation is based on PyTorch and trained on 4 Titan X GPUs. Networks are trained using Adam optimizer [28] with learning rate set to be 1e-4. The batch size is 64 when training VIAI-A and 16 for VIAI-AV. The decay parameter $\eta_1(t)$ and $\eta_2(t)$ are set to be $\max(0.1, 0.9^{step/1000})$. The synchronization loss $\mathcal{L}_{Sync}$ only updates video encoder E_v as this stabilizes training.

Competing Methods. We validate our spectrogram inpainting is superior to deep learning-based autoregressive audio generation methods with the listed baselines. **SampleRNN** [32] has the ability to predict long-term audios with or without input conditions. We adopt it as an audio inpainting baseline. Then we reproduce **Visual2Sound** [54] as audio-visual baseline. Note that in original [54] paper, only ImageNet and action recognition pretrain network is used for feature extraction. For a fair comparison, we initialize their video extraction network to be our synchronizing pretrained ones. Also, we train an inverse SampleRNN model and fuse the outputs from both sides to create a **bi-directional SampleRNN** model. A similar **bi-Visual2Sound** model is also implemented. These approaches are compared with our method VIAI-A, VIAI-AV and a particular reference result of VIAI-AA’, which is described in section 3.2 when reconstructing $s^r_{aa'}$. All experiments are conducted on the same set of data with the same pre-processing steps as described above. Note that we also reproduce state of the art traditional audio inpainting method which can handle the longest distortion [7], but it fails to generate any results on our setting.

5.1. Quantitative Evaluation

Due to the limitation of paper length, we specifically show the results of **cello**, as a particular case study for quantitative evaluation. During the evaluation, the distorted length is fixed to be 0.8s for comparison. In the same way as training, the whole input audio length is selected to be 4 seconds. Only the missing area is considered under this setting, and 20 corrupted segments are sampled from each video in the test set for evaluation.

Evaluation for Spectrograms. We first evaluate the directly inpainted results of spectrograms by regarding them as images in the criterion of PSNR and SSIM [47] (larger is better). For our baselines [32] and [54], the audios are first generated then converted to Mel-spectrogram.

Evaluation for Audios. We adopt audio evaluation protocols SDR and OPS from the audio-source separation community to evaluate the final inpainted raw audio results. SDR is the Signal to Distortion Ratio that directly com-

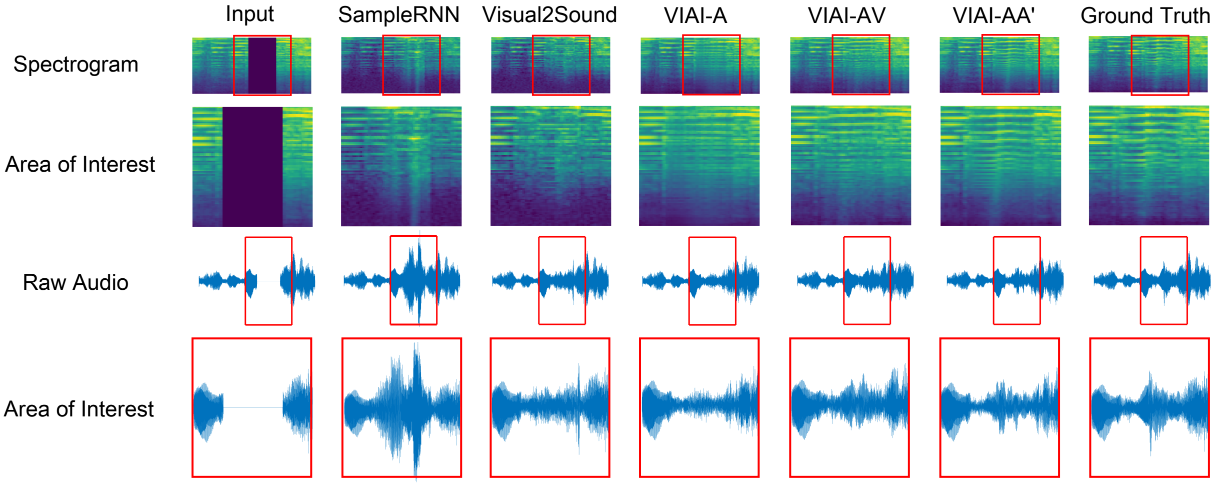


Figure 5. Qualitative results of a 0.8s distortion at an arbitrary position for different methods. The area of interest is the part in its corresponding red bracket above. Better viewed with zoom-in.

Table 2. Users' Mean Opinion Scores. Lager is higher, with the maximum value to be 5.

MOS on \ Approach	SampleRNN [32]	Visual2Sound [54]	VIAI-A	**VIAI-AV**	**VIAI-AA'**
Audio Quality	2.51	2.20	3.05	**3.93**	**4.35**
Audio-Visual Coherence	2.22	2.23	3.02	**3.96**	**4.40**
Similarity with Target	2.35	2.20	2.97	**4.01**	**4.46**

Table 3. Users' MOS with bi-directional baselines.

MOS \ Approach	bi-SampleRNN	bi-Visual2Sound	VIAI-A	**VIAI-AV**
Audio Quality	2.89	2.92	3.12	**3.86**
Audio-Visual Coherence	2.76	2.90	3.05	**3.93**
Similarity with Target	2.31	2.65	3.11	**3.96**

paring the data samples numerically. Defined in PEMO-Q auditory model [25], OPS is the Overall Perceptual Score, which is also an objective assessment of audio quality proposed in [15].

It can be observed from Fig. 4 that except for the model directly borrow intact audio information for inpainting (VIAI-AA'), results with video assistance surpass that of audio-only. And our VIAI system outperforms purely autoregressive models.

5.2. Qualitative Evaluation

We visualize a case in the form of spectrogram and raw audio at Fig 5. The areas of interests are shown explicitly. The comparison shows that while autoregressive models fail to keep smoothness, our proposed VIAI-A generates visually reasonable and continuous results. Moreover, with the presence of visual information, our VIAI-AV model captures more details than VIAI-A. The result of VIAI-AA' reaches the best, which proves that information in the bottle-neck layer has indeed been used. For auditory results please refer to our video.

User Study. Numerical numbers are hard to measure the true quality of audio signal, so we conduct a user study as a complimentary evaluation. The users are asked to evaluate the results with respect to the following three criteria; (1) Audio quality. The users mark how well the inpainting qualities are by listening to audios only. (2) Audio and Visual Coherence. To evaluate how well the inpainted audios are associated with the given videos. (3) The similarity to the ground truth. Compare the inpainted results with the ground truths and decide to what extent they are similar.

We utilize the widely used Mean Opinion Scores (MOS) rating protocol. There are overall 20 users taking part in the evaluation. The procedure for audio generation is the same as quantitative evaluation. We generate 50 different inpainted audio clips with all methods shown, and randomly assign 10 of them to one of the users. The users then give the ratings ranging from 1-5 with 5 to be the highest. Finally, all opinions are averaged.

The main results are listed in Table 2. and results for bi-directional methods are conducted additionally, listed in Table 3. As illustrated, users prefer our VIAI system comparing to baselines by significant margins. Apparently, with video information infused, the system can inpaint audios that are coherent with their corresponding videos.

5.3. Ablation Study

Audio-Visual Synchronizing. We propose that the audio-visual synchronizing part is the core of extracting desired visual information into the bottle-neck feature. Theoretically, the network will directly take the short-cut of the

MOS on \ Approach	VIAI-AV' (no sync)	VIAI-AV (no prob)	VIAI-AV (no con)	**VIAI-AV**
Audio Quality	2.90	3.59	3.00	**3.93**
Audio-Viusal Coherence	2.95	3.65	3.17	**3.96**
Similarity with Target	3.00	3.56	3.49	**4.01**

Table 4. Ablation study with Mean Opinion Scores.

Class	Violin	Accordion	Guitar	Flute	Xylophone	Trumpet	Saxophone	Tuba	Average
VIAI-A	21.1|0.64	22.2|0.59	21.3|0.58	22.4|0.60	20.2|0.56	20.2|0.62	21.5|0.59	20.0|0.57	21.2|0.60
VIAI-AV	22.4|0.66	23.6|0.61	21.9|0.61	23.5|0.63	21.1|0.58	21.0|0.64	22.5|0.60	21.2|0.57	22.2|0.62

Table 5. PSNR|SSIM results on all classes.

Approach \ Score	PSNR	SSIM
VIAI-A $\eta_1(0)$	21.8	0.60
VIAI-A $\eta_1(+\infty)$	21.6	0.59
VIAI-A (old ini)	21.5	0.58
VIAI-A	**22.2**	**0.61**
VIAI-AV' (no sync)	21.8	0.62
VIAI-AV (no prob)	22.5	0.63
VIAI-AV	**23.2**	**0.64**

Table 6. Ablation study with PSNR and SSIM metrics.

original VIAI-A path to inpaint base on spectrograms. We believe to use solely the reconstruction loss on VIAI-AV will render results similar to VIAI-A. The network trained without it is denoted by VIAI-AV'.

Probe Loss of VIAI-AA'. Then we investigate the help of the probe loss term $\mathcal{L}_{Gen}^{aa'}$. Besides the already shown results in Section 5.1 and 5.2, which demonstrate that latent information can be extracted from the bottle-neck, we further explore the influence of the existence of the loss term. The model is called VIAI-AV (no prob).

Weight Adjusting for Reconstruction To validate the effectiveness of the weight adjusting term $\eta_1(t)$ and interpolation initialization, we train extra experiments on AIVI-A by setting the coefficients of the loss term to be $\eta_1(0)$ and $\eta_1(+\infty)$. The experiment with the traditional fix value initialization is also performed as VIAI-A (old ini).

WaveNet Conditioning. Lastly, we use WaveNet to condition the generation on past results to further ensure smoothness. The training outcome without the conditioning term is addressed as VIAI-V (no con).

Ablation Results. The results regarding the metric of PSNR and SSIM are shown in Table 4. Note that VIAI-AV (no con) shares the same inpainted spectrogram as VIAI-AV. We only perform subjective studies on these extra modified VIAI-AV methods at Table 6. As depicted in the tables, our final setting reaches optimal results regarding all kinds of criteria.

5.4. Further Analysis

Analysis for Baselines. The baseline methods are designed to generate continuous and reasonable results directly or following a probe input segment. However, the task of inpainting requires the generated parts to be coherent with both sides of the existing audio parts. Particularly, Visual2Sound [54] fails to capture fine-grained visual information when applied to instrument playing data during our re-implementation.

Results on All Classes. We conduct inpainting experiments on all 9 classes of our collected MUSICES dataset. The PSNR|SSIM results of the rest classes are shown in Table 5.

Failure Cases. Failure could happen when the ground truth is already contaminated by noise, or the changing of music notes is too severe, which can be improved in the future.

6. Conclusion

In this paper, we have studied a new task and proposed an effective system called Vision-Infused Audio Inpainter (VIAI), which is capable of inpainting realistic and varying audio segments to fill in the corrupted audio. Our model integrates the intact corresponding video information into our framework to create inpainting results, which are coherent with the videos. Specifically, we formulate the problem of audio inpainting in the form of deep spectrogram semantic inpainting, and leverage the audio-visual synchronizing supervision to create a joint space for reconstruction. The novel usage of WaveNet decoder that conditions on both previous data and the reconstructed spectrogram enables the generation of high-quality raw audio data. Compared to prior methods, our approach can handle extreme inpainting settings that could not be processed by existing works, and it achieves audio-visual coherence audio-inpainting for the first time. Furthermore, an enhanced multi-modality dataset named MUSICES is contributed to the community for future audio-visual research.

Acknowledgements. We thank Yu Liu and Yu Xiong for their helpful assistance. This work is supported in part by SenseTime Group Limited, and in part by the General Research Fund through the Research Grants Council of Hong Kong under Grants CUHK14202217, CUHK14203118, CUHK14205615, CUHK14207814, CUHK14213616.

References

[1] Amir Adler, Valentin Emiya, Maria G Jafari, Michael Elad, Rémi Gribonval, and Mark D Plumbley. Audio inpainting. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):922–932, 2012. 1, 2

[2] Yang Ai, Hong-Chuan Wu, and Zhen-Hua Ling. Samplernn-based neural vocoder for statistical parametric speech synthesis. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5659–5663. IEEE, 2018. 2

[3] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 609–617. IEEE, 2017. 2

[4] Relja Arandjelović and Andrew Zisserman. Objects that sound. In *ECCV*, 2018. 2

[5] Sercan O Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al. Deep voice: Real-time neural text-to-speech. *arXiv preprint arXiv:1702.07825*, 2017. 2

[6] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*, pages 892–900, 2016. 2

[7] Yuval Bahat, Yoav Y Schechner, and Michael Elad. Self-content-based audio inpainting. *Signal Processing*, 111:61–72, 2015. 1, 2, 6

[8] Giannis Chantas, Spiros Nikolopoulos, and Ioannis Kompatsiaris. Sparse audio inpainting with variational bayesian inference. In *Consumer Electronics (ICCE), 2018 IEEE International Conference on*, pages 1–6. IEEE, 2018. 1

[9] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *ECCV*, 2018. 1

[10] 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. ACM, 2017. 1

[11] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? In *BMVC*, 2017. 1, 2

[12] Joon Son Chung, Andrew W Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *CVPR*, pages 3444–3453, 2017. 2

[13] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103. Springer, 2016. 2

[14] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016. 2, 4

[15] Valentin Emiya, Emmanuel Vincent, Niklas Harlander, and Volker Hohmann. Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2046–2057, 2011. 7

[16] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. Neural audio synthesis of musical notes with wavenet autoencoders. *arXiv preprint arXiv:1704.01279*, 2017. 2

[17] 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. *arXiv preprint arXiv:1804.03619*, 2018. 2

[18] Ariel Ephrat and Shmuel Peleg. Vid2speech: speech reconstruction from silent video. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017. 1, 2

[19] 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)*, pages 35–53, 2018. 2

[20] Ruohan Gao and Kristen Grauman. 2.5d-visual-sound. *CVPR*, 2019. 2

[21] Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. Deep voice 2: Multi-speaker neural text-to-speech. In *Advances in Neural Information Processing Systems*, pages 2962–2970, 2017. 2

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[23] Wangli Hao, Zhaoxiang Zhang, and He Guan. Cmcgan: A uniform framework for cross-modal visual-audio mutual generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 1, 2

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[25] Rainer Huber and Birger Kollmeier. Pemo-qa new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions on audio, speech, and language processing*, 14(6):1902–1911, 2006. 7

[26] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)*, 36(4):107, 2017. 2, 3

[27] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 3

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[29] Bruno Korbar, Du Tran, and Lorenzo Torresani. Co-training of audio and video representations from self-supervised temporal synchronization. In *Advances in neural information processing systems*, 2018. 2, 4

[30] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4463–4471, 2017. 4

[31] André Marafioti, Nicki Holighaus, Piotr Majdak, Nathanaë Perraudin, et al. Audio inpainting of music by means of neural networks. In *Audio Engineering Society Convention 146*. Audio Engineering Society, 2019. 2

[32] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. Samplernn: An unconditional end-to-end neural audio generation model. In *ICLR*, 2017. 2, 6, 7

[33] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. In *Advances in Neural Information Processing Systems*, pages 360–370, 2018. 2

[34] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017. 2

[35] Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis C Cobo, Florian Stimberg, et al. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, 2018. 2, 4

[36] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. *European Conference on Computer Vision (ECCV)*, 2018. 2

[37] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2405–2413, 2016. 2

[38] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016. 2, 3

[39] Nathanael Perraudin, Nicki Holighaus, Piotr Majdak, and Peter Balazs. Inpainting of long audio segments with similarity graphs. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018. 1, 2

[40] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: 2000-speaker neural text-to-speech. In *ICLR*, 2018. 2

[41] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018. 2

[42] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018. 2, 4

[43] Kai Siedenburg, Monika Dörfler, and Matthieu Kowalski. Audio inpainting with social sparsity. *SPARS (Signal Processing with Adaptive Sparse Structured Representations)*, 2013. 2

[44] Ichrak Toumi and Valentin Emiya. Sparse non-local similarity modeling for audio inpainting. In *ICASSP-IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018. 1

[45] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *SSW*, page 125, 2016. 2, 4

[46] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks. In *Advances in neural information processing systems*, 2018. 3

[47] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[48] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. Shift-net: Image inpainting via deep feature rearrangement. In *ECCV*, 2018. 3

[49] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5485–5493, 2017. 2

[50] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[51] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint Pattern Recognition Symposium*, pages 214–223. Springer, 2007. 5

[52] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. *European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 5

[53] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. *AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 1, 2

[54] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *CVPR*, 2018. 1, 2, 4, 6, 7, 8