Unpaired Image-to-Speech Synthesis with Multimodal Information Bottleneck*

Shuang Ma
SUNY Buffalo
Buffalo, NY
shuangma@buffalo.edu

Daniel McDuff
Microsoft
Redmond, WA
damcduff@microsoft.com

Yale Song
Microsoft
Redmond, WA
yalesong@microsoft.com

Abstract

Deep generative models have led to significant advances in cross-modal generation such as text-to-image synthesis. Training these models typically requires paired data with direct correspondence between modalities. We introduce the novel problem of translating instances from one modality to another without paired data by leveraging an intermediate modality shared by the two other modalities. To demonstrate this, we take the problem of translating images to speech. In this case, one could leverage disjoint datasets with one shared modality, e.g., image-text pairs and text-speech pairs, with text as the shared modality. We call this problem "skip-modal generation" because the shared modality is skipped during the generation process. We propose a multimodal information bottleneck approach that learns the correspondence between modalities from unpaired data (image and speech) by leveraging the shared modality (text). We address fundamental challenges of skip-modal generation: 1) learning multimodal representations using a single model, 2) bridging the domain gap between two unrelated datasets, and 3) learning the correspondence between modalities from unpaired data. We show qualitative results on image-to-speech synthesis; this is the first time such results have been reported in the literature. We also show that our approach improves performance on traditional cross-modal generation, suggesting that it improves data efficiency in solving individual tasks.

1. Introduction

Recent advances in deep generative models have shown impressive results across many cross-modal generation tasks, including text-to-image [36], text-to-speech [27], image-to-video [29], video-to-sound [56] synthesis. Training these models typically requires a large amount of paired samples with direct correspondence between instances from the different modalities, which limits their applicability to new ("unseen") modalities. Some attempts have been made

---

*Code: https://github.com/yunyikristy/skipNet
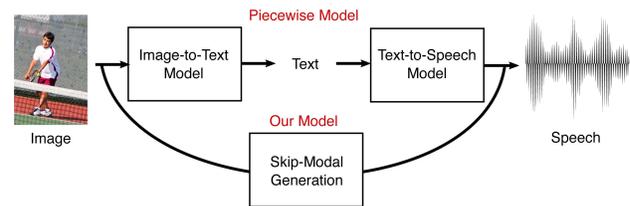
Figure 1. Cross-modal generation typically requires paired data with direct correspondence between modalities. However, this data is not always available (e.g., image-to-speech), in which case generation could be done by bridging two existing datasets via an intermediate modality (text). We propose an approach that directly generates outputs by learning multimodal correspondences from unpaired data provided by multiple disjoint datasets.

to eliminate such constraint in the context of image-to-image cross-*domain* translation, training a network on unpaired examples with the cycle consistency constraint [58, 59, 9]. However, those methods generally assume that two domains come from the same modality, e.g., images of horses and zebras; as we show later, these methods tend to fail in a cross-*modal* scenario (such as image-to-speech) where the assumption no longer holds.

In this work, we aim to learn a mapping from one modality to another without using paired samples. Our main idea is to leverage readily available datasets that do not directly provide paired samples of the two modalities we are interested in, but have "skip" correspondence between the two desired modalities via a shared one. For example, for image-to-speech synthesis we may leverage two existing datasets with image-text and text-speech pairs, where text serves as the shared modality. A naive solution to this would be training two networks separately, each solving either of the tasks with the paired data, and running them sequentially, e.g., given an image, generate text, and use it to generate speech. However, this approach is not trainable end-to-end and suffers from several issues such as domain discrepancy and information loss between the two models.

We introduce a new task *skip-modal generation* that aims to translate one modality to another by "skipping" an inter-

mediate modality shared by two different datasets. There are several reasons why this is an interesting problem to solve. From a practical standpoint, leveraging readily available datasets for solving new tasks allows for new applications. Also, training a single model with multiple datasets could potentially improve data efficiency, improving performance on the tasks each dataset was originally designed for; later, we empirically show this is indeed the case with our proposed model. From a theoretical standpoint, an ability to translate across multiple modalities may suggest that the model is one step closer to finding a unified abstract representation of different sensory inputs [33, 12]. Achieving this means information from one can be translated into any of the other modalities. Our experiments show our proposed approach can translate instances across different combinations of image, text, and speech modalities.

We focus on addressing three key challenges in skip-modal generation: learning to represent multimodal data in a uniform manner, resolving multi-dataset domain discrepancies, and learning the correspondence from unpaired data. To this end, we propose a novel generative model trainable on multiple disjoint datasets in an end-to-end fashion. Our model consists of modality-specific encoders/decoders and a multimodal information bottleneck (MIB) that learns how to represent different modalities in a shared latent space. The MIB transforms each modality-specific encoder output into the shared modality space (e.g., text) and further processes it through a memory network that serves as an information bottleneck [43]. This helps us obtain unified abstract representations of multiple modalities, capturing *"the most meaningful and relevant information"* [43] regardless of modalities or datasets. We train our model by solving two cross-modal generation tasks through the shared modality, enabling the model to learn multimodal correspondence.

We evaluate our approach on image-to-speech synthesis using two existing datasets – the COCO [6] dataset that provides image-text pairs, and an in-house text-to-speech (TTS) dataset that provides text-speech pairs – and demonstrate a superior performance over current baselines. To the best of our knowledge, this is the first time image-to-speech synthesis results have been reported. We also evaluate our approach on each of the cross-modal generation tasks the datasets were originally collected for, and show that we outperform previous state-of-the-art methods on each task, suggesting our method also improves data efficiency.

To summarize our contributions, we: (1) introduce skip-modal generation as a new task in multimodal representation learning; (2) propose an approach that learns the correspondence between modalities from unpaired data; (3) report realistic image-to-speech synthesis results, which has not been reported in the literature before; (4) show our model improves data efficiency, outperforming previous results on cross-modal generation tasks.

2. Related Work

Cross-Modal Synthesis: There has been much progress in cross-modal synthesis involving language, vision, and sound. For vision and language, image-to-text synthesis (image captioning) has been a popular task, where attention mechanisms have shown particularly strong results [48, 52, 54, 28, 32, 2]. In text-to-image synthesis, most existing methods are based on deep generative models [14, 23]. Reed *et al.* [36] and Zhang *et al.* [55] were some of the first to show promising results. Further improvements have been reported using attention mechanisms [53, 26]. For language and sound, speech-to-text (ASR) is perhaps the most mature topic of research, and great advances have been made with deep learning [17]. Text-to-speech synthesis using deep neural networks has gained much attention recently, with methods such as WaveNet [44], DeepVoice [4, 13, 34], VoiceLoop [42, 30], Char2Wav [39], and Tacotron [49, 50]. Our work is distinct from all existing lines of research in cross-modal synthesis in that we do not require paired samples to train a model. Instead, we leverage a shared modality between different datasets to learn the skip-correspondence between modalities where no paired data is available.

Cross-Domain Synthesis: Cross-domain within-modality synthesis has also been a topic of extensive study. Pix2pix [19] was the first attempt at translating across different image domains by training on paired data (e.g., sketches to photos). Since then, numerous methods have tackled the problem from an unsupervised learning perspective, eliminating the need for paired data [58, 41, 26, 5, 21]. Methods based on cycle consistency [58] have been particularly effective in this problem space. Unfortunately, cross-domain synthesis methods tend to fail on cross-modal scenarios because of the larger domain gap between different modalities. We empirically validate this in our experiments. Instead of using the cycle consistency loss, Lior *et al.* [41] translate between human faces and emojis. They leverage the fact that a face has a rigid low-dimensional structure (e.g., facial landmarks), and use a pretrained human face classifier to obtain effective representations of both human faces and emojis. Unlike their approach, in this work we make no assumption about the types of data.

3. Approach

Given two cross-modal datasets with one shared modality – e.g., a text-image dataset $\mathcal{A} = \{(\mathbf{x}_{A,i}^{txt}, \mathbf{x}_{A,i}^{img})\}_{i=1}^{N}$ and a text-speech dataset $\mathcal{B} = \{(\mathbf{x}_{B,i}^{txt}, \mathbf{x}_{B,i}^{spch})\}_{i=1}^{M}$, with text as a shared modality – our goal is to learn a network that can model data from all three modalities. We design our network with modality-specific encoders and decoders $\mathbf{E}^j$ and $\mathbf{D}^j$, respectively, with $j = \{\text{text, image, speech}\}$. Note that the definition of our model is agnostic to modali-
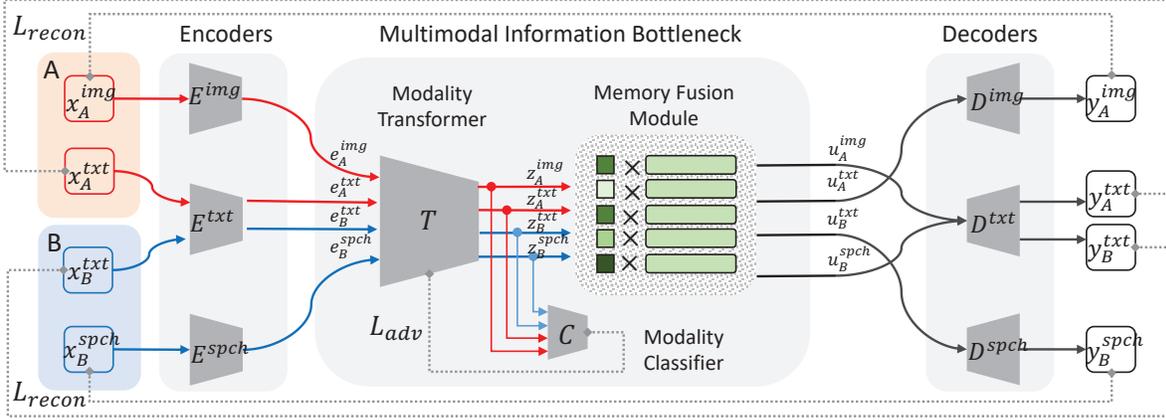
Figure 2. Our model consists of modality-specific encoders $\mathbf{E}^{(\cdot)}$ and decoders $\mathbf{D}^{(\cdot)}$, and a multimodal information bottleneck that learns to represent different modalities in a shared embedding space. We train the model end-to-end using two disjoint datasets involving image, text, and speech (A and B) by solving cross-modal reconstruction tasks. The modality classifier $\mathbf{C}$ is used only during training.

ties; the encoders/decoders can be swapped out for different modalities depending on the dataset and application.

Our main technical contribution is the multimodal information bottleneck (MIB), which consists of a modality transformer $\mathbf{T}$ and a memory fusion module $\mathbf{M}$ (see Figure 2); the modality classifier $\mathbf{C}$ is used only during training. The $\mathbf{T}$ transforms output from different encoders into the shared modality space (text); the $\mathbf{M}$ acts as an information bottleneck [43] and further processes the signals to produce compact, unified abstract representations. We use the output to generate an instance in different modalities.

3.1. Modality-Specific Encoders

Image encoder: We feed images to a three-layer CNN and perform max-pooling to obtain the output $\mathbf{e}^{img} \in \mathbb{R}^{512}$.

Text encoder: We process text into a sequence of 128-D character-level embeddings via a 66-symbol trainable look-up table. We then feed each of the embeddings into two fully-connected (FC) layers. The output sequence is fed into the CBHG [49] to obtain a sequence of 128-D embeddings; we use the original parameter settings of [49]. Finally, we apply average pooling over the sequence and feed it into one FC layer with 512 units to obtain the output $\mathbf{e}^{txt} \in \mathbb{R}^{512}$.

Speech encoder: We extract mel-spectrograms, a time-frequency representation of sound, from audio waveforms using 80 frequency bands. We treat this as a single-channel image of dimension t-by-80, where t represents the time. We feed it into a two-layer fully convolutional network and further process it using a GRU [8] with 512 units, feeding in a 5-by-80 chunk at a time. We take the last state of the GRU as the output $\mathbf{e}^{spch} \in \mathbb{R}^{512}$.

3.2. Multimodal Information Bottleneck

Neuroscientists have developed theories that the brain forms unified representations of multimodal signals [12, 33]. Modelling this computationally is very challenging because information contained in different modalities are often not directly comparable. The mapping of instances between modalities are not bijective, nor injective, nor surjective. This is especially true between text and image/speech; a sentence "There is a little blue bird" can map to images depicting different shapes and poses of a bird, or to speech signals with different intonation, tone, stress, and rhythm. Conversely, certain imagery and sounds are indescribable.

To tackle our problem of modeling multimodal data despite these challenges, we focus on how *structured* and *compact* textual representations are; image and audio contain richer information with considerably higher degrees of variability than text. Thus, we use text as a *conduit* to learn the correspondence between image and speech. This has an effect of an information bottleneck [43], which limits the flow of certain modality-specific information and helps the model learn to align image and speech from unpaired data.

Modality transformer: We start by transforming instances from image and speech modalities into a shared latent space induced by the text modality. The modality transformer $\mathbf{T}$ is a three-layer residual network that maps embeddings of each modality $\mathbf{e}^{j}$ to $\mathbf{z}^{j} \in \mathbb{R}^{256}$.

To ensure the desired transformation is performed, we use an adversarial objective that encourages $\mathbf{z}^{j}$ to be indistinguishable from each other with respect to the text modality. To this end, we design a modality classifier $\mathbf{C}$ with two FC layers and a 3-way softmax classifier representing the three modalities. We then define an adversarial loss as

$$\mathcal{L}_{adv} = \min_{\mathbf{T}} \max_{\mathbf{C}} \mathcal{L}_{\mathbf{T}} + \mathcal{L}_{\mathbf{C}} \qquad (1)$$
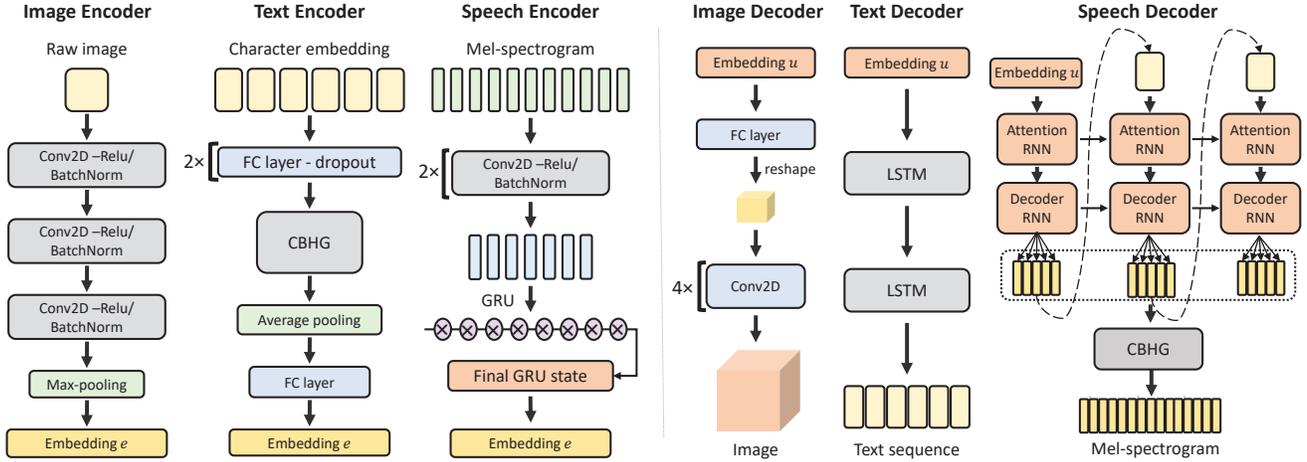
Figure 3. **Architectures of the modality-specific encoders and decoders.** We provide parameter settings in the supplementary material.

where the mini-max game is defined with two terms

$$\mathcal{L}_{\mathbf{T}} = -\mathbb{E}\big[\log \mathbf{C}(\mathbf{T}(\mathbf{e}_A^{img}))_{txt}\big] - \mathbb{E}\big[\log \mathbf{C}(\mathbf{T}(\mathbf{e}_B^{spch}))_{txt}\big]$$
$$\mathcal{L}_{\mathbf{C}} = -\mathbb{E}\big[\log \mathbf{C}(\mathbf{z}_A^{img})_{img}\big] - \mathbb{E}\big[\log \mathbf{C}(\mathbf{z}_B^{spch})_{spch}\big]$$
$$\quad - \mathbb{E}\big[\log \mathbf{C}(\mathbf{z}_A^{txt})_{txt}\big] - \mathbb{E}\big[\log \mathbf{C}(\mathbf{z}_B^{txt})_{txt}\big]$$

where $\mathbf{C}(\cdot)_j$ means we take the value from the corresponding category. To make an analogy to GAN training [14], $\mathbf{C}$ acts as a modality discriminator and $\mathbf{T}$ tries to fool $\mathbf{C}$ into believing that all $\mathbf{z}^j$ are from the text modality. In practice, we add the gradient reversal layer [11] to train our model without having to alternative between min-max objectives.

Memory fusion module: Next, we extract the uniform abstract representation $\mathbf{u}^j$ which has the most relevant information shared between paired modalities. A principled way to achieve this is through the information bottleneck (IB) approach [43], which seeks a coding mechanism that maximally preserves information in the input signal when represented using a set of external variables.

The design of our memory fusion module is partly inspired by memory networks [51] and multi-head self-attention [45]. In a nutshell, we define an external memory M that stores basis vectors representing modality-agnostic "abstract concepts," which is shared by all the modalities involved. The model reads from the memory during the forward pass, and writes to it during back-propagation. We use multi-head self-attention [45] as our coding mechanism, encoding $\mathbf{z}^j$ into $\mathbf{u}^j$ with respect to the shared M.

Formally, we define an external memory $M \in \mathbb{R}^{n_k \times d_k}$, where n_k is the number of basis vectors and d_k is the dimension of each basis vector. We also define an intermediate variable $K \in \mathbb{R}^{n_k \times d_k}$ which we use with M to form the "⟨key, value⟩ pairs" for the multi-head self-attention (K is the key, M is the value). We compute K by convolving M with 256 1D kernels of size one. Finally, we compute $\mathbf{u}^j$
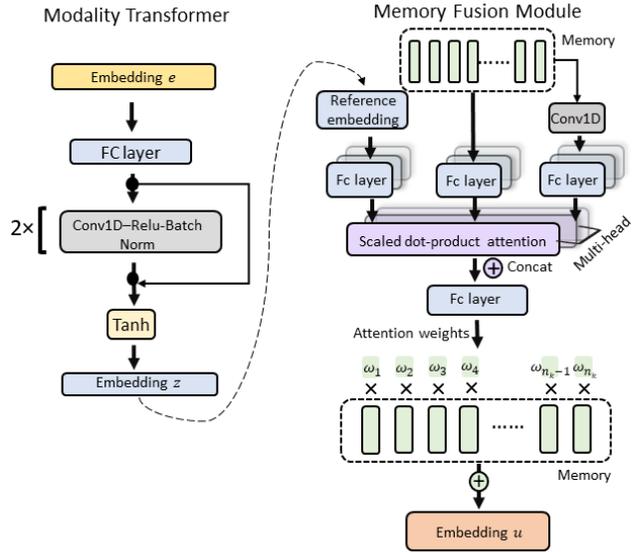
Figure 4. **Architecture of multimodal information bottleneck.**

as a linear combination of basis vectors in M with weights given by the scaled dot-product attention [45],

$$\mathbf{u}^j = \text{softmax}\left(\mathbf{z}^j K^T / \sqrt{d_k}\right) M \qquad (2)$$

Intuitively, $\mathbf{z}^j$ serves as a query to search the relevant keys to determine *where to read* from the memory. The scaled dot-product inside the softmax can be understood as a compatibility function between a query and the keys, which gives us attention scores for attending to different parts of the memory. We use multi-head self-attention with four parallel heads to make the module jointly attend to information from different subspaces at different positions.

Training of the memory fusion module: Enabling the

desired information bottleneck effect requires a careful design of the learning objective. One popular way is to impose an autoencoder-type reconstruction loss for each modality. In our scenario, this corresponds to, e.g., $\mathbf{u}^{img}$ reconstructing $\mathbf{x}^{img}$. While this would help the network learn to bottleneck superfluous information present within each modality, it would miss out on the opportunity to learn *cross-modal correspondences*. Most crucially, this will prevent the network from learning *modality-agnostic* representations, which are important for skip-modal generation, and instead learn redundant concepts presented by different modalities.

Therefore, we solve the cross-modal generation tasks as provided by two paired datasets. Specifically, we aim to reconstruct $\mathbf{x}_A^j$ from $\mathbf{u}_A^j$ and $\mathbf{x}_B^j$ from $\mathbf{u}_B^j$ in a cross-modal fashion. We define our loss as

$$\mathcal{L}_{recon} = \mathcal{L}_A^{img} + \mathcal{L}_B^{spch} + \mathcal{L}_{A,B}^{txt} \quad (3)$$

We use the l_1 loss for both image and speech modalities:

$$\mathcal{L}_A^{img} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}_{A,i}^{img} - \mathbf{D}^{img}(\mathbf{u}_{A,i}^{txt})\|_1 \quad (4)$$

$$\mathcal{L}_B^{spch} = \frac{1}{M} \sum_{i=1}^{M} \|\mathbf{x}_{B,i}^{spch} - \mathbf{D}^{spch}(\mathbf{u}_{B,i}^{txt})\|_1 \quad (5)$$

For text modality we use the cross-entropy loss:

$$\mathcal{L}_{A,B}^{txt} = -\frac{1}{N} \sum_{i=1}^{N} \mathrm{CE}\left(\mathbf{x}_{A,i}^{txt}, \mathbf{D}^{txt}(\mathbf{u}_{A,i}^{img})\right)$$
$$-\frac{1}{M} \sum_{i=1}^{M} \mathrm{CE}\left(\mathbf{x}_{B,i}^{txt}, \mathbf{D}^{txt}(\mathbf{u}_{B,i}^{spch})\right) \quad (6)$$

where we compare two sentences character-by-character according to 66 symbol categories. Note that the computation of $\mathcal{L}_{A,B}^{txt}$ depends on both $\mathcal{A}$ and $\mathcal{B}$, and the text decoder must serve a dual-purpose as an image-to-text generator and a speech-to-text generator. This allows our network to learn the skip-modal correspondence between image and speech. It also maximizes the information bottleneck effect in our memory fusion module because the external memory is conditioned on all three combinations of the modalities.

Interpretation of the multimodal information bottleneck: The two components in the MIB compensate each other with related yet different objectives. The modality transformer "drags" any given modality-specific embedding $\mathbf{e}^j$ into a shared latent subspace induced by the text modality. This helps us further process the signals in a more stable manner; otherwise the memory fusion module must deal with signals coming from three different spaces, which may have different statistical properties.

The memory fusion module then encourages $\mathbf{u}^j$ to contain the most relevant correspondence information between modalities. We share the external memory to encode embeddings from different modalities. Trained with our cross-modal reconstruction objectives, the use of a shared memory provides a strong "bottleneck" effect so that (1) it suppresses modality-specific information that does not contribute to cross-modal generation, and (2) it focuses on finding a highly-structured latent multimodal space. This allows us to obtain compact representations of the multimodal data. In Section 4, we show this improves not only the generalization ability for skip-modal generation, but also the data efficiency for each individual cross-modal generation task.

3.3. Modality-Specific Decoders

Image decoder: We feed $\mathbf{u}^j$ into one FC layer with 1024 units and reshape the output to be in $R^{4 \times 4 \times 64}$. We then upsample it with four deconvolutional layers to generate an image of size 128×128 pixels. During training, we feed $\mathbf{u}^{txt}$ to the decoder for cross-modal generation.

Text decoder: We use a two-layer LSTM as our text decoder. After initializing it with $\mathbf{u}^j$, we unroll it to generate a sentence until we get the end-of-sentence token. During training, we feed either $\mathbf{u}^{img}$ or $\mathbf{u}^{spch}$ to the decoder.

Speech decoder: We use the attention-based decoder of [47] that contains an attention RNN (a two-layer residual GRU with 256 cells) and a decoder RNN (a single-layer GRU with 256 cells). We initialize both RNNs with $\mathbf{u}^j$ and unroll them to generate a series of t-by-80 mel-spectrogram chunks. At each step, we predict multiple, non-overlapping chunks, which has been shown to speed up the convergence [49]. We convert the predicted mel-spectrogram into an audio waveform using the Griffin-Lim algorithm [15]. During training, we feed $\mathbf{u}^{txt}$ to the decoder, while at inference time we feed $\mathbf{u}^{img}$ for skip-modal generation.

3.4. Learning Objective and Optimization

We train our model by minimizing a loss function

$$\mathcal{L} = \mathcal{L}_{recon} + \alpha \mathcal{L}_{adv} \quad (7)$$

where we set $\alpha = 0.1$ in our experiments. We train the whole network end-to-end from scratch using the ADAM optimizer [22] with an initial learning rate of 0.002. We train our model for 100 epochs using a batch size of eight.

4. Experiments

We evaluate our proposed approach from two perspectives: 1) image-to-speech synthesis; 2) the effectiveness of multimodal modeling. We train our model on two datasets: COCO [6] that contains image-text samples, and an in-house dataset EMT-4 that contains 22,377 American-English audio-text samples, with a total of 24 hours. All the audio samples are read by a single female speaker.

(a) **Piecewise**: *"Three men sitting on a bench."*
Ours: *"Three men sitting on a bench in black formals."*

(b) **Piecewise**: *"A cat in a chair with keyboards."*
Ours: *"A cat and remotes sitting in the chair."*

(c) **Piecewise**: *"A child in red is eating a birthd cake."*
Ours: *"A child with red shirt is in front of a birthday cake."*

(d) **Piecewise**: *"A zebra stands beside wall."*
Ours: *"Zebras standing on ground with flowers."*

(e) **Piecewise**: *"A white cake and berr and some grap and plates."*
Ours: *"A white cake topped with berries and a plate of grapes and breads."*

(f) **Piecewise**: *"A pizza with hess on serving tray."*
Ours: *"A pizza on the table with green herbs."*

(g) **Piecewise**: *"A man is flying on snow hill."*
Ours: *"A person in red coat is skiing in snow."*

(h) **Piecewise**: *"Blue cars driving on the side road."*
Ours: *"A motorcycle and a blue car are driving on the road."*

Figure 5. **Image-to-speech synthesis results.** For the purpose of presentation, we manually transcribed audio results. Red: incorrect word predictions, green: correct/more fine-grained word predictions compared with the baseline, yellow: incorrect word pronunciation, and blue: correct/better word pronunciation compared with the baseline. Audio samples are available at `https://bit.ly/2U7741S`

4.1. Skip-Modal Generation

We validate skip-modal generation on image-to-speech synthesis both qualitatively and quantitatively, comparing ours with two baselines: the piecewise approach and CycleGAN [58]. The piecewise approach uses two individual models [54] sequentially, e.g., image-to-text followed by text-to-speech. CycleGAN [58] was originally proposed for image-to-image translation from unpaired data. To see how the model generalize to the cross-modal case, we train it directly on images and audio samples from both datasets. For a fair comparison, we design both baselines using the same encoder-decoder architectures as ours, and train them end-to-end from scratch using the same dataset.

Qualitative evaluation. We had seven human judges evaluate the generated speech from our skip-generation model and the two baselines. Twenty speech samples were generated using the models, leading to 140 independent evaluations. The judges were shown the source image and listened to the speech. They were asked to select the audio sample that had the most accurate content and the sample with speech that was closest to a human voice. They also selected the sample they felt had the highest overall quality. Examples of the samples can be found here: `https://bit.ly/2U7741S`

On average 78.6% (sd = 27.6%) of the subjects picked ours for the highest quality content. Based on audio quality,

65.0% (sd = 35.7%) of the subjects picked ours as the highest quality. Based on overall quality, 74.3% (sd = 33.9%) of the subjects picked ours. In summary, our subjects picked ours three times more frequently than either of the other baselines based on all three quality metrics.

Figure 5 shows some of the samples used in our user study; we manually transcribed the synthesized audio results for the purposes of presentation. We analyze the results by focusing on two aspects: 1) does the speech sample correctly describe the content as shown in the image? 2) is the quality of pronunciation in the speech sample realistic?

The piecewise approach sometimes incorrectly predicted words, e.g., in Fig. 5 (b) keyboards vs. remotes. We also see that our approach produces results with more fine-grained details, e.g., (g) flying vs. skiing, (h) motorcycle is missed by the baseline. These suggest that our approach is superior to the baseline in terms of modeling multimodal data.

One limitation of the piecewise approach is the inability to deal with the *domain gap* between datasets, e.g., certain concepts appear in one dataset but not in the other. This is indeed our case: the vocabularies of the two datasets overlap by only 26% (COCO has 15,200 words and EMT-4 has 17,946 words; 6,874 words overlap). This domain gap issue is reflected in our results: (e) the pronunciation of 'berries' and 'grapes' are incorrect in the baseline result, and similarly for (c) and (f). These words (berries,

	B@1	B@2	B@3	B@4	WER
CycleGAN [52]	26.2	20.1	11.3	9.2	12.1
Piecewise [54]	68.2	51.9	39.2	**30.1**	4.1
Ours	**69.2**	**52.1**	**40.8**	29.9	**3.9**

Table 1. **Skip-modal generation results.** B@k are BLEU scores.

	B@1	B@2	B@3	B@4	CIDEr	SPICE
ATT [54]	70.9	53.7	40.2	30.4	–	–
SAT [52]	71.8	50.4	35.7	25.0	–	–
RFNet [3]	76.4	60.4	46.6	35.8	112.5	20.5
UD [20]	77.2	–	–	36.2	113.5	20.3
Ours	74.1	55.2	41.1	30.6	–	–
w/ [20]	78.9	63.2	48.1	37.0	116.2	22.4
w/ [3]	**79.8**	**64.0**	**48.9**	**37.1**	**117.8**	**22.5**

Table 2. **Image-to-text generation results on COCO.**

	WER			WER
Policy [57]	5.42	DeepVoice3 [35]		10.8
DeepSpeech2 [1]	5.15	Tacotron [49]		10.6
GateConv [25]	4.80	Tacotron2 [38]		10.5
Seq2Seq [7]	4.01	GST [50]		**10.2**
Ours	**3.88**	Ours		10.5

Table 3. **Speech-to-text (left) and text-to-speech (right) results.**

	BLEU-1 / I2T	WER / S2T
Ours w/o **M**	65.2	6.99
Ours **M** $\rightarrow$ FC	65.9	6.32
Ours w/o **T**	68.6	6.01
Ours w/o L_{adv}	69.8	5.87
Ours	**74.1**	**3.88**

Table 4. **Ablation results** on image-to-text (I2T) and speech-to-text (S2T), evaluating contributions of different modules including the memory fusion module (M), the modality transformer (T), the adversarial loss (L_{adv}). **M** $\rightarrow$ FC means we replace **M** with two FC layers to match the number of parameters.

grapes, birthday, herb) do not appear in the EMT-4 dataset, which means the text-to-speech model must perform zero-shot synthesis. This is reflected in Fig. 5 (c, e, f) - see the yellow words. Our results show superior quality on those out-of-vocabulary words despite being trained on the same datasets. To quantify the word expressivity of our model, we analyzed the vocabulary size of the synthesized speech using ASR [44]. Our model produced a vocabulary of 2,761 unique words, while the piecewise baseline produced 1,959 unique words; this is 802 more words, a 40% increase over the baseline.

Finally, we show additional results in Figure 6 where we synthesize both speech and text from the same image as an input (speech results are manually transcribed). We see that the text and speech results are semantically very similar in that they describe the same content. This, together with other results above, suggests the model has learned to extract a unified abstract representation of multimodal data because different decoders can reliably synthesize samples that contain similar content – despite the speech decoder having never seen the image embeddings during training.

Quantitative evaluation. To evaluate image-to-speech synthesis results quantitatively, we use a pretrained ASR model based on WaveNet [44] and compare the text output with the ground-truth sentence corresponding to an input image from COCO. We report the results using the BLEU scores and the word error rate (WER). Table 1 shows our approach achieving the lowest WER with the highest BLEU scores (except for BLEU-4).

4.2. Cross-Modal Generation

To evaluate our approach in an objective manner, we turn to cross-modal generation where there exist state-of-the-art approaches and widely used metrics for each task.

Image $\rightarrow$ **Text:** We compare with four recent image cap-
tioning models: ATT [54], SAT [52], RFNet [20], and Up-Down (UD) [3]. The results are shown in Table 2. Note that our approach (**Ours**) uses a 3-layer CNN as the image encoder while all four baselines use deeper CNNs with pre-training/finetuning on extra datasets. Specifically, both ATT and SAT use the GoogleNet [40] which pretrained on ImageNet [10] as the image encoder. Our model, despite using a much shallower CNN, outperforms ATT and SAT by a large margin. The other two baselines use even more sophisticated image encoders: RFNet [20] combines ResNet-101 [16], DenseNet [18], Inception-V3/V4/Resnet-V2 [40], all pretrained on ImageNet [10]. UpDown (UD) [3] uses a Faster R-CNN [37] with Resnet-101 [16] pretrained on ImageNet [10] and finetuned on Visual Genome [24] and COCO [6]. For fair comparisons, we replace the 3-layer CNN with RFNet (**Ours w/ [20]**) and UD (**Ours w/ [3]**). This improves performance compared to the baselines and shows the data efficiency of our approach: Because our model can handle multimodal data effectively, it can leverage external data sources *even if* the modalities do not match. This helps our model learn more powerful multimodal representations from a large variety of data, which is not possible with the conventional bi-modal models.

Speech $\rightarrow$ **Text:** We compare our method with four ASR models, DeepSpeech2 [1], Seq2Seq [7], Policy Learning [57], and Gated Convnets [25]. All the models are trained end-to-end on the LibriSpeech corpus [31]. Particularly, similar to ours, the Seq2Seq model incorporates multi-head attention [46] to attend to multiple locations of the encoded features. For a fair comparison, we fine-tune our model on the LibriSpeech corpus.

Table 3 shows that our model outperforms the baselines
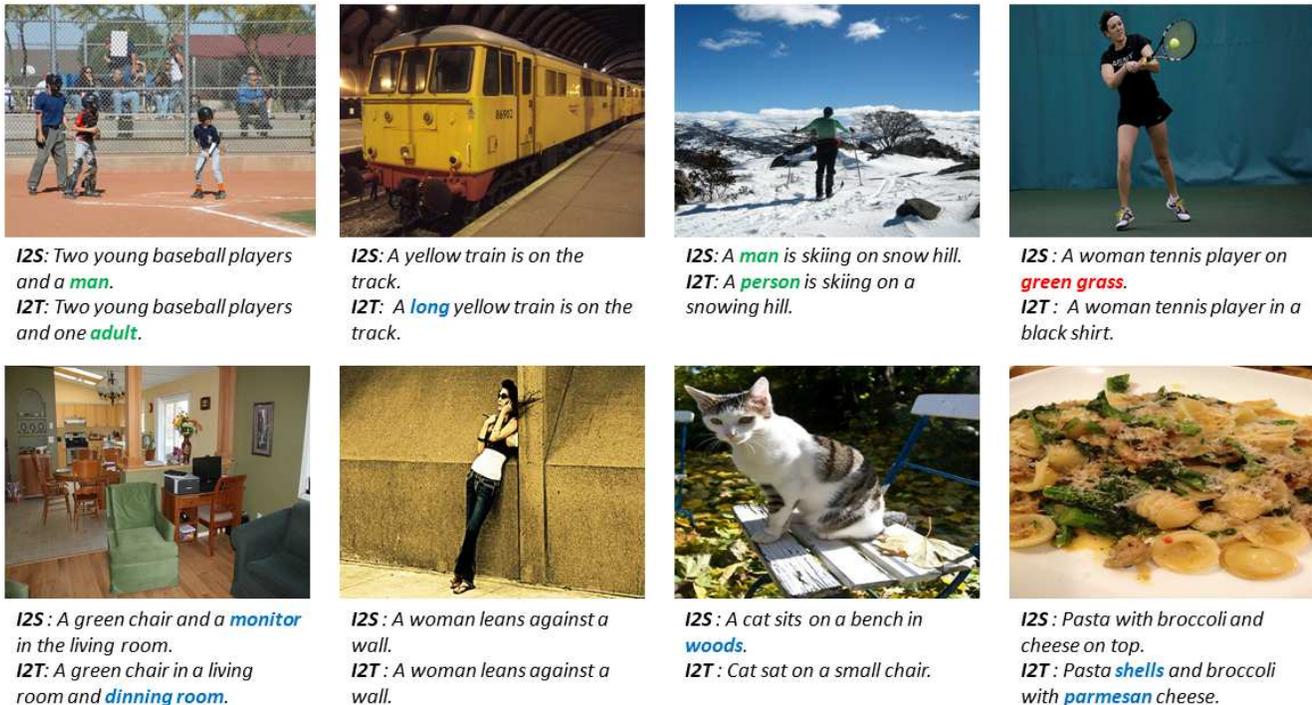
Figure 6. **Image-to-speech (I2S) and image-to-text (I2T) synthesis results**. I2S results are manually transcribed for presentation.

on speech-to-text tasks. Our multimodal information bottleneck is trained on a larger variety of data, which helps them learn more powerful representations. Also, as the text modality comes from two unrelated datasets, the cross-modal reconstruction loss (Eqn. (3)) enforces the model to solve more challenging optimization problems, which leads to improved results as seen in our experiments.

Text → Speech: We compare with four text-to-speech (TTS) models: Tacotron [49], Tacotron 2 [38], DeepVoice3 [35], and GST [50]. Tacotron [49] is an RNN-CNN auto-regressive model trained only on reconstruction loss, while GST extends it by incorporating information bottleneck layers (which they call global style tokens). For both baselines we used the same Griffin-Lim algorithm [15] as a vocoder. To evaluate the quality of the synthesized results quantitatively, we again use a pretrained ASR model based on WaveNet [44] to compute the Word Error Rate (WER) for the samples synthesized by each model. Table 3 shows that all three methods perform similarly. We believe one limiting factor is in the vocoder and expect to get better results with deep vocoders such as WaveNet [44].

Ablation Study. We investigate the contribution of the modality transformer $\mathbf{T}$ and the memory fusion module $\mathbf{M}$, evaluating on image-to-text and speech-to-text tasks.

Table 4 reports BLEU-1 scores for the image-to-text experiments and WER for the speech-to-text experiments. In both cross-modal generation tasks, the performance drops significantly when we remove the memory fusion module $\mathbf{M}$ (ours w/o $\mathbf{M}$). This suggests that the $\mathbf{M}$ plays the most

significant role in modeling multimodal data. We also replace $\mathbf{M}$ with two FC layers that have a similar number of parameters as $\mathbf{M}$. This marginally improved the performance (B@1/I2T 65.2 vs. 65.9, WER/S2T 6.99 vs. 6.32). Our model still outperforms this baseline by a large margin (74.1 and 3.88). When we remove the modality transformer $\mathbf{T}$, we also see the performance drop significantly. This shows the importance of pushing modality-specific embeddings into a shared latent space; without this component, the $\mathbf{M}$ must deal with signals coming from three different modalities, which is a considerably more difficult task. We also test the contribution of the adversarial loss (Eqn. (1)) that we use to train $\mathbf{T}$. Without this loss term, the performance is similar to that in the setting without $\mathbf{T}$, which shows the adversarial loss plays a crucial role in training $\mathbf{T}$.

5. Conclusion

We propose a novel generative model for skip-modality generation. We demonstrate our approach on a challenging image-to-speech synthesis task where no paired data is available. Unlike conventional cross-modal generation, which relies on the availability of paired data, our model learns the correspondence between image and speech directly from two unrelated datasets, image-to-text and text-to-speech, using text as a shared modality. We show promising results on image-to-speech synthesis, as well as various cross-modal generation tasks, suggesting the model also benefits from increased data efficiency.

References

[1] Dario Amodei et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *ICML*, 2016.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[4] Sercan Ömer Arik, Mike Chrzanowski, Adam Coates, Greg Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Jonathan Raiman, Shubho Sengupta, and Mohammad Shoeybi. Deep voice: Real-time neural text-to-speech. In *ICML*, 2017.

[5] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. In *NIPS*, 2017.

[6] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015.

[7] Chung-Cheng Chiu, Tara Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Katya Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, and Michiel Bacchiani. State-of-the-art speech recognition with sequence-to-sequence models. In *ICASSP*, 2018.

[8] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, 2014.

[9] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018.

[10] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[11] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.

[12] Marie H Giard and Franck Peronnet. Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *Journal of cognitive neuroscience*, 11(5):473–490, 1999.

[13] 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 *NIPS*, 2017.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*. 2014.

[15] D. Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1984.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[17] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, et al. Deep neural networks for acoustic modeling in speech recognition. *Signal processing magazine*, 2012.

[18] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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

[20] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV*, 2018.

[21] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML, 2017*.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014.

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 2017.

[25] Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert. Letter-based speech recognition with gated convnets. *CoRR*, abs/1712.09444, 2017.

[26] Shuang Ma, Jianlong Fu, Chang Wen Chen, and Tao Mei. DA-GAN: instance-level image. In *CVPR*, 2018.

[27] Shuang Ma, Daniel Mcduff, and Yale Song. Neural TTS stylization with adversarial and collaborative games. In *ICLR*, 2019.

[28] Alexander Mathews, Lexing Xie, and Xuming He. Semstyle: Learning to generate stylised image captions using unaligned text. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[29] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In *ICLR*, 2016.

[30] Eliya Nachmani, Adam Polyak, Yaniv Taigman, and Lior Wolf. Fitting new speakers based on a short untranscribed sample. In *ICML*, 2018.

[31] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *ICASSP, 2015*. IEEE.

[32] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. In *CVPR*, 2017.

[33] Pietro Pietrini, Maura L Furey, Emiliano Ricciardi, M Ida Gobbini, W-H Carolyn Wu, Leonardo Cohen, Mario Guazzelli, and James V Haxby. Beyond sensory images: Object-based representation in the human ventral pathway. *Proceedings of the National Academy of Sciences*, 101(15):5658–5663, 2004.

[34] 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.

[35] 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.

[36] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016.

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. 2015.

[38] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. *CoRR*, abs/1712.05884, 2017.

[39] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. Char2wav: End-to-end speech synthesis. In *ICLR, workshop*, 2017.

[40] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[41] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In *ICLR*, 2017.

[42] Yaniv Taigman, Lior Wolf, Adam Polyak, and Eliya Nachmani. Voiceloop: Voice fitting and synthesis via a phonological loop. In *ICLR*, 2018.

[43] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. In *Allerton Conference on Communication, Control, and Computing*, 1999.

[44] Aron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alexander Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 2017.

[47] Oriol Vinyals, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey E. Hinton. Grammar as a foreign language. *NIPS*, 2015.

[48] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[49] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. In *INTERSPEECH*, 2017.

[50] Yuxuan Wang, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A. Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *ICML*, 2018.

[51] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In *ICLR*, 2015.

[52] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[53] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018.

[54] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. *CVPR*, 2017.

[55] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *CVPR*, 2017.

[56] 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.

[57] Yingbo Zhou, Caiming Xiong, and Richard Socher. Improving end-to-end speech recognition with policy learning. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[58] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[59] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NIPS*, 2018.