The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

CMCGAN: A Uniform Framework for Cross-Modal Visual-Audio Mutual Generation

Wangli Hao,[1,4] **Zhaoxiang Zhang,**[1,2,3,4,*] **He Guan**[1,4]
[1]Research Center for Brain-inspired Intelligence, CASIA
[2]National Laboratory of Pattern Recognition, CASIA
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, CAS
[4]University of Chinese Academy of Sciences
{haowangli2015,zhaoxiang.zhang,guanhe2015}@ia.ac.cn

Abstract

Visual and audio modalities are two symbiotic modalities underlying videos, which contain both common and complementary information. If they can be mined and fused sufficiently, performances of related video tasks can be significantly enhanced. However, due to the environmental interference or sensor fault, sometimes, only one modality exists while the other is abandoned or missing. By recovering the missing modality from the existing one based on the common information shared between them and the prior information of the specific modality, great bonus will be gained for various vision tasks. In this paper, we propose a Cross-Modal Cycle Generative Adversarial Network (CMCGAN) to handle cross-modal visual-audio mutual generation. Specifically, CMCGAN is composed of four kinds of subnetworks: audio-to-visual, visual-to-audio, audio-to-audio and visual-to-visual subnetworks respectively, which are organized in a cycle architecture. CMCGAN has several remarkable advantages. Firstly, CMCGAN unifies visual-audio mutual generation into a common framework by a joint corresponding adversarial loss. Secondly, through introducing a latent vector with Gaussian distribution, CMCGAN can handle dimension and structure asymmetry over visual and audio modalities effectively. Thirdly, CMCGAN can be trained end-to-end to achieve better convenience. Benefiting from CMCGAN, we develop a dynamic multimodal classification network to handle the modality missing problem. Abundant experiments have been conducted and validate that CMCGAN obtains the state-of-the-art cross-modal visual-audio generation results. Furthermore, it is shown that the generated modality achieves comparable effects with those of original modality, which demonstrates the effectiveness and advantages of our proposed method.

Video mainly contains two symbiotic modalities, the visual and the audio ones. Information embedded in these two modalities owns both common and complementary information respectively. Common information can make the translation over visual and audio modalities be possible. Meanwhile, complementary information can be adopted as the prior of one modality to facilitate the associative tasks. Thus, sufficient utilization of these common and complementary information will further boost the performances of related video tasks. However, due to the environmental disturbance and sensor fault, one of the modality may be missing or damaged, which will bring significant inconveniences such as silent film and screen blurred. If we can restore the missing modality from the remaining modality based on the cross-modal prior, great bonus will be gained for various multimedia tasks and many traditional single-modal databases can be reused in conjunction to gain better performance.

Generative Adversarial Networks (GANs) have gained extraordinary popularity because of their ability in generating high-quality realistic samples, which is superior to other generative models. Compared to numerous work focusing on static information translation, such as image-to-image (Isola et al. 2016; Zhu et al. 2017) and text-to-image (Reed et al. 2016), few of methods concern dynamic visual-audio modality conversion and generation. Chen et al. firstly design Conditional GANs for cross-modal visual-audio generation. Drawbacks of their work are that the mutual generation process relies on different models and it cannot be trained end-to-end.

Inspired by (Isola et al. 2016; Zhu et al. 2017), we propose Cross-Modal Cycle Generative Adversarial Network (CMCGAN) to achieve cross-modal visual-audio mutual generation. Compared to CycleGAN, CMCGAN introduces a latent vector to handle dimension and structure asymmetry among different modalities. Moreover, another two generation paths are coupled with CycleGAN to facilitate cross-modal visual-audio translation and generation. Finally, a joint corresponding adversarial loss is designed to unify the visual-audio mutual generation in a common framework. In addition, CMCGAN can be trained end-to-end to obtain better convenience.

Benefiting from CMCGAN, a dynamic multimodal classification network is developed for double modalities. Once only single modal as input, we will supplement the absent one in the aid of CMCGAN and then perform the subsequent classification task. In summary, we make the following contributions:

- We propose a Cross-Modal Cycle Generative Adversarial Network(CMCGAN) to simultaneously handle cross-modal visual-audio mutual generation in the same model.
- We develop a joint adversarial loss to unify visual-audio mutual generation, which makes it possible not only to

[*]Corresponding author. (zhaoxiang.zhang@ia.ac.cn)

distinguish training data from generated or sampling but also to check whether image and sound pairs matching or not.

- We develop a multimodal classification network for different modalities with dynamic loading.

Related Works

Our work closely draws on recent studies in generative adversarial network (GAN), cross-domain translation and cross-modal transfer.

Generative Adversarial Network

Generative Adversarial Network (GAN) (Goodfellow et al. 2014), has a wide applications and can be utilized to generate ”unseen” and fancy samples. To obtain better synthetic results, numerous models have been proposed to improve the performance of GAN. For example, Conditional GAN (Mirza and Osindero 2014), Deep Convolutional GAN (Radford, Metz, and Chintala 2015),Wasserstein GAN (Arjovsky, Chintala, and Bottou 2017) and CycleGAN (Zhu et al. 2017).

In this paper, we establish our cross-modal visual-audio mutual generation model based on CycleGAN. Different from CycleGAN, our model is performing cross-modal generation other than image-to-image generation.

Cross-Domain Translation

Cross-domain translation refers to exploring mapping relationship between two different domains. Based on conditional GAN, Isola et al. develop a ”pix2pix” framework to learn a mapping between paired input and output images (Isola et al. 2016). Shared with similar ideas, other various tasks have been established. For example, generating images of outdoor scenes from attributes and semantic layouts (Karacan et al. 2016) and generating photographs from sketches (Sangkloy et al. 2016). Moreover, Zhu et al. develop a CycleGAN (Zhu et al. 2017) to build mapping relationship across different image domains. Further, based on CycleGAN, Lu et al. (Lu, Tai, and Tang 2017) propose a Conditional CycleGAN, which is utilized to perform image-to-image translation subjected to specific attribute condition.

Common properties of the above cross-domain translation frameworks are that the domains they performing translations are from the same modality and share the similar dimension and structure. These methods can not be applied to cross-modal generation effectively, which is because that samples from different modalities are dimension and structure asymmetric.

If we want to take advantages of the promising cross-domain translation capacity of the CycleGAN (Lu, Tai, and Tang 2017), dimension and structure asymmetry across two different modalities need to be handled. In our work, a latent vector is introduced into CycleGAN to figure out this problem.

Cross-Modal Transfer

Recently, various cross-modal transfer tasks have been developed. In (Owens et al. 2016), sound is utilized as a supervisor to guide visual learning. While in (Aytar, Vondrick, and Torralba 2016), Yusuf Aytar et al. adopt a visual network as a teacher. Knowledge learned from this teacher network can be transferred to a student sound network. In (Feng, Wang, and Li 2014; Pereira et al. 2014; Rasiwasia et al. 2010; Wang et al. 2016), people focus on cross-modal indexing and retrieval. Although these methods attempt to build a joint representation and correlation over cross-modalities’ data, samples retrieved via these methods are in the database. They cannot deal with the ”unseen” samples effectively.

Taking inspiration from generating images from text captions (Reed et al. 2016), Lele Chen et al. first design Conditional GANs for cross-modal visual-audio generation. Specifically, they develop two separate networks, such as I2S (image to sound) and S2I (sound to image) to perform visual→audio and audio→visual generation (Chen et al. 2017) respectively. Although images/sounds can be generated from sounds/images by S2I/I2S, their whole cross-modal visual-audio generation framework suffers from several problems. For example, their visual-audio mutual generation are realized by two separate networks, which is inefficient. Further, each cross-modal generation path cannot be trained end-to-end. Specifically, a classification network is first trained to obtain discriminant information from one modality. Then, based on the extracted discriminant features, another modality is generated via I2S or S2I.

To overcome shortcomings of the conventional cross-modal visual-audio generation model (Chen et al. 2017), we build a Cross-Modal Cycle Generative Adversarial Model (CMCGAN) to perform cross-modal visual-audio generation. Our model unifies visual-to-audio and audio-to-visual into a common framework by a joint corresponding adversarial loss. Moreover, CMCGAN can be trained end-to-end.

Cross-Modal Cycle Generative Adversarial Network

In this section, we depict our CMCGAN in detail. The overall diagram of CMCGAN is presented in Figure 1. CMCGAN is composed of two groups of generation paths. One group of generation paths are from one modality to the same modality (blue arrow streams), including visual-audio-visual and audio-visual-audio. The other group of generation paths are from one modality to the other modality (red arrow streams), including visual-visual-audio and audio-audio-visual. Components under the same kind of rectangles are sharing weights.

Four Subnetworks and Discriminator Network

Audio-to-Visual (A2V) subnetwork: The A2V subnetwork is dubbed as: $G_{A\rightarrow V}$. The raw audio wave is first transferred to its Log-amplitude of Mel-Spectrum (LMS) representation with size 128×44 and resized to 128×32. LMS of audio sample is then passed through a sound encoder (EncoderA) with some continuous convolutional layers to obtain a feature map F_A. F_A is concatenated with a latent vector Z to obtain the embedding vector E_A. Finally, E_A is passed through an image decoder (DecoderV) with

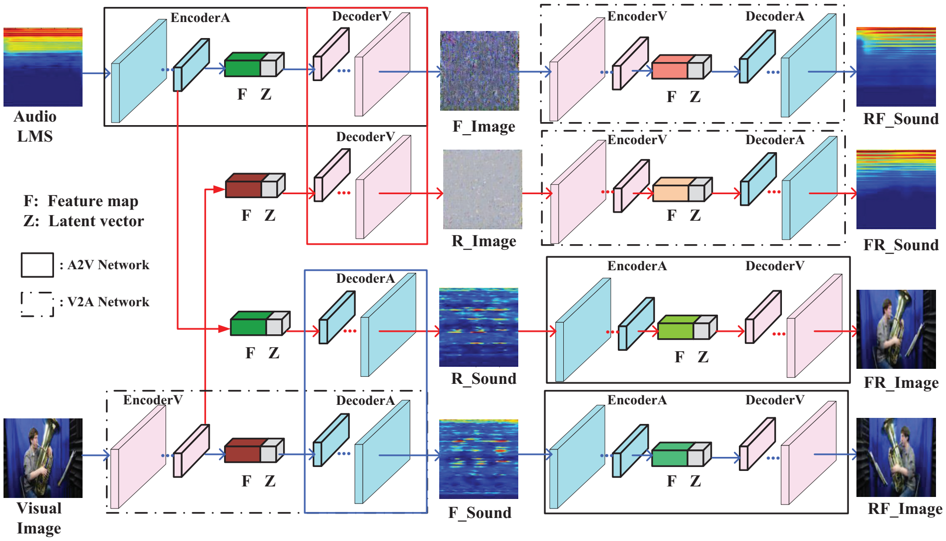

Figure 1: The overall framework of CMCGAN. Components under the same kind of rectangles are sharing weights. F indicates feature map of the corresponding samples, Z denotes the Latent vector. Blue arrow streams indicate the generation paths visual-audio-visual and audio-visual-audio. Red arrow streams denote the generation paths visual-visual-audio and audio-audio-visual. F_Image/F_Sound indicates the generated image/sound from original sound/image. R_Image/R_Sound denotes the recovered image/sound from original image/sound. FR_Image/FR_Sound indicates the generated image/sound from R_Sound/R_Image. RF_Image/RF_Sound denotes the recovered image/sound from F_Sound/R_Image.

several continuous deconvolutional layers to generate a synthetic image with size 128 × 128 × 3. Specifically, size of the input sound LMS is 128 × 32 × 1 and size of the output generated image is 128 × 128 × 3.

Visual-to-Audio (V2A) subnetwork: The V2A subnetwork is dubbed as: $G_{V\rightarrow A}$. Organization of this subnetwork is similar with that of A2V subnetwork, which contains an image encoder (EncoderV) and a sound decoder (DecoderA). This subnetwork takes an image as input and outputs a sound LMS.

Audio-to-Audio (A2A) subnetwork: The A2A subnetwork is dubbed as: $G_{A\rightarrow A}$. Organization of this subnetwork is similar with that of A2V subnetwork, which contains a sound encoder (EncoderA) and a sound decoder (DecoderA). This subnetwork takes a sound LMS as input and outputs a sound LMS.

Visual-to-Visual (V2V) subnetwork: The V2V subnetwork is dubbed as: $G_{V\rightarrow V}$. Organization of this subnetwork is similar with that of A2V subnetwork, which contains an image encoder (EncoderV) and an image decoder (DecoderV). This subnetwork takes an image as input and outputs an image.

Four Generation paths: The generation path visual-audio-visual is denoted as $G_{V\rightarrow A\rightarrow V}$, which is formed by concatenating $G_{V\rightarrow A}$ and $G_{A\rightarrow V}$ subnetworks sequentially. $G_{A\rightarrow V\rightarrow A}$, $G_{V\rightarrow V\rightarrow A}$ and $G_{A\rightarrow A\rightarrow V}$ share the similar meaning.

Discriminator: The discriminator network is depicted as: $\mathbb{R}^{|\phi_D(a)|} \times \mathbb{R}^{|\varphi_D(x)|} \mapsto [0,1]$. An image x and a sound LMS a are taken as input. They are separately passed through several continuous convolutional layers to get corresponding encoded feature maps E_{DV} and E_{DA} respectively. E_{DV} and E_{DA} are then concatenated together to produce a scalar probability s. s is adopted to judge whether this pair of image and sound is real or not. Where ϕ_D and φ_D are the encoding functions of audio and image samples respectively.

Network Architectures Image/sound encoder has seven continuous convolutional layers. Each layer is followed by a batch normalization layer (BN) and a Relu layer. The numbers of kernels for all convolutional layers in image/sound encoder are 3/1-64-128-256-512-512-256-64 respectively. Image/sound decoder has seven continuous deconvolutional layers. Each deconvolutional layer is followed by a batch normalization layer (BN) and a Leaky Relu layer. The numbers of kernels for all deconvolutional layers in image/sound decoder are 256-512-512-256-128-64-3/1 accordingly. In addition, the image/sound classifier has the similar architecture with that of image/sound encoder, except a fully connected layer is attached in the final. The detailed structure of image/sound classifier is presented as 3/1-64-128-256-512-512-256-64-fc(13) and the fc indicates the fully connected layer. The kernel size and the stride for each convolutional or deconvolutional layer is 5 and 2 respectively.

Joint Corresponding Adversarial Loss To unify visual-audio mutual generation into a common framework, we develop a joint corresponding adversarial strategy and our discriminator aims at maximizing the following loss term $\mathcal{L}_D$:

$$\begin{aligned}
\mathcal{L}_{Dw} &= log(S_{rvra}) + \frac{1}{2} * (log(1 - S_{rvwa}) + log(1 - S_{wvra})) \\
\mathcal{L}_{Df} &= log(S_{rvra}) + \frac{1}{2} * (log(1 - S_{rvfa}) + log(1 - S_{fvra})) \\
\mathcal{L}_D &= \mathcal{L}_{Dw} + \mathcal{L}_{Df}
\end{aligned} \tag{1}$$

where $\mathcal{L}_{Dw}$ is utilized to justify whether the image-sound pair is from the same instrument category or not, $\mathcal{L}_{Df}$ is adopted to justify whether the image-sound pair is sampled from generated or real sets. Specifically, S_{rvra} is the scalar probability for the true pair of image and sound, S_{rvwa} denotes the scalar probability for the pair of real image and wrong sound that sampled from wrong category of instruments, S_{rvfa} indicates the scalar probability for the pair of real image and fake sound generated by real image or real sound. S_{wvra}/S_{fvra} shares the similar meaning with S_{rvwa}/S_{rvfa} respectively.

Meanwhile, the loss of our generator is formulated as:

$$\mathcal{L}_G = log(S_{rvfa}) + log(S_{fvra}) \tag{2}$$

Consistency Loss We also restrict our model to the following consistency loss, expecting to generate plausible images and sound LMS.

$$\begin{aligned}
\mathcal{L}_{Cons} = {} & \mathcal{L}_{l1}(G_{A\to V\to A}, GT\_s) + \mathcal{L}_{l1}(G_{V\to A\to V}, GT\_i) \\
& + \mathcal{L}_{l1}(G_{A\to A\to V}, GT\_i) + \mathcal{L}_{l1}(G_{V\to V\to A}, GT\_s)
\end{aligned} \tag{3}$$

where $\mathcal{L}_{l1}$ indicates the l_1 loss, subscripts of G denote the generation paths from corresponding input to output, GT_s and GT_i represent the ground truth sound and image samples respectively.

Our cross-modal visual-audio generation training algorithm is presented in Algorithm 1.

Experiments

Dataset and Implementation Details

To validate the performance of CMCGAN for cross-modal visual-audio mutual generation, Sub-URMP (University of Rochester Musical Performance) dataset (Li et al. 2016; Chen et al. 2017) is adopted. This dataset contains 13 music instrument categories. For each category, different music pieces are played by 1 to 5 persons. In detail, there are total 17,555 sound-image pairs in Sub-URMP dataset.

Network parameters are learned by SGD algorithm for discriminators and Adam for generators. The batch size is set to 64 and momentum as 0.9. The learning rate in our experiments is 0.001. We stop our training procedure at 200 epochs. The size of Gaussian latent vector is 100.

Performance Evaluations

In this section, several experiments are designed to evaluate the performance of our model CMCGAN.

Algorithm 1 CMCGAN training algorithm for cross-modal visual-audio generation.

```
Input: minibatch images $x$, minibatch sounds LMS $a$,
    minibatch images $\hat{x}$ that mismatched with $x$, minibatch
    sounds LMS $\hat{a}$ that mismatched with $a$, latent vector z,
    number of training batch steps $S$, number of generator
    loss training steps $K$.
for each $i \in [1, S]$ do
    Sample a minibatch image $x$ and sound $a$
    Forward $x$, $a$ and $z$ through network
    Sample a minibatch mismatched image $\hat{x}$ and sound $\hat{a}$
    Forward (image, sound), (generated image, sound),
        (wrong image, sound), (image, generated sound) and
        (image, wrong sound) pairs through discriminator
        separately
    Compute discriminator loss $\mathcal{L}_D$ (Equation 1)
    Update $D$
    for each $j \in [1, K]$ do
        Compute generator loss $\mathcal{L}_G$ (Equation 2)
        Compute consistency loss $\mathcal{L}_{Cons}$ (Equation 3)
        Update $G$
    end for
end for
```

Evaluate the Performances of CMCGAN and Conventional Cross-Modal Visual-Audio Generation Models To validate the superiority of our model CMCGAN, recent benchmark models S2IC and I2S (Chen et al. 2017) are utilized as comparisons. S2IC/I2S indicates generating image/sound from the corresponding sound/image separately. Comparison results are presented in Figure 2.

Figure 2 reveals that CMCGAN can obtain better cross-modal generated images and sounds when compared to S2IC and I2S models respectively. Specifically, images generated by our model have less noise and sounds LMS are more similar to the ground-truth ones.

To further validate the performance of our model, we also evaluate the classification accuracies of generated images/sounds of CMCGAN and S2IC/I2S (Chen et al. 2017) respectively. Comparison results are displayed in Table 1.

From Table 1, we can see that CMCGAN exceeds S2IC/I2S in a large extent, which further verifies the superiority of our model.

It is interesting to give manual judgments. We collect observers from our lab (21 people) to see the satisfactory and acceptable rate of generated modalities. The metrics are: image/sound average satisfactory rate (I-AST/S-AST) and image/sound average acceptable rate (I-AAT/S-AAT). We compare our model and the existing model as: I-AST/S-AST: 0.714/0.619 vs. 0.429/0.190, I-AAT/S-AAT: 0.952/0.857 vs. 0.619/0.381 respectively. It further illustrates the advantages of our approach.

Evaluate the Performances of Models with or without Attached Cross-Modal Generation Path Besides the generation paths which have the same input and output modality (visual-audio-visual and audio-visual-audio), two

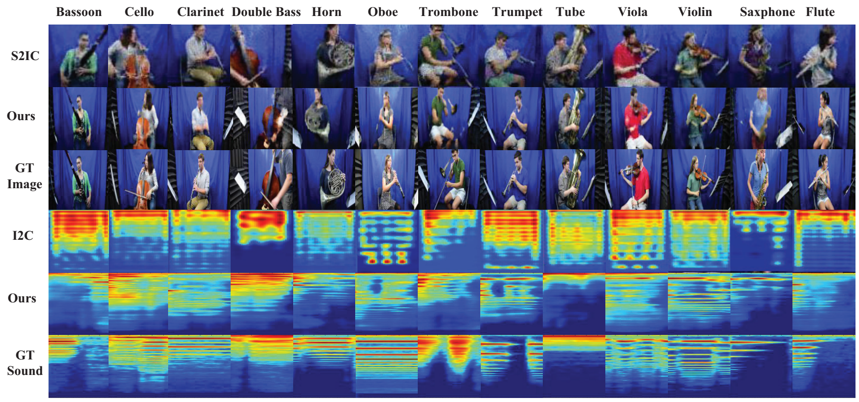

Figure 2: Cross-modal generated images and sounds based on different models. GT indicates ground truth.

Models (Sound-to-Image)	S2IC	CMCGAN
Training Accuracy	0.8737	**0.9105**
Testing Accuracy	0.7556	**0.7661**
Models (Image-to-Sound)	I2S	CMCGAN
Training Accuracy	-	0.8109
Testing Accuracy	0.1117	**0.5189**

Table 1: The classification accuracies for generated images and sounds based on different models.

extra paths which have the different input and output modalities (visual-visual-audio and audio-audio-visual), are attached in our model CMCGAN to perform cross-modal visual-audio mutual generation. To validate the effectiveness of our attached cross-modal generation paths, we compare CMCGAN with model LCGAN. LCGAN shares the similar structure with CMCGAN but has no cross-modal generation paths. Among them, audio-visual-audio indicates first generating the fake image (F_Image) from the original sound, then recovering the sound (RF_Sound) from the fake image. visual-audio-visual has the similar meaning. Comparison results are shown in Figure 3.

Figure 3 reveals that fake sound and image (F_Sound and F_Image) generated by LCGAN perform worst, which validates that directly transferring from one modality to the other is inappropriate. On the other hand, recovered image and sound (RF_Image and RF_Sound) from fake sound and image by LCGAN perform best, which may be because that the intermediate samples F_Sound and F_Image carry information from original image and sound in some extent. Moreover, our model achieves comparable cross-modal generated results with RF_Image and RF_Sound from LCGAN respectively, which validates the effectiveness of our attached cross-modal generation path.

Evaluate the Performances of Models with or without Latent Vectors in SubNetworks' Encoder To validate the effects of latent vectors in handling dimension and structure asymmetry across visual and audio modalities, we compare CMCGAN with NLCMCGAN. NLCMCGAN shares the same structure with CMCGAN, expect that latent vectors are not integrated into its subnetworks.

Comparison results are displayed in Figure 4. Specifically, CMCGAN-i indicates encoders/decoders of subnetworks in CMCGAN have i convolutional/deconvolutional layers. NLCMCGAN-i shares the similar meaning with CMCGAN-i.

Figure 4 demonstrates the generated images of CMCGAN contain more plausible pixel distributions and the generated sounds LMS of CMCGAN are more similar with the ground truth ones when compared to those of NLCMCGAN. Moreover, the more convolutional and deconvolutional layers the CMCGAN subnetworks have, the more reasonable pixel distributions for the generated images and the more similar sound LMS for the generated sound LMS respectively.

These results validate that latent vector is effective for handling dimension and structure asymmetry across different modalities. Further, the more abstract the features are extracted, the smaller gaps the two symbiotic modalities suffer.

Evaluate the Performances of Models with Different Adversarial Losses Our joint corresponding adversarial loss is adopted to optimize image/sound and sound/image matching in addition to the image and sound realism. We believe that this joint corresponding adversarial loss may introduce additional information to the corresponding generators, which will further enhance the qualities of generated image and sound. To verify the effectiveness of our joint corresponding adversarial loss function, another model with standard adversarial loss is applied as a comparison. Standard adversarial loss refers to optimize the image and sound realism. Comparison results are shown in Figure 5. Due to space constraint, we only present the comparison results of the generated images.

Figure 5 demonstrates that images generated by model with standard adversarial loss seem suffer more severe aliasing effects, while our model with joint corresponding adversarial loss obtains better generated images.

Dynamic Multimodal Classification Network

Model Benefiting from cross-modal visual-audio generation, we develop a dynamic multimodal classification net-

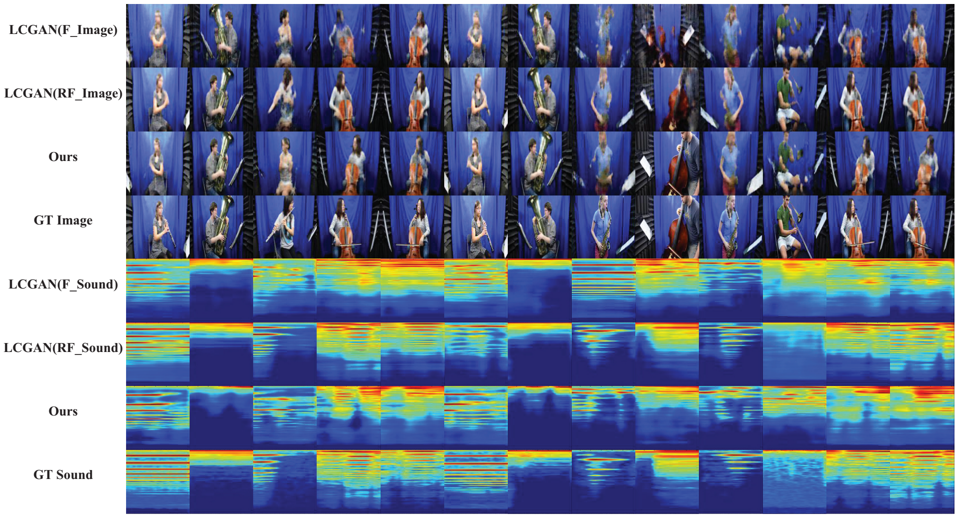

Figure 3: Cross-modal generated images and sounds based on various generation paths. GT means ground truth. F_Image is the generated image from sound. RF_sound is the recovered sound from F_Image. F_Sound and RF_Image have similar meanings.

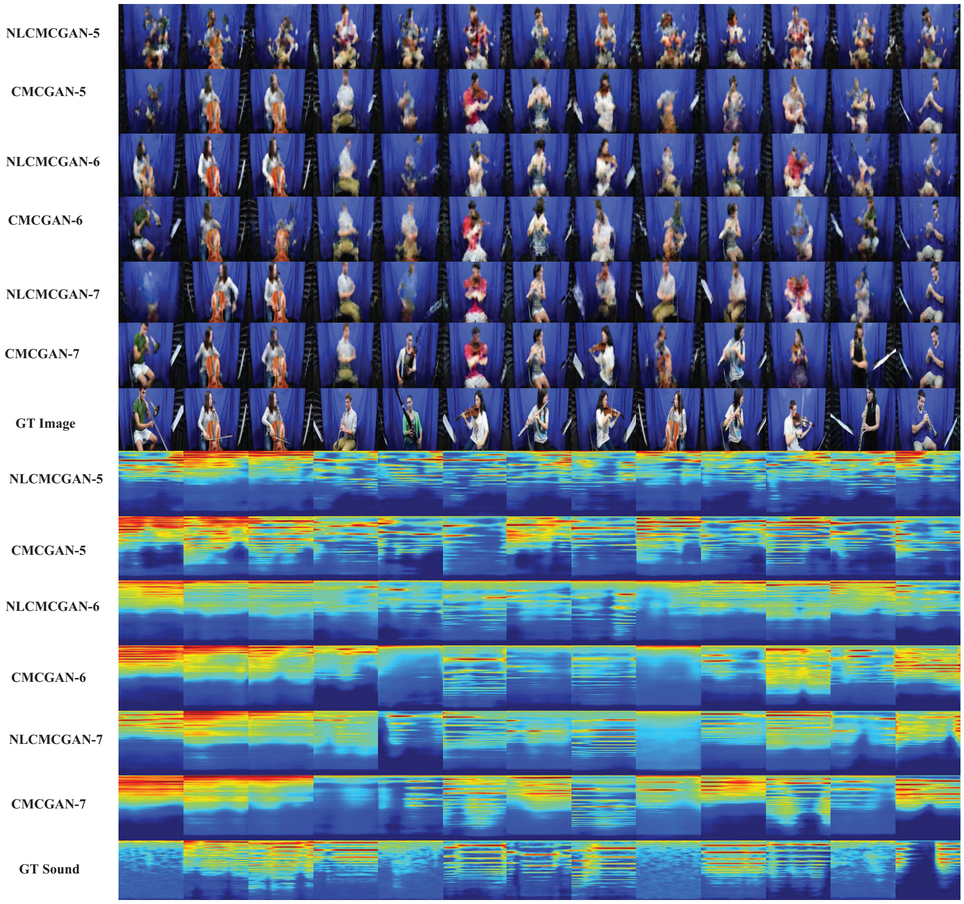

Figure 4: Cross-modal generated images and sounds based on models with different convolutional and deconvolutional layers.

work for different input modalities, which is displayed in Figure 6. If we have both image and sound samples, they are concatenated directly and sent to the subsequent classification network (solid arrows). If one of the modalities is lost,

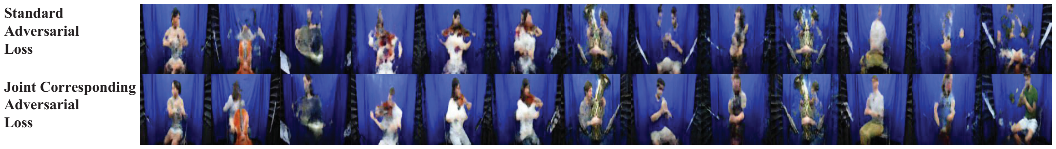

Figure 5: Generated images based on models with different loss. First row shows the generated images from model with standard adversarial loss. Second row displays the generated images from model with joint corresponding adversarial loss.

Models	V	A	V-A	GV-A	GA-V
Acc	0.9531	0.8863	0.9741	0.9804	0.9861

Table 2: The classification accuracies of different models.

it is generated from the other modality via CMCGAN, then the original and the generated modalities are concatenated together and sent to the subsequent classification network (dotted line/dotted arrows).

Results To verify the effectiveness of our dynamic multimodal classification network, we compare the classification accuracies of models with different inputs. They are V model (input only has image), A model (input only has sound), V-A model (input have both image and sound), GA-V model (input have image and generated sound) and GV-A model (input have sound and generated image). Among them, V-A, V-GA and GV-A models are executed via dynamic multimodal classification network. Comparison results are shown in Table 2.

Table 2 reveals that V-A model is superior to V and A models, which indicates the sufficient utilization of both visual and audio information underlying videos can boost the performance of specific tasks (such as classification). In addition, performances of GV-A and GA-V models are comparable or even better than that of V-A model, which denotes cross-modal generation can handle modality absent problem effectively. Moreover, GV-A and GA-V models surpass V-A model may due to the generated image/sound is more similar with those from training dataset other than testing dataset, which lead to better classification accuracies.

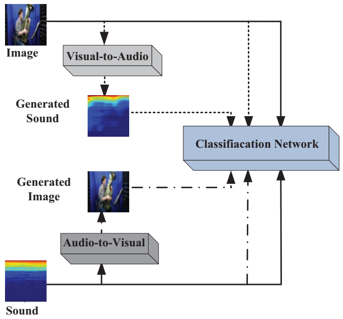

Figure 6: Dynamic multimodal classification network.

Discussions

Cross-modal visual-audio generation refers to restoring one modality from the other, which has attracted extensive attentions due to its potential capacity in handling modality absent problem. Although high-order information representation shared by visual and audio modalities may make their mutual generation be possible, information mapped from image to sound or its inversed is often expanded irrationally owing to dimension and struction discrepancies of different modalities. To solve this problem, a common Gaussian latent vector is combined with the high-level abstraction information of the source sample, which can refine the texture details of images and high frequency of sounds. Based on the same high-order feature representation, even a small perturbation of the Gaussian latent variable can be amplified into a significant image/sound difference. That is the generalization capabilities that we need to fit by our CMCGAN. With the extraction of modal features more abstract, CMCGAN tends to own more abundant and reasonable information for restoring missing modality.

Compared to standard adversarial loss, the joint corresponding adversarial loss in CMCGAN can provide extra image-sound mutual matching information in addition to image and sound realism to corresponding generators. These extra matching information will provide a finer constraint for image and sound generation, thus avoiding the information aliasing resulting from limit prior introduced by standard adversarial loss. Further, joint corresponding adversarial loss can unify visual-audio mutual generation into a common framework through mutual matching strategies.

Besides the group of generation paths that have same input and output modality (visual-audio-visual and audio-visual-audio), our model CMCGAN attaches another group of generation paths that conducting cross-modal visual-audio generation (visual-visual-audio and audio-audio-visual). The former kind of generation paths achieve promising generated image and sound, which is because that information underlying them is transferred from one modality to the same modality in some extent. Through sharing weights with the former generation paths at corresponding components, knowledge learned from them can be transferred to our cross-modal visual-audio generation paths, resulting in better cross-modal image and sound generation.

In addition to common information, two symbiotic modalities also contain complementary information. Fusing these complementary information will further enhance the related video tasks (here, we verify the classification task). However, due to environment inference or sensor fault, sometimes, only one modality exists while the other is abandoned or missing. Benefiting from cross-modal visual-audio mutual generation, our dynamic multimodal classification network can handle modality absent problem effectively. When one modality is missing, our dynamic multimodal classification network first generates it from the other modality through CMCGAN. Then, this modality and the generated one are fused together and fed to the subsequent classifier. Experimental results validate that the model with generated modality can obtain comparable or even better classification accuracy compared to the model with original modality. Reasons are that cross-modal generated image and sound via our CMCGAN can capture the promising discriminant information underlying original data. Moreover, the superiority of our model with generated data may due to the reason that the generated image/sound are more similar with those from training dataset, other than testing dataset.

Finally, the performance seems better in some particular genre of music. Based on our analysis on the experimental results, we can see our method gives better results when the training samples and the test samples have small variance, and vice versa. The advantage of our approach is that it can capture the essence of the cross-modal visual-audio mutual information to generate reasonable results even when the training samples and the test samples have significant gaps.

Conclusions

This paper proposes a CMCGAN model for cross-modal visual-audio mutual generation. Through introducing latent vectors, CMCGAN can handle dimension and structure asymmetry across two different modalities effectively. By developing a joint corresponding adversarial loss, CMCGAN can unify visual-audio mutual generation into a common framework and introduce more prior information for better cross-modal generation. Further, CMCGAN can be trained end-to-end to obtain better convenience. Numerous experiments have been conducted and our model CMCGAN achieves the state-of-art cross-modal generated images/sounds. Moreover, taking benefits from cross-modal visual-audio generation, we develop a dynamic multimodal classification network, which can deal with modality absent problem effectively.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 61773375, No. 61375036, No. 61602481, No. 61702510), and in part by the Microsoft Collaborative Research Project.

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

Aytar, Y.; Vondrick, C.; and Torralba, A. 2016. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, 892–900.

Chen, L.; Srivastava, S.; Duan, Z.; and Xu, C. 2017. Deep cross-modal audio-visual generation. *arXiv preprint arXiv:1704.08292*.

Feng, F.; Wang, X.; and Li, R. 2014. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, 7–16. ACM.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2016. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*.

Karacan, L.; Akata, Z.; Erdem, A.; and Erdem, E. 2016. Learning to generate images of outdoor scenes from attributes and semantic layouts. *arXiv preprint arXiv:1612.00215*.

Li, B.; Liu, X.; Dinesh, K.; Duan, Z.; and Sharma, G. 2016. Creating a musical performance dataset for multimodal music analysis: Challenges, insights, and applications. *arXiv preprint arXiv:1612.08727*.

Lu, Y.; Tai, Y.-W.; and Tang, C.-K. 2017. Conditional cyclegan for attribute guided face image generation. *arXiv preprint arXiv:1705.09966*.

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

Owens, A.; Wu, J.; McDermott, J. H.; Freeman, W. T.; and Torralba, A. 2016. Ambient sound provides supervision for visual learning. In *European Conference on Computer Vision*, 801–816. Springer.

Pereira, J. C.; Coviello, E.; Doyle, G.; Rasiwasia, N.; Lanckriet, G. R.; Levy, R.; and Vasconcelos, N. 2014. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(3):521–535.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Rasiwasia, N.; Costa Pereira, J.; Coviello, E.; Doyle, G.; Lanckriet, G. R.; Levy, R.; and Vasconcelos, N. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, 251–260. ACM.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*.

Sangkloy, P.; Lu, J.; Fang, C.; Yu, F.; and Hays, J. 2016. Scribbler: Controlling deep image synthesis with sketch and color. *arXiv preprint arXiv:1612.00835*.

Wang, K.; Yin, Q.; Wang, W.; Wu, S.; and Wang, L. 2016. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*.