Diverse Audio-to-Image Generation via Semantics and Feature Consistency

Pei-Tse Yang and Feng-Guang Su and Yu-Chiang Frank Wang
Department of Electrical Engineering, National Taiwan University, Taiwan

***Abstract*—Humans are capable of imagining scene images when hearing ambient sounds. Therefore, audio-to-image synthesis becomes a challenging yet practical topic for both natural language comprehension and image content understanding. In this paper, we propose an audio-to-image generation network by applying the conditional generative adversarial networks. Specifically, we utilize such generative models with the proposed feature consistency and conditional adversarial losses, so that diverse image outputs with satisfactory visual quality can be synthesized from a single audio input. Experimental results on sports audio/visual data verify that the effectiveness and practicality of the proposed method over the state-of-the-art approaches on audio-to-image synthesis.**

***Index Terms*—audio-to-image generation, conditional generative adversarial network, cross-modal generation**

I. Introduction

In daily life, people can imagine pictures or things based on external stimuli, such as text descriptions and sounds. With the rapid development of machine learning, computer vision, and natural language processing, many researchers have been trying to equip machines with the ability of imagination. For instance, the recent research on text-to-image [1] or the more challenging problem, audio-to-image generation. This also drives research progress in many research areas, such as cross-modality and multi-modal learning. In this paper, we demonstrate that the machine can recall multiple corresponding pictures after hearing a sound.

The recent advances in Generative Adversarial Networks (GAN) [2] also motivates the developments of recent audio-to-image generation models. Rather than solely depending on a noise vector as the input, the generators are designed to be conditioned on audio segments, which is the main idea of conditional GANs (cGANs) [3]. However, even though some techniques have been adopted to generate realistic images, such as projection discriminator [4] and recurrent adversarial network [5], this task remains challenging for synthesizing the precisely corresponding images. To address similar problems of text-to-image synthesis, many methods [6], [7], [8], [9] have been proposed to generate text-relevant images by applying the discriminator to distinguish between the ground truth image and corresponding text pair and the generated image and corresponding text pair. In order to solve text-to-image-to-text problems with better semantics-preserving ability, Qiao *et al.* proposed MirrorGAN [1] to improve the generated images by redescription.

Inspired by the above methods, we aim to develop a deep learning model for audio-to-image synthesis. More specifically, our model is able to produce diverse image outputs with consistent audio and visual semantic information can be produced, by observing the input audio data. We redesign the conditional discriminators and the regenerate of audio segments in our proposed framework, which enforces the output visual and audio content features, respectively. In our experiments, we consider real-world sports videos and apply three evaluation metrics: inception score [13], R-precision, and classification score for quantitative evaluation. We also conduct cross-modal synthesis and the multi-modal visualization for qualitative evaluation. Both evaluations support the effectiveness of our model in realizing diverse audio-to-image generation while performing favorably against recent approaches.

We highlight the contributions of our works below:

- We are among the first to address audio-to-image generation with diverse outputs with satisfactory quality.
- Our proposed generative model is realized by a condition discriminator and observing feature consistency, preserving both semantic (categorical) information and content authenticity.
- Experiments on sports video data confirm the effectiveness of our approach in generating diverse image outputs when conditioned on audio inputs.

II. Proposed Method

A. Notations and Problem Definition

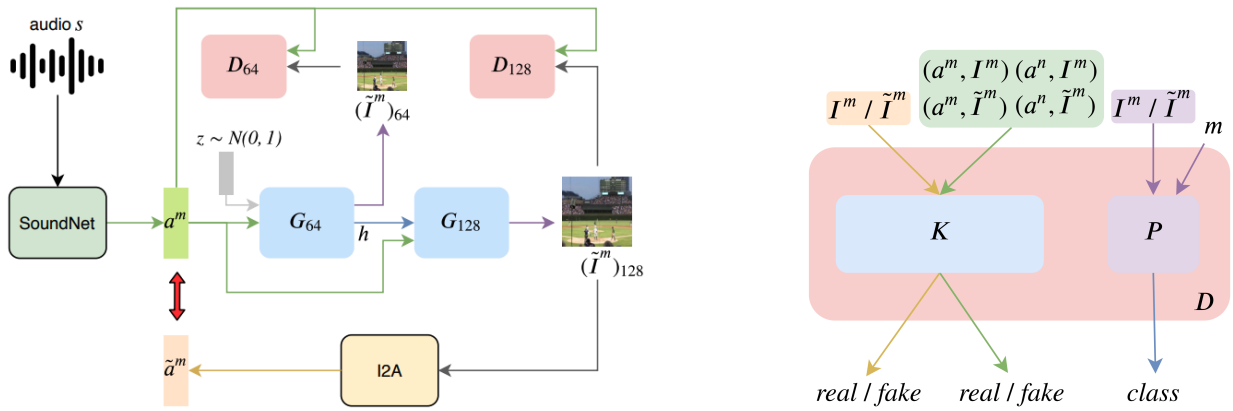
To begin with, we define the notations used in this paper. We assume that a set of N audio segment and image pairs $\{(S_i^m, I_i^m)\}_{i=1}^N$ are collected from M categories, where I_i represents the corresponding image of the audio segment S_i, and $\{m\}_{m=1}^M$ indicates the category that the set belongs to. As depicted in Fig. 1, we apply SoundNet [14] to extract the acoustic features for the audio segments $a_i^m \in \mathbb{R}^d$ (d denotes the dimension of feature). Coupled with a noise vector z, the generator G_{64} first extracts the low-dimensional latent representation $h \in \mathbb{R}^e$ (e indicates the dimension after extraction) and recovers the image:

$$(\tilde{I}_i^m)_{64} = G_{64}(a_i^m, z), \tag{1}$$

where $(\tilde{I}_i^m)_{64} \in \mathbb{R}^{64\times64\times3}$. To leverage the images with better quality, the generator G_{128} further refines the image by:

$$(\tilde{I}_i^m)_{128} = G_{128}(a_i^m, h). \tag{2}$$

We see that $(\tilde{I}_i^m)_{128} \in \mathbb{R}^{128\times128\times3}$ is conditioned on the acoustic feature a_i^m and h, which is the visual feature extracted



(a) (b)

Fig. 1. Overview of our proposed model. (a) The generators (G_{64} and G_{128}) take the audio input for producing image outputs I at different scales, whose semantic information and authenticity are enforced by conditional discriminators (D_{64} and D_{128}). We have pretrained SoundNet and image-to-audio (I2A) models in our framework; the former is to extract the audio feature a, while the latter is to preserve the resulting audio content consistency. (b) Design of the conditional discriminator D, in which K assesses the authenticity of the input images (conditioned on the audio features), while P serves as an auxiliary classifier verifying the categorical label m. Note that n denotes a category label different from m.

by G_{64}. The synthesized images, $(\tilde{I}_i^m)_{64}$ and $(\tilde{I}_i^m)_{128}$, are fed to the introduced conditional discriminator D_{64} and D_{128}, respectively. The design and properties of such discriminators will be explained in Sec. II-C.

B. Conditional Image Generation with Audio Consistency

Previously, [15], [16] exploited audio spectrograms to address audio-to-image synthesis, while recent cross-modality tasks such as text-to-image generation associated text features with the corresponding images [17], [18], [19], [1]. It is reported in [4] that, among sound feature representations, including Spectrogram, Fbank, mel-frequency cepstral coefficients (MFCCs), and SoundNet feature embedding, *SoundNet* features are preferable when it comes to image generation. Therefore, our acoustic features a are encoded by a pretrained SoundNet, which will be fed into our generator G_{64} as inputs. Moreover, in order to realize diverse image outputs, we have random noise z fed into G_{64} as well. To further refine the generated image output, we have G_{64} take $(\tilde{I}_i^m)_{64}$ and the visual latent representation h by G_{64} as the inputs to the second generator G_{128}. Furthermore, we design a unique architecture in the conditional discriminator D as depicted in Fig. 1 to make the generated images more semantically consistent.

The objective function of each generator $\mathcal{L}_{G_j}$ (j denotes the image dimension) is:

$$\mathcal{L}_{G_j} = -\mathbb{E}[K_j((\tilde{I}^m)_j, a^m, a^n) - log P_j(m|(\tilde{I}^m)_j)], \tag{3}$$

where m and n denote the category labels. Note that (K, P) are the network modules in our discriminator D, where K assesses the authenticity of the input image $\tilde{I}^m$ with the corresponding audio feature a^m, while P serves as an auxiliary classifier verifying the categorical label of the input image.

To further strengthen the relation between sound features and the corresponding synthesized images $(\tilde{I}^m)_{128}$, we propose to observe audio consistency between the acoustic feature $\tilde{a}$ extracted from $(\tilde{I}^m)_{128}$ and the input a. More specifically, as shown in Fig. 1, we apply a I2A model [20] pretrained on the training pairs (a_i^m, I_i^m) which extracts the reconstructed acoustic feature $\tilde{a}_i^m \in \mathbb{R}^d$ from the high-quality image $(\tilde{I}_i^m)_{128}$. With the above design, the acoustic feature based consistency loss is denoted as:

$$\mathcal{L}_{con} = L_2(a, \tilde{a}), \tag{4}$$

which calculates the L_2 distances between the audio feature of the input signal and that encoded by the generated image. Thus, the objective function of the generator is the summation of the above generator and audio consistency losses.

C. Conditional Discriminator with Auxiliary Classifier for Semantics Consistency

As depicted in Fig. 1, we have a conditional discriminator in our framework, which serves as a multi-task learning model for solving different tasks. Here, we design two different modules (K, P) in the discriminator D. First, K manages to distinguish the generated images from the real ones for preserving image quality. In detail, other than general loss function defined in GAN [2], we have different combinations of audio-image pairs to enforce the *audio-image* consistency. Specifically, only the pairs $(a^m, I^n)|_{m=n}$ is regarded as true, the other pairs such as $(a^m, I^n)|_{m\neq n}$ and $(a^m, \tilde{I}^m)$ are regarded as false. Second, similar to AC-GAN [10], we apply an auxiliary classifier P to recognize the associated image categories, which recognizes the categorical label m of the generated image outputs. We note that, we have discriminators introduced at the outputs of both G_{64} and G_{128}, with the only difference as the number of down-sampling blocks due to distinct image sizes. Thus, the loss function can be presented as $\mathcal{L}_{D_j}$ where j denotes the single dimension of the generated

TABLE I
QUANTITATIVE EVALUATION IN TERMS OF INCEPTION SCORE AND CLASSIFICATION ACCURACY.

	inception score	classification accuracy(%)
Ground Truth	4.79 ± 0.52	96
VAE	1.06 ± 0.01	50
Wan *et al.* [4]	2.25 ± 0.22	78
Ours	$\mathbf{3.37 \pm 0.17}$	**86**

images:

$$\begin{aligned}\mathcal{L}_{D_j} = \mathbb{E}[max(0, 1 - K_j((I^m)_j, a^m, a^n))] \\ +\mathbb{E}[max(0, 1 + K_j((\tilde{I}^m)_j, a^m, a^n))] \\ -\mathbb{E}[logP_j(m|(I^m)_j)].\end{aligned} \tag{5}$$

With loss functions for both generator and discriminator components defined, we apply standard techniques including hinge loss [12] and spectral normalization [11] to train our proposed model.

III. EXPERIMENTS

A. Dataset

The dataset that we apply for evaluation is collected by [4], which consists of 9 categories with $10,701$ sound-image pairs for training and 248 audio segments for testing. We choose two categories, baseball and soccer, to evaluate our proposed model. We note that, the original data in this dataset are with noisy labels (which were originally classified by SoundNet), we remove the data with incorrect labels before training and testing. After data cleaning and randomly splitting them into train/test sets, our dataset contains $2,065$ sound-image training pairs, and 218 audio segments for testing.

B. Quantitative Results

For comparisons, we first train a VAE model for audio-to-image generation tasks as our baseline. Then, we reproduce the state-of-the-art method [4] with the same settings. We apply the inception score [13] to measure the objectiveness and diversity of the synthesis images. Moreover, we train a classifier on ground truth images for classification accuracy evaluation. The inception score and classification accuracy are shown in Table I. Our proposed model achieved the highest performance in both evaluation metrics. Compared with the previous work [4], we improved the inception score from 2.25 to 3.37 and outperformed by 8% in the classification accuracy. In other words, our model can generate images with better quality and diversity, and at the same time, the generated images can also be more correlated with the input audio segments.

To further evaluate the visual-sound similarity between the generated images and their corresponding audio segments, we consider the metric of R-precision which has been widely used for retrieval evaluation and calculate the top-k ranked retrieval results. For each audio segment, we first form an image pool with its generated image and 99 randomly selected mismatched generated images. Then, we compute the cosine similarities between the acoustic feature and the image features in the pool. Lastly, candidates are ranked in descending similarity and we found the top-k (k = 1, 5, 10) relevant results for calculating the R-precision. From Table II, it is clear that our model obtained the highest score in each ranked k, when performing favorably against both VAE and [4].

TABLE II
RETRIEVAL PERFORMANCE EVALUATION IN TERMS OF R-PRECISION (%).

top-k	$k = 1$	$k = 5$	$k = 10$
VAE	3.2	9.2	16.5
Wan *et al.* [4]	5.5	17.4	28.4
Ours	**6.0**	**18.3**	**34.9**

C. Qualitative Results with Diverse Image Outputs

We now qualitatively examine the images generated by our proposed model. Subsequently, We present samples generated by VAE, Wan *et al.*, and our proposed model conditioned on testing audio segments in Fig. 2. For each row, the first two images belong to the soccer category, while the remaining two images belong to the baseball category. The athletes can be apparently seen in the images generated by our proposed model, and the scenes are highly associated with its audio categories.

In addition, we investigate whether the generated images are correlated with the input audio segments, and also evaluate the multi-modal visualization. We selected three audio segments from the testing data, one of which belongs to soccer category, others belong to the baseball one. Specifically, the two baseball audio segments are different. One of them is a broadcast of a baseball game, while the other is a segment in which players are playing baseball. The results are shown in Fig. 3. In each row, the first image denotes the ground truth, while the remaining images demonstrate the multi-modal results using the same audio segments. Evidently, our model can generate multi-modal samples, conditioning on the same audio segments. Specifically, the audio category and sound information are precisely related to the generated images. For instance, in the second and third rows, although the categories of input audio segments are both baseball, our results show that the former is a broadcast view, while the latter is a play view which is consistent with the ground truth images. The audio segments used for testing here can be found in https://peitseyang.github.io/audio-to-image/.

The above quantitative and qualitative results successfully verify the effectiveness of our proposed model, which is able to produce cross-modality and multi-modal image results based on the audio input.

IV. CONCLUSION

We propose a unique generative model for diverse audio-to-image generation. Conditioned on the audio input, our model is able to produce the corresponding image outputs with

(a) Soccer Samples (b) Baseball Samples

Fig. 2. Example audio-to-image generation results and comparisons between VAE, Wan *et al.* [4], and our proposed model. Note that for each column, the outputs are produced by the same audio input.

Fig. 3. Example multi-modal image outputs produced by our model. Note that each row is conditioned on a single audio input, while more than one possible image output with proper visual content can be generated by our model.

both audio and semantics consistency. The former is achieved by observing the audio feature consistency, while the latter is realized by the conditional discriminator, which not only preserves the output image quality but also simultaneously ensures the visual content to match the associated semantic label. The integration of the generative model with a coarse-to-fine architecture allows us to output diverse images with satisfactory quality. From the experiments on real-world sports audio-video data, we confirm that our model quantitatively and qualitatively performs favorably against baseline and recent methods on this challenging task.

REFERENCES

[1] T. Qiao, J. Zhang, D. Xu and D. Tao, "Mirrorgan: Learning text-to-image generation by redescription," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1505-1514.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[3] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR*, vol. abs/1411.1784, 2014.

[4] C.H. Wan, S.P. Chuang, and H.Y. Lee, "Towards audio to scene image synthesis using generative adversarial network," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 496–500.

[5] Y. Song, J. Zhu, D. Li, A. Wang, and H. Qi, "Talking face generation by conditional recurrent adversarial network," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 919–925.

[6] S. Hong, D. Yang, J. Choi, and H. Lee, "Inferring semantic layout for hierarchical text-to-image synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7986–7994.

[7] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proceedings of the 33rd International Conference on Machine Learning*, 2016, pp. 1060–1069.

[8] Z. Zhang, Y. Xie, and L. Yang, "Photographic text-to-image synthesis with a hierarchically-nested adversarial network," *CoRR*, vol. abs/1802.09178, 2018.

[9] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D.N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5908–5916.

[10] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[11] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems*, 2016, pp. 1945–1954.

[12] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in *Proceedings of the on Thematic Workshops of ACM Multimedia*, 2017, pp. 349–357.

[13] W.L. Hao, Z. Zhang, and H. Guan, "CMCGAN: A uniform framework for cross-modal visualaudio mutual generation," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018, pp. 6886–6893.

[14] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proceedings of the 33rd International Conference on Machine Learning*, 2016, pp. 1060–1069.

[15] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1947–1962, 2019.

[16] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional

generative adversarial networks," in*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1316–1324.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[18] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 2642–2651.

[19] D. Tran, R. Ranganath, and D.M. Blei, "Deep and hierarchical implicit models," *ArXiv*, vol. abs/1702.08896, 2017.

[20] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *CoRR*, vol. abs/1802.05957, 2018.