Cascade Attention Guided Residue Learning GAN for Cross-Modal Translation

Bin Duan
b_d208@txstate.edu

Wei Wang
wei.wang@epfl.ch

Hao Tang
hao.tang@unitn.it

Hugo Latapie
hlatapie@cisco.com

Yan Yan
y_y34@txstate.edu

Abstract

Since we were babies, we intuitively develop the ability to correlate the input from different cognitive sensors such as vision, audio, and text. However, in machine learning, this cross-modal learning is a nontrivial task because different modalities have no homogeneous properties. Previous works discover that there should be bridges among different modalities. From neurology and psychology perspective, humans have the capacity to link one modality with another one, e.g., associating a picture of a bird with the only hearing of its singing and vice versa. Is it possible for machine learning algorithms to recover the scene given the audio signal?

*In this paper, we propose a novel **C**ascade **A**ttention-Guided **R**esidue GAN (CAR-GAN), aiming at reconstructing the scenes given the corresponding audio signals. Particularly, we present a residue module to mitigate the gap between different modalities progressively. Moreover, a cascade attention guided network with a novel classification loss function is designed to tackle the cross-modal learning task. Our model keeps the consistency in high-level semantic label domain and is able to balance two different modalities. The experimental results demonstrate that our model achieves the state-of-the-art cross-modal audio-visual generation on the challenging Sub-URMP dataset [18]. Code will be available at https://github.com/tuffr5/CAR-GAN.*

1. Introduction

Cross-modal learning involves multiple modalities, aims at learning knowledge from one modality to facilitate the tasks (*e.g.*, retrieval and generation) from another correlated modality. Cross-modal learning gains long-lasting interest in multimedia. Recently, with the increasing popularity of Generative Adversarial Networks (GANs) [8], cross-modal research is not only limited to retrieval [25, 27] but also makes the cross-modal generation possible, such as text-to-image [23, 35], image-to-image [14, 43], story visualization

and generation [19, 26]. Recently, radio signals [42] have also been successfully applied to human pose prediction. The radio signals which are a form of waves that are robust to occlusions so that it can predict human poses behind the wall. Another special wave is an audio signal, which has also been explored to reconstruct the scene [1, 9, 29, 21, 22].

Generating image from audio using GANs was first described by Chen *et al*. [1] where they introduce a conditional GANs (cGANs) [20] model to tackle the problem. Later, Hao *et al*. presented the Cross-Modal Cycle GAN (CMC-GAN) [9] to solve the cross-modal visual-audio mutual generation problem. Although this paper conducts an interesting exploration, we still observe unsatisfactory artifacts and missing contents in the generated images, which are due to several reasons. First, even if previous works [1, 9] showed there were truly some connections between audio and visual modalities, there is still a huge gap between different modalities, *e.g.*, the sound of the wind blowing trees and the image of shaking leaves. Thus, without prior knowledge about this scenario, it is hard to associate them together. Second, a random latent vector was employed to assist the learning process. They tried to represent the properties of the input audios accompanied by some manually defined random latent vectors. However, we argue that these latent vectors cannot represent the information of the audios accurately since they are random Gaussian noises, *i.e.*, they are not directly withdrawn from the audios. Consequently, we avoid employing random latent vectors in our designed model.

Based on the above observations, in this paper, we propose a novel **C**ascade **A**ttention-Guided **R**esidue GAN (CAR-GAN) to handle the audio-to-image translation task. The proposed CAR-GAN contains two generation stages and the overall framework of CAR-GAN is depicted in Fig. 1, in the first stage, the Log-amplitude of Mel-Spectrum (LMS) [7] image A (a representation of the raw audio) concatenated with L_a (one-hot class label code for the audio) is fed into the first self-attention guided generator (G_1) and G_1 outputs a coarse result I'. Different from previous works [1, 9], we deprecate the latent vector in our model. Note that
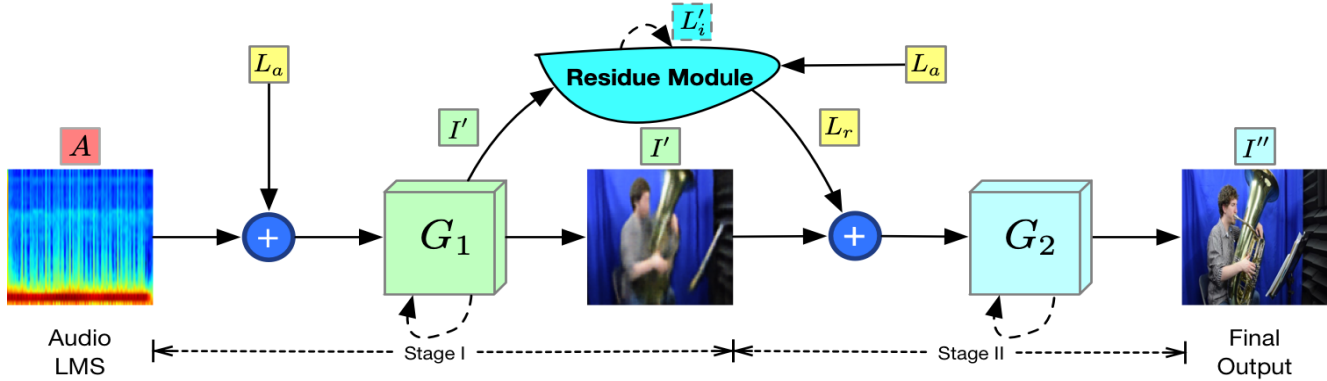
Figure 1. Overview of CAR-GAN framework. Generator G_1 takes the audio LMS (A) and its class label (L_a) as input to generate coarse image I'. Generator G_2 takes I' and the output of our specific-designed residue module (L_r) as inputs to synthesize fine-grained image I''. Note that to detect the most distinguished part in different modality spaces, we introduce self-attention mechanism to both generators. Two generators of different stages are jointly optimized in an end-to-end fashion that aims at enjoying the mutually improved benefits from different modalities, *i.e.*, audio and image. ⊕ denotes channel-wise concatenation.

to efficiently model relationships between widely separated spatial regions, we introduce self-attention to both generator and discriminator.

The coarse output I' from the first generation network is taken as input to the proposed residue module to obtain the corresponding class label vector L_i'. Then we subtract L_i' from label L_a to obtain the residue class label vector L_r. L_r reflects the difference between the generated image and the audio label in the semantic domain. Next, the coarse output I' from the first stage, together with the residue class label L_r, are input into the second stage network and generate more fine-grained final results. The intuition behind the residue module is that the second generator G_2 can flexibly preserve the similarities between audio and image space and only model the differences when it is necessary, which can be regarded as a progressive generation strategy. Finally, to optimize the proposed CAR-GAN in an end-to-end fashion, the cascaded residue classification loss is further used to generate more realistic images and preserve the consistency between two stages. It is worth noting that, the classifier is pre-trained with the real images from Sub-URMP dataset. During the training of our CAR-GAN model, the classifier parameters are fixed, and only the gradients of the synthesized images will be back propagated to guide the generator to synthesize images with correct semantic labels. In this way, the label consistency between the synthesized images and the audio labels could be preserved. Through extensive experimental evaluations, we demonstrate that CAR-GAN produces better results than the baselines such as S2IC [1] and CMCGAN [9].

Overall, the contributions of this paper are summarized as follows:

- A novel Cascade Attention-Guided Residue GAN framework (CAR-GAN) for the cross-modal audio-to-image translation task is proposed. It explores cascaded at-

tention guidance with a coarse-to-fine generation, aims at producing a more detailed synthesis from the jointly learned representation of both audio and image spaces.

- A novel residue module is presented, which is utilized to smooth the gaps between different modalities at class label space and is able to find correlations between different modalities. We also propose a new cascaded residue classification loss for more robust optimization. It not only helps the model generate more realistic images but also keeps the consistency between two stages' generation processes.

- Qualitative and quantitative results demonstrate the effectiveness of the proposed CAR-GAN on the cross-modal audio-to-image translation task, and show state-of-the-art performance on the challenging Sub-URMP dataset [18] with remarkable improvements.

2. Related work

Generative adversarial Networks (GANs) is proposed by Goodfellow *et al.* [8], complemented with adversarial method. A vanilla GAN is composed of a generator and a discriminator. The discriminator is trying to discriminate whether an image is real or fake. Conversely, the generator is to learn to output images that can fake the discriminator. Since the GANs appeared, plenty of works such as [14, 26, 38] on computer vision are based on GANs. With the success of GANs, conditional GANs encode additional information as a reference into the GAN framework which will make sure the generator can run more straightforward to the target. cGANs have achieved remarkable results in image-to-image translation [14, 43, 3], super-resolution image generation [17, 39] and style transfer [13, 2].

Image-to-Image Translation adopt input-output data to learn a translation mapping between input and output do-

mains. For instance, Isola *et al.* propose Pix2Pix [14], which is a general-purpose solution to image-to-image translation problems. For further improving the quality of the generated images, works such as [33, 40] try to employ the attention mechanism to force the generator to pay more attention to the distinguished content between different input and output domains. In this paper, we embed the proposed attention mechanism into our cross-modal GAN model, which allows the generator to effectively pay attention to the most distinguished representations between audio and image modalities. Moreover, previous works such as [31, 41] generate images using residual images which is different from ours, we employ the residual class-label to guide the generator for producing photo-realistic images. In this way, the generator only needs to focus on the high-level difference between the audio representation and image representation.

Cross-Modal Learning represents any kind of learning that involves information obtained from more than one modality. Earlier work such as [5, 36] show cross-modal perception phenomena from the perspective of neurological and psychological filed. They try to figure out the mutual relation between auditory and visual information from a neurologist's or psychologist's view. Later, cross-modal multimedia retrieval starts booming since the advance of multimedia technology. Works such as [27, 25] take advantage of cross-modal learning to help retrieving. Afterwards, cross-modal learning is popular together with generative models, *e.g.*, Variational Autoencoders (VAEs) [32] and GANs [23, 12].

Audio-to-Image Translation. There are few works to address audio-to-image translation. Existing methods [37, 1, 9] on audio-to-image translation take the advantage of latent Gaussian vector, where they design convolution neural networks encoder to extract feature map out of input audio. Later, the extracted feature maps are concatenated with the latent vector. The combined feature map is feed into the generator to produce corresponding images. The latent vector plays an important role in this translation. However, different from these existing methods, we propose replacing the latent vector with the proposed residue class-label since it contains more meaningful representations between different modalities.

3. Cascade attention guided residue learning GAN

In this section, we start with the model formulation, and then introduce the proposed objectives. Finally, we present the implementation details including network architecture and training procedure. The overall framework of the proposed CAR-GAN is illustrated in Fig. 1.

In stage one, we present a cascade attention guided generation sub-network, which utilizes both the audio signal A and its class label L_a as inputs to generate an image. The generated image I' is further fed into the proposed residue module to obtain the corresponding image class label L'_i. Next, we calculate the residual cross-modal label L_r between the audio label L_a and the image label L'_i. L_r reflects the distance between the generated images and the real images in the semantic domain.

In stage two, the coarse synthesis I' and the residual cross-modal label L_r are combined together as the input. In this way, the semantic difference between the generated images and audio signals, L_r can be employed to guide the generator to further refine the generated image I'. As a result, the cross-modality semantic distance could be further reduced after the refinement.

3.1. Stage I: Cascade attention guided generation

Class-Label Guided Generation with Self-Attention. Translating audio into image is an extremely challenging task since it is difficult to tell any relationship between audio and image modalities directly. To handle this challenge, previous works such as S2IC [1] and CMCGAN [9] tried to employ a random Gaussian noise vector as input to guide the generator to produce a synthetic image. We argue that the Gaussian noise vector will introduce some errors misguiding the generator. Different from them, we input a more accurate audio class-label into the generator similar to [3, 34]. Specifically, as shown in Fig. 1, we concatenate input LMS A from audio space with audio class-label L_a, and input them into the first generator G_1 and synthesize its corresponding coarse image I' as $I' = G_1(A, L_a)$. In this way, the audio class-label L_a provides stronger supervision to guide cross-modal translation in the deep network. Moreover, to force the generators paying more attention to the most distinguished content between different modalities, we further introduce self-attention mechanism into the generators. Zhang *et al.* [40] proposed the Self-Attention Generative Adversarial Network (SAGAN) for image generation tasks. Differently, in this paper, we propose a self-attention image-to-image translation network which allows long-range dependency modeling for cross-modal image translation task with drastic domain change. Once the generators know which part they should pay attention to, the next goal is to generate images with more fine-grained details. Therefore, we cascade two generators and train them simultaneously.

Coarse-to-Fine Cascade Generation. Due to the complexity of the cross-modal audio-to-image translation task, we observe that the first stage generator G_1 only outputs a coarse synthesis with blurred artifacts, missing content and high pixel-level dis-similarity. This thus inspires us to explore a cascade generation strategy to boost the synthesis performance from the coarse predictions. The Coarse-to-fine strategy has been used in different computer vision applications achieving promising performances, such as se-
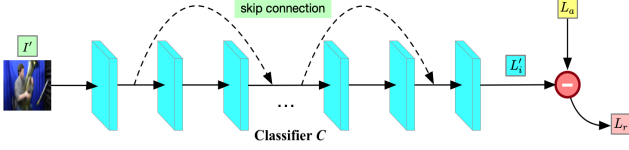
Figure 2. Residue Module: I' denotes the output of the first generator G_1. The output of the classifier C is L'_i, which is the predicted label of I'. L_a is the label of the audio, L_r denotes the difference between two different modalities, *i.e.*, audio label L_a and image label L'_i. $\odot$ denotes channel-wise subtraction.

mantic segmentation [15] and object detection [24, 4]. In this paper, we adapt the coarse-to-fine strategy to handle a more challenging audio-to-image translation task. We observe significant improvement using the proposed cascade coarse-to-fine strategy, which is illustrated in the experimental section.

3.2. Stage II: Cross-modal residue label guided generation

The overview of proposed residue cross-modal label guided generation module is shown in Fig. 2. This module consists of a pre-trained classifier to preserve the cross-modal label cycle consistency and a cross-modal residue label guided generation sub-network.

Cross-Modal Label Cycle Consistency. The coarse output I' of stage I is fed into the classifier C to generate an image classification label L'_i. To further reduce the space of possible mismatch between audio and image modalities, we hypothesize that the learned mapping functions should be cycle-consistent in cross-modal translation. For the audio class label L_a, the translation cycle should be able to bring it quite close to the image label L'_i, *i.e.*, $G_1(A, L_a) \to I' \to C(I') \to L'_i \approx L_a$. We name this as cross-modal label cycle consistency since L'_i is a label representation of the coarse output I' in the image modality and L_a is a label representation of the audio A in the audio modality. Note that the proposed cross-modal cycle consistency is different from the cycle consistency in CycleGAN [43] which adapts the cycle-consistency between the input image and the reconstruction image in the image space, in this paper, we employ making two different modalities cycle consistent in the class-label space.

Cross-Modal Residue Label Guided Generation. Previous works have shown that residual images can be effectively learned and used for image generation task. For instance, Shen and Liu [31] used the learned residual image as the difference between images before and after the face attribute manipulation. Zhao *et al.* [41] trained networks to learn residual motion between the current and future frames for the image-to-video generation task, which avoids learning motion-irrelevant details. Instead of manipulating the whole image, both approaches proposed to learn the resid-

ual images. In this way, the manipulation can be operated efficiently with modest pixel modification. However, in the paper, we propose the residue label rather than residue image for cross-modal image translation task. Specifically, we first obtain the residue label L_r between the image label L'_i and the audio label L_a by calculating $L_r = L_a - L'_i$. Then we intend to generate the missing information L_r in the second generator stage, which can be expressed as,

$$I'' = G_2(I', L_r) = G_2(G_1(A, L_a), L_r). \qquad (1)$$

In this way, the generation process can be operated efficiently with modest pixel modification, *i.e.*, the generator G_2 can flexibly preserve the similarities between the audio and image representations, and only model the differences between them.

3.3. Label consistency: Backpropagating via the classifier

A vanilla cGAN conducts backpropagation mainly determined by the discriminator. The discriminator judges from the image-level information, but not the label-level semantic information. We argue that different but corresponding modalities can match with the same semantic label. Therefore, apart from performing backpropagation from the discriminator, we also backpropagate from the label classifier to make sure the generated images belong to the same label domain with input audios. During the training, the corresponding label, *i.e.*, the instrument type of the input audio signals, is fed to the classifier. When we update the model during backpropagation, the parameters of the pre-trained classifier are fixed. Only the gradients of input images are passed back such that the images can be revised accordingly to match its semantic label. In this way, the label consistency can be guaranteed. The classifier is re-trained using a pre-trained model using ImageNet [6]. The backpropagating path is described in Fig. 3. During the backpropagation of the classifier, we design a joint loss $\mathcal{L}_C$ for our two-stage generations. $\mathcal{L}_C$ is composed of two parts: the classification loss of stage I and stage II,

$$\mathcal{L}_C = \mathcal{L}(I', I'') = \lambda_{I'}\mathcal{L}(I') + \lambda_{I''}\mathcal{L}(I''), \qquad (2)$$

where $\mathcal{L}$ is the cross-entropy loss function. $\lambda_{I'}$ and $\lambda_{I''}$ are coefficients to control the relative importance of the two objectives.

3.4. Universal optimization objective

Adversarial Loss. Our adversarial loss is composed of two parts since we adapt two stages generation. During the stage I, the adversarial loss of discriminator D for differentiating generated audio-image pairs $[A, I']$ from real audio-image pairs $[A, I]$ is formulated as following:

$$\mathcal{L}_{cGAN}(A, I') = \mathbb{E}_{A,I}\left[\log D(A, I)\right] + \\ \mathbb{E}_{A,I'}\left[\log(1 - D(A, I'))\right], \qquad (3)$$
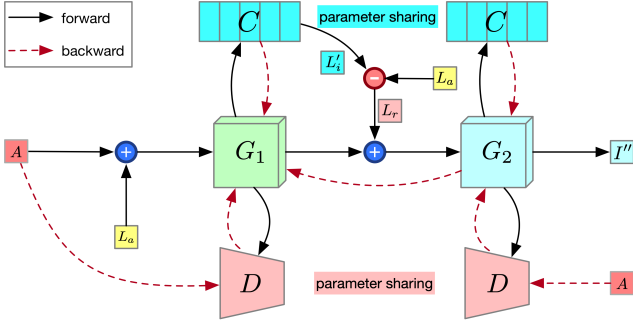
Figure 3. The backpropagation path of the proposed CAR-GAN. The solid line denotes the path for the forward process, the red dashed line denotes the backpropagation path. We use the same pretrained classifier C twice in our model, so does the discriminator D but without pretrained. I'' denotes the final generated cross-modal image. $\oplus$, $\ominus$ denote channel-wise concatenation and subtraction, respectively.

where I is the ground truth image. During the stage II, the adversarial loss of discriminator D for differentiating generated audio-image pairs $[A, I'']$ from real audio-image pairs $[A, I]$ is formulated as following:

$$\mathcal{L}_{cGAN}(A, I'') = \mathbb{E}_{A,I}\left[\log D(A, I)\right] + \mathbb{E}_{A,I''}\left[\log(1 - D(A, I''))\right]. \tag{4}$$

The above two adversarial losses both target to reduce the disagreements with ground truth images and generate more realistic synthesized images. Therefore, our adversarial loss is the total sum of the two stages:

$$\mathcal{L}_{cGAN} = \lambda_{G_1}\mathcal{L}_{cGAN}(A, I') + \lambda_{G_2}\mathcal{L}_{cGAN}(A, I'') \tag{5}$$

Universal Loss. Besides the adversarial loss $\mathcal{L}_{cGAN}$, we also introduce the classification loss $\mathcal{L}_C$ and $L1$ loss function for better optimizing our CAR-GAN. Our final loss function is a combination of the three losses.

$$\min_{\{G1,G2,C\}} \max_{\{D\}} \mathcal{L} = \lambda_c\mathcal{L}_C + \lambda_{cGAN}\mathcal{L}_{cGAN} + \lambda_{L1}\mathcal{L}_{L1}, \tag{6}$$

where $\mathcal{L}_{L1} = \mathcal{L}_{L1}(I, I') + \mathcal{L}_{L1}(I, I'')$. λ_c, λ_{cGAN} and λ_{L1} denote the trade-off parameters to control the significance of its corresponding loss function, respectively. Our model is trying to balance the min-max problem while training.

3.5. Implementation details

Network Architecture. Inspired by the work of Isola *et al.* [14], we employ U-Net [28] as the backbone of our generators G_1 and G_2. U-Net is a Convolution Neural Network (CNN) architecture with skip connections between a down-sampling encoder and an up-sampling decoder, and it retains complex textual information of the input. We share the same network architecture in both generators. The convolutions of down-sampling layers and up-sampling layers are

4×4 kernel with stride 2 and padding 1. The filters in attention convolution layers are 1×1 with stride 1. For the discriminator D, we adapt PatchGAN as in [14, 43]. The kernel size of the attention convolution layers is also 1×1. Both the generators and discriminator have attention layers before the last two convolution layers. The other convolution layers in the discriminator have a kernel size of 4×4 kernel with stride 2 and padding 1. Batch normalization is used in our model. As for the classifier, we employ ResNet50 [10] architecture which is pre-trained in the ImageNet. Then we add a fully connected layer at the end of the network and conducted transfer learning in Sub-URMP dataset for high classification accuracy. The classifier is fixed while training our model.

Training Details. First, we employ preprocessing for every audio, where the audio are converted from waveform pattern to LMS pattern, which is a frequency warping pattern that allows for better representation of audio clips. Our proposed CAR-GAN is trained and optimized in an end-to-end style. C is pre-trained and fixed while training. We first train G_1, G_2 with D fixed, and update parameters of G_1 and G_2 by the sum of gradients from C's and D's backpropagation, and then we train D with G_1 and G_2 fixed, but the backpropagation of C has no influence on the optimization of D, *i.e.*, the optimization is only determined by G_1 and G_2. We apply Adam algorithm [16] for optimizing both the generators (G_1, G_2) and discriminator (D) jointly. The betas of the Adam algorithm set to 0.9 and 0.999, respectively. Weights are initialized from a Gaussian distribution with standard deviation 0.2 and mean 0.

4. Experiments

4.1. Experimental settings

Datasets. Following [1], we adapt the widely used Sub-URMP (University of Rochester Musical Performance) dataset [18] to evaluate the proposed model. This dataset consists of 17,555 pairs of audios and images, and has 13 kinds of instruments played by different people. It maps an image to a half second long audio, and the image is the first frame of the half second long audio.

Parameter Settings. We resize images to 256×256 resolution as inputs. We implement with Pytorch and the experiments are running at 4 Nvidia GeForce GTX 1080 Ti GPUs with batchsize 64. We set the learning rate to 0.0008, and stop our training at the epoch of 200. Both $\lambda_{I'}$ and $\lambda_{I''}$ in Eq. (2) are set to 0.5. We set λ_{G_1}, λ_{G_2} in Eq. (5), and λ_c, λ_{cGAN} in Eq. (6) all equal to 1, and λ_{L1} equal to 100 in Eq. (6).

Evaluation Metrics. To compare with previous work [1, 9], we employ the classification accuracy as the metric, the only metric used in the previous two papers. The way we measure our model is that we first train a model
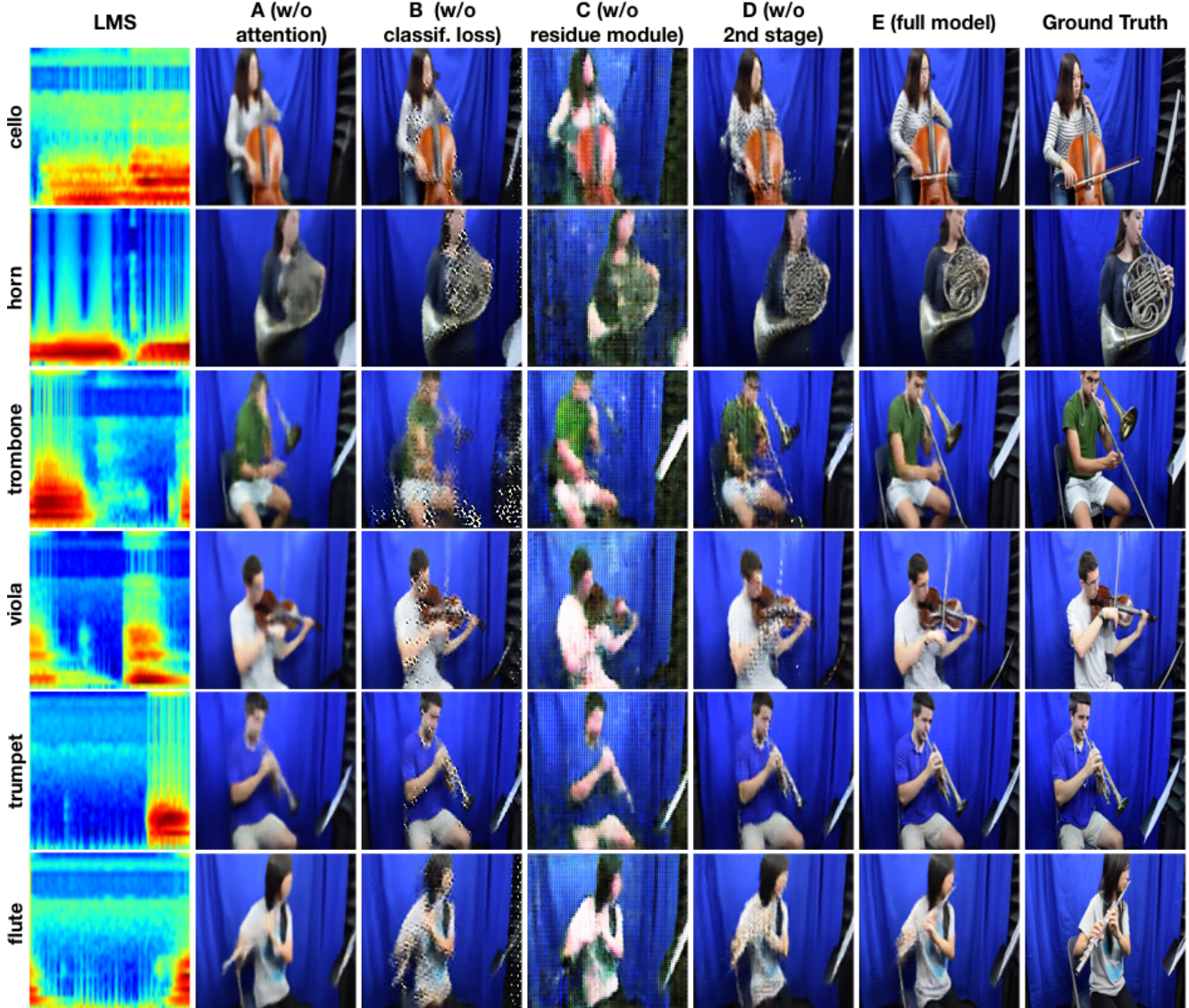
Figure 4. Ablation study: synthesized images by different models of the proposed CAR-GAN. LMS represents the LMS of the input audios.

with 99.56% classification accuracy using ResNet50 trained on Sub-URMP dataset, and then we test the pre-trained ResNet50 on our generated results. The intuition behind this is that if the generated images are realistic, the classifier trained on the real images will also achieve decent accuracy on the generated images during the testing stage. In addition to classification accuracy, we also employ Fréchet Inception Distance (FID) [11] and Inception Score (IS) [30] metrics to further evaluate the fidelity of the generated images. Due to the lack of pretrained model and released code of previous work, we are not able to get FID and IS of their works.

4.2. Experimental results

Settings of Ablation Study. We perform ablation study on the proposed CAR-GAN. We break down our model and assemble it into five different models. The following five models share a similar backbone, but a particular part is abandoned. Model A avoids using attention guided generation. Model B drops out the $\mathcal{L}_c$ loss. Residue module is taken out from model C. Model D is running without stage II (no second generator). And the last model E is our fully proposed model. Fig. 4 shows the corresponding cross-modal generated images of different models. Table 1 depicts how models perform based on the metrics we employed.

Influence of Attention Mechanism. Compared with our

Baseline	A	B	C	D	E	GT
FID	279.3022	332.1574	380.1104	307.3725	**240.7771**	-
IS	3.2221	3.4337	2.0699	3.7215	**4.2023**	4.7552

Table 1. Results of the proposed CAR-GAN for FID and IS metrics.

Figure 5. Generated images of different stages of our model.

Method	Accuracy	
	Training	Testing
S2IC	0.8737	0.7556
CMCGAN	0.9105	0.7661
Ours	**0.9954**	**0.9033**

Table 2. The classification accuracy of different methods.

proposed full model E, model A which avoids attention mechanism performs slightly worse. The results of model A shows the overall contour of images, but some details are left out. That is, attention mechanism does enhance the representation ability of our model.

Influence of Classification Loss. If we add our proposed classification loss into the model, we make an improvement by 27.51% in FID and 22.39% in IS since the classification loss $\mathcal{L}_c$ awards the generators strong guidance towards the
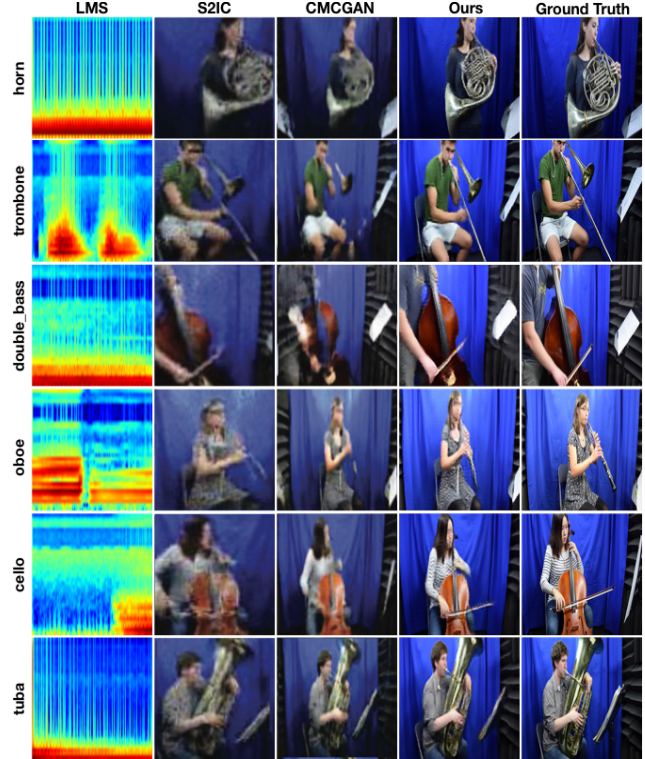

Figure 6. Synthesized images of different methods on the Sub-URMP dataset.

ground-truth. This tells generators from an overall classification view. Therefore, the generators know the appropriate direction to go from the label domain.

Influence of Residue Module. With our residue module, we achieve an amazing improvement on FID by 36.66% and on IS by 103.01%. Our residue module supplements the missing information during the generation of Stage I, making that the generated images belong to the same domain and keep balance in the label domain between inputs and outputs.

Influence of Two-Stage Generation. Our two stages of generators lead to the improvement of 21.67% in FID and 13.01% in IS. The single generator has weak representation ability for a complicated cross-modal generation. Thus, the union of two or more generators can progressively improve representation ability and performance. To visualize the influence of two-stage generation, we display synthesized images by different stages in Fig. 5. The generated images by the second stage have more fine-grained details and are more realistic, i.e., it certifies that two-stage generation is

beneficial for the whole generation procedure.

State-of-the-art Comparisons. We show the quantitative and qualitative results with comparison methods in Table 2 and Fig. 6, respectively. We make an improvement in training accuracy by 13.93%, 9.32% compared to S2IC [1] and CMCGAN [9], respectively. As for the testing, we achieve an increase of 19.55% and 17.91% compared with S2IC and CMCGAN. Furthermore, we produce more realistic and detailed images compared with the previous methods as illustrated in Fig. 6.

Failure Cases and Analysis. During our experiments, we find there are some synthesized images which are randomly combined by learned features. When we start furtherly looking into these cases, we finally find out that the corresponding inputs of these failure cases are more like noises that are randomly distributed, and the GT images show people mostly hold the instrument still and wait, that is, there is no sound making by instruments, the audios mainly consist of background and other noises.

5. Conclusion

In this paper, we design a novel Cascade Attention Guided Residue Learning GAN (CAR-GAN) to solve the challenging cross-modal audio-to-image translation task. Particularly, it employs cascaded attention guidance and a coarse-to-fine generation strategy. A novel residue learning model is also proposed to tackle the cross-modal class-label dis-match problem between audio and image modality. By introducing the residue module, generators learn to produce residue feature between two stages, which pushes the output closer to its corresponding real image in a high-level semantic space. Finally, the proposed joint classification loss facilitates the model generation and keeps consistency in the label domain. Experimental results show the state-of-the-art performance on the cross-modal translation task.

References

[1] L. Chen, S. Srivastava, Z. Duan, and C. Xu. Deep cross-modal audio-visual generation. In *ACM MM Workshop*, 2017.

[2] X. Chen, C. Xu, X. Yang, L. Song, and D. Tao. Gated-gan: Adversarial gated networks for multi-collection style transfer. *TIP*, 28(2):546–560, 2019.

[3] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018.

[4] D. Cristinacce, T. F. Cootes, and I. M. Scott. A multi-stage approach to facial feature detection. In *BMVC*, 2004.

[5] R. K. Davenport, C. M. Rogers, and I. S. Russell. Cross modal perception in apes. *Neuropsychologia*, 11(1):21–28, 1973.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[7] L. Deng, J. Droppo, and A. Acero. Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise. *IEEE Transactions on Speech and Audio Processing*, 12(2):133–143, 2004.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

[9] W. Hao, Z. Zhang, and H. Guan. Cmcgan: A uniform framework for cross-modal visual-audio mutual generation. In *AAAI*, 2018.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[11] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017.

[12] C. Hu, D. Li, Y.-Z. Song, and T. M. Hospedales. Now you see me: Deep face hallucination for unviewed sketches. In *BMVC*, 2016.

[13] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[15] N. Khan, B.-W. Hong, A. Yezzi, and G. Sundaramoorthi. Coarse-to-fine segmentation with shape-tailored continuum scale spaces. In *CVPR*, 2017.

[16] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[17] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.

[18] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma. Creating a multi-track classical musical performance dataset for multimodal music analysis: Challenges, insights, and applications. *TMM*, pages 522–535, 2016.

[19] Y. Li, Z. Gan, Y. Shen, J. Liu, Y. Cheng, Y. Wu, L. Carin, D. Carlson, and J. Gao. Storygan: A sequential conditional gan for story visualization. *arXiv:1812.02784*, 2018.

[20] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv:1411.1784*, 2014.

[21] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *ICML*, 2011.

[22] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik. Speech2face: Learning the face behind a voice. In *CVPR*, 2019.

[23] H. Park, Y. Yoo, and N. Kwak. Mc-gan: Multi-conditional generative adversarial network for image synthesis. In *BMVC*, 2018.

[24] M. Pedersoli, A. Vedaldi, J. Gonzalez, and X. Roca. A coarse-to-fine approach for fast deformable object detection. *PR*, 48(5):1844–1853, 2015.

[25] J. C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, and N. Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *TPAMI*, 36(3):521–535, 2014.

[26] T. Qiao, J. Zhang, D. Xu, and D. Tao. Mirrorgan: Learning text-to-image generation by redescription. In *CVPR*, 2019.

[27] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM MM*, 2010.

[28] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[29] A. Rouditchenko, H. Zhao, C. Gan, J. McDermott, and A. Torralba. Self-supervised audio-visual co-segmentation. In *ICASSP*, 2019.

[30] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *NIPS*, 2016.

[31] W. Shen and R. Liu. Learning residual images for face attribute manipulation. In *CVPR*, 2017.

[32] A. Spurr, J. Song, S. Park, and O. Hilliges. Cross-modal deep variational hand pose estimation. In *CVPR*, 2018.

[33] H. Tang, D. Xu, N. Sebe, Y. Wang, J. J. Corso, and Y. Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *CVPR*, 2019.

[34] H. Tang, D. Xu, W. Wang, Y. Yan, and N. Sebe. Dual generator generative adversarial networks for multi-domain image-to-image translation. In *ACCV*, 2018.

[35] Y. Verma and C. V. Jawahar. Im2text and text2im: Associating images and texts for cross-modal retrieval. In *BMVC*, 2014.

[36] J. Vroomen and B. d. Gelder. Sound enhances visual perception: cross-modal effects of auditory organization on vision. *Journal of experimental psychology: Human perception and performance*, 26(5):1583, 2000.

[37] C.-H. Wan, S.-P. Chuang, and H.-Y. Lee. Towards audio to scene image synthesis using generative adversarial network. In *ICASSP*, 2019.

[38] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. In *NeurIPS*, 2018.

[39] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018.

[40] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. In *ICML*, 2019.

[41] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. Metaxas. Learning to forecast and refine residual motion for image-to-video generation. In *ECCV*, 2018.

[42] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi. Through-wall human pose estimation using radio signals. In *CVPR*, 2018.

[43] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.