Acoustic Scene Classification with Imagined Images

Yang Liu1,2, *AlexandrosNeophytou*[2], Sunando Sengupta[2], and Eric Sommerlade[2] [1]CVSSP, University of Surrey, UK
[2]Microsoft Corporation, Reading, UK yangliuav@gmail.com
{Alexandros.Neophytou, Sunando.Sengupta, Eric.Sommerlade}@microsoft.com

No Institute Given

Abstract. Acoustic scene classification is a challenging task since the recorded sound is sensitive to the kind of environment and susceptible to contamination from sounds produced by other objects in the scene vicinity. In this paper, we introduce an audio-visual generative adversarial network (AVGAN) to classify acoustic scenes and reconstruct visual scenes from the sound. For training AVGAN, an audio-visual dataset, called AV-10 where includes 24k sequences for 10 categories, is proposed[1]. In the training step, visual features are generated by passing frames through an encoder network. Audio information is used to generate *imagined* visual features with a generator network. These two features are distinguished by a discriminator. With the *imagined* features, an extra decoder is used to reconstruct the *imagined* images, and a classifier gives the category of the sound. We test our purposed AVGAN on the scene classification task of DCASE and ESC-50, where our method outperforms the state-of-the-art approaches.

Keywords: Scene classification, GAN, Image reconstruction, Audio-visual cross-modal learning

1 Introduction

Acoustic scene classification is a popular task to classify scenes with audio recordings into different classes. It has many applications in surveillance of abnormal sound [10], sound event detection [18] and navigation [8]. However, it is difficult to classify scenes only with audio information, especially in unknown cities, since the field of sounds is sensitive to environments, time and recording equipment. Various sounds can be transformed by the frequency domain properties and changing times, while subjected to interference from noise from surrounding objects in the immediate neighbourhood. In contrast to machines, humans can easily classify acoustic scenes with a short-time sound, especially if they have been there before. Based on the knowledge and experience, humans can imagine

[1] The dataset would be released shortly. The download link would be added in the camera-ready copy.

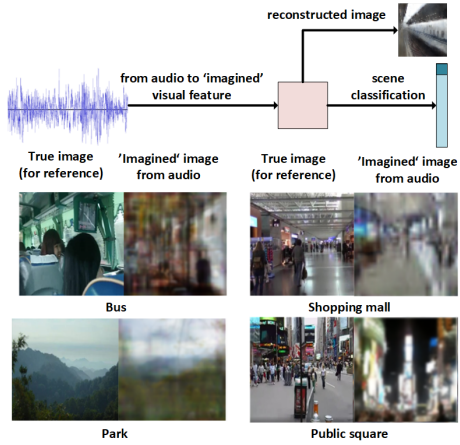

Fig. 1. Top: We consider the task of acoustic scene classification with *imagined* visual features reconstructed from short audio recordings; Bottom: We show four images pairs. The left images of each pair are examples class images shown for reference, and are unknown to our proposed model. The right images of each pair are the reconstructed *imagined* images generated only with audio recording on AV-10.

the scene and the *imagined* image is schematic and abstract [16]. The *imagined* image can eliminate the effects of time-varying sound and help humans detect the scene accurately. This is the main motivation of our work, where we mimic human behaviour by creating imagined scene features to aid the classification task.

In this work, we train a visual reconstruction model to eliminate errors caused by environmental change and proposed an audio-visual scene dataset AV-10. Our goal is to genetic *imagined* feature from a short audio recording to classify the scene and reconstruct an *imagined* image from the *imagined* feature. Fig. 1 shows sample results of our method. Note that our goal is not to generate the exact scene of the audio recording, but rather to use the *imagined* image to classify the audio recording.

We design an audio-visual generative adversarial network (AVGAN) including an encoder, decoder, generator, discriminator and classifier, shown in Fig.2. More specifically, the encoder can convert the visual frames to lower-dimensional and discrete features. The generator takes a short audio recording as input and predicts *imagined* features representing the scene. The discriminator is used to discriminate the between *imagined* features and the output of the encoder trained on images from an audio visual clip. The decoder converts these *imagined* features to the *imagined* images. Finally, the classifier is used to predict the category of audio recordings based on the imagined features. The encoder, decoder, generator and discriminator are trained on our AV-10 dataset, comprised of 24k 10-second videos with ten different classes. A large number of videos recorded by different equipment in various cities makes the *imagined* features robust to

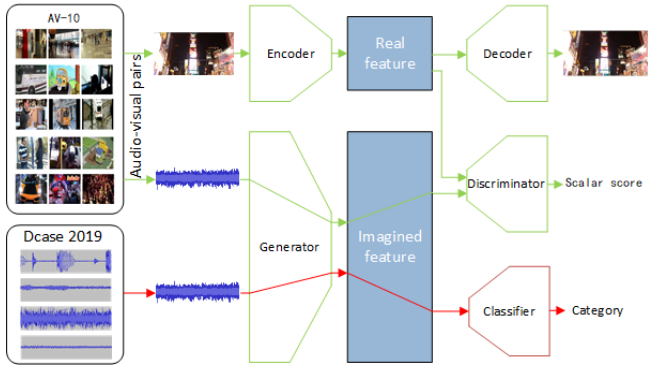

Fig. 2. AVGAN model and training pipeline. Audio-visual pairs of the AV-10 dataset are used to train the encoder, decoder, generator and discriminator (Green trapezoids). The goal of using a video dataset is to reconstruct an image from the audio recording. The *real* features and *imagined* features are shown as blue boxes. The audio recording of DCASE dataset is used to train a classifier (Red trapezoid) which predicts the category of the audio recording when the generator is fixed.

equipment and environment. We use DCASE [26], recorded by three types of equipment at ten places of six cities, and ESC-50 [30] including 50 categories to train and test the classifier.

We are not the first to use videos to improve acoustic scene classification. [4] propose audio-visual representations trained on videos and use them for classification. Although [6, 13] don't directly implement acoustic scene classification, they reconstruct images from the sound on the Sub-URMP dataset [6], which is also very close to our work. Compared to these methods and the state-of-the-art acoustic scene classification methods, our proposed AVGAN has four main contributions:

- Our proposed method performs well even in unseen cities due to the use of *imagined* features. In fact, the performance of AVGAN is similar for seen and unseen cities.
- *Imagined* features can be observed as the *imagined* images. The images are reconstructed from audio recordings without being restricted to traits. Compared to [6, 13, 4], our encoder and decoder are trained on YouTube data, resulting in addressing vanishing gradient and the mode collapse.
- Our proposed AVGAN does not need any pre-trained visual network as teachers for the audio network.
- A novel audio-visual scene dataset with 24k 10-second videos with ten different classes, which is a subset of Youtube 8M [1]. For each recording, apart from the video-level label given by the Youtube 8M dataset, fives visual scene labels and five acoustic scene labels with their weights are provided by two scene classifiers pre-trained on Places 365 [33] and Audio Set [11] dataset, respectively.

We test our method on ten different scenes in the DCASE dataset and ESC-50 dataset. The experimental results show our proposed AVGAN outperforms the state-of-the-art methods in DCASE 2019 and human performance in ESC-50 dataset.

2 Related Work

2.1 Acoustic scene classification

Many audio classification methods are based on deep learning technologies, such as fully-connected neural networks [20], convolutional neural network (CNN) [7] and recurrent neural networks (RNN) [19]. Kong et al. propose a cross-task learning method for audio-tagging, sound event detection and spatial localisation based on a nine-layer CNN [18]. Koutini et al. propose a novel frequency-aware CNN layer and adaptive weighting for addressing the noisy labels problem [21], which gets the top accuracy for the unknown cities in DCASE-2019 challenges. These CNN based methods usually use the log-mel spectrogram of audio recordings as input and output the presence probability of acoustic scene in either frame-level or video-level. However, the local spectrograms (lower frequencies or higher frequencies) of recordings belonging to the same scene are not the same in different cities, especially when the recording equipment is not matched [21]. To address this problem, Auxiliary Classifier GAN (ACGAN) is also used to generate additional samples into the training dataset [5]. As the experiments show, the GAN scheme can improve performance from 0.5% to 4%. However,the accuracy of ACGAN [5] decreased 8.8% in unseen cities.

2.2 Audio-visual cross-modal learning

Sound and vision are two basic sources through which humans can understand this world. Based on prior experience, a short sound bite is sufficient for humans to imagine a fitting environment. Audio-visual data offer a wealth of resources for knowledge transfer between different modalities [4, 24]. Our work is closely related to the reconstruction from audio recordings. For example, [27, 3] has used speech to reconstruct the speakers' face. [6, 13] uses conditional GANs and cycle GAN to achieve cross-modal audio-visual generation of musical performances. However, these works focus on special cases, such as faces or musical instruments, where the visual object is located at the centre of the images. For less constrained test data, the reconstructed images may be noisy, due to large variation. Although [2, 4] can not generate images, they use audio-visual representations trained on videos. [2] propose a generic audio-visual model to classify if a given video frame and a short audio clip correspond to each other. However, acoustic scene information is not enriched by visual information, due to the fusion layers only selecting the common features of the recording and images. [4] propose a student-teacher training procedure where a pre-trained visual recognition model transfers the knowledge from the visual modality to the sound

modality. Compared to [4], our proposed method has two main benefits: firstly, AVGAN does not need any pre-trained visual recognition model as the teachers of the audio model. Secondly, the reconstructed *imagined* images contain more visual information than other state-of-the-art methods, which can help researchers observe and understand the contribution of each hyperparameter and choose the optimal hyperparameters accordingly.

3 AVGAN

Our goal is to generate an image from an audio recording with a GAN network and to classify the audio recording using the visual information. However, the main challenge for training such a GAN network is that the distributions of the audio samples and visual samples are different and their Jensen–Shannon (JS) divergence [25] is large, which leads to vanishing gradient and mode collapse. Vanishing gradient leads the generator to learn extremely slow. Mode collapse causes the generated images from different audio recordings to be similar. Therefore, our work is not to generate images in a high-dimensional space, but rather to generate a low-dimensional intermediate representation, mapping audio and images. Using a low-dimensional space increases the probability of having overlaps of audio and visual distributions and decreases their JS divergence, thus addressing the vanishing gradient and the mode collapse problem [32].

Our AVGAN pipeline illustrated consist of 5 main components: 1. an encoder, which takes images of YouTube videos as input and produces low-dimensional *real* features in a discrete distribution 2. a decoder, which takes (*real* or *imagined*) features as input and reconstructs an image for observation. 3. a generator, which takes a complex spectrogram of sound as input, and predicts *imagined* features. 4. a discriminator to discriminate whether the low-dimensional features are real or imagined. 5. a classifier, which takes the (*real* or *imagined*) features as input and predicts the category of the features.

During training, we use two datasets, including the AV-10 dataset and an audio scene dataset (such as DCASE dataset) The training schedule can be described as follows: first, we only train the encoder and decoder to ensure images can be correctly reconstructed. Second, the generator is trained to generate low-dimensional features. Third, we only train the discriminator to distinguish between *real* and *imagined* features with the fixed encoder. These steps are repeated on the video dataset until the generator can generate features that closely match the output of the encoder. Finally, the generator is fixed, and we only train the classifier on the audio scene dataset. Now, we describe these five models in detail.

3.1 Encoder network and Decoder network

It is a hard and challenging task to generate continuous features such as colour information of images from the audio recording. Therefore, the encoded feature should be low-dimensional, abstracted and discrete to reduce the difficulty of

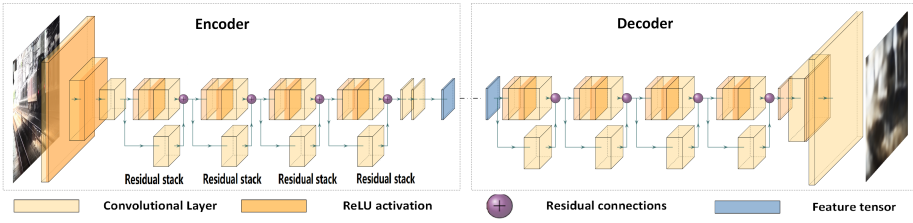

Fig. 3. Encoder and decoder architectures. For the encoder, the input image dimensions are 128×3×3 and the output feature dimensions are 16×16×1 (the blue part). For the decoder, the input is the feature (the blue part), and the output is the reconstructed image.

the generator. The encoder and decoder are modified from VQ-VAE [28], whose architecture is shown in Fig. 3. The convolutional layer is shown as light yellow, while ReLU activation is shown as dark orange. We chose a residual stack depth of four, which outputs a set of deep and discrete features. In each residual stack, the number of filters is 128. The input image of the encoder and output image of the decoder has the same dimension 128×128×3. The VQ-layer is used to get discrete features, where there are 16 embedding layers. The output of the encoder is the *real* features, while the input of the decoder is the *real* (or *imagined*) features of the audio recording. These features are shown as blue tensors. The training objective is:

$$L_{E,De} = \frac{\|v - De(E(v))\|_2^2}{\delta_v} + \|\text{sg}\left[E(v)\right] - e\|_2^2 + \beta \left\|E(v) - \text{sg}[e]\right\|_2^2 \qquad (1)$$

where v is the input image and δ_v is the diagonal covariance of training images. $De(E(v))$ is the reconstructed images from the *real* features $E(v)$, where E is the encoder and De is the decoder. e represents the embedding vectors and sg represents the stop-gradient operator that is defined as an identity at the forward computation time and has zero partial derivatives. The decoder with the embedding layers optimises the first two loss-terms, while the encoder optimises the first and the last loss terms. The weight of the latent loss of VQ-layer β is 1.

3.2 Generator network

In our experiments, when the generator is used to generate images directly, mode collapse frequently occurs, since the divergence between the audio and visual distributions is large. As a result, the output images are noisy and similar. To address this problem, the feature has been compressed to 16×16 by the encoder. As opposed to an image, the feature decreases the difficulty of the generation. The architecture of the generator is shown as Fig. 4. The input is the audio log-mel spectrograms [15]. For getting the deep features, 1024×1 vector (the layer before the dense layer) is calculated by three pooling layers and eight convolutional layers. This vector is further appended with the 16×1 vector. This

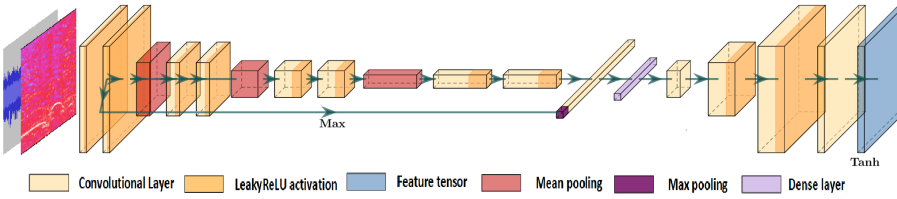

Fig. 4. Generator architectures. The input spectrogram dimensions are 64×64 (frequency×time) for audio recording and the output is *imagined* features. (Note, for better visualization, we shorten the depth of tensors)

16×1 vector is the max value of the spectrogram along the temporal dimension to preserve the local audio characteristics, since they are better contained in the frequency space, whereas linguistic information usually spans longer time duration [17]. The 16×1 vector is shortcut connected with 1024×1 vector for addressing vanishing gradients. The vector further passes a full connection layer, and batch normalization layer (the dense layer) and its length is decreased to 512. At the end of these blocks, we apply the tanh activation and obtain output of the generator as 16×16 feature map (see Fig. 4).

3.3 Discriminator network

The goal of the discriminator, also called "critic", is to distinguish between the *real* features encoded from the input images and the *imagined* features generated from the audio recording. Therefore, the output of the discriminator is a scalar score. Fig. 5 shows the discriminator architecture. For training stability, we use Minibatch Discrimination. When the similarity between the *imagined* features and *real* features increases, the mode collapses. The discriminator uses the value to detect generated images and penalise the generator. The training objective is:

$$\begin{aligned} L_{D,G} = & \mathbb{E}[\mathbb{E}[D(E(v)) - D(G(P(v)))]] \\ & - \lambda \mathbb{E}\left[(\|\nabla D(wG(P(v)) + (1-w)E(v))\|_2 - 1)^2\right] \\ & + \gamma \mathbb{E}[\frac{\|v - De(G(P(v)))\|_2^2}{\delta_v}] \end{aligned} \tag{2}$$

where $P(v)$ is the paired audio recording of image v. D is the discriminator and G is the generator. $G(P(v))$ represents the imaged feature and $E(v)$ represents the real feature. $\mathbb{E}$ is the expected value over all videos. Inspired by WGAN-PG [12], the third term is penalty gradient, where ∇ is the gradient operator and w is random value from 0 to 1. $wG(P(v)) + (1-w)E(v)$ represents the penalty sampled uniformly along straight lines between pairs of points sampled from the real feature distribution and the fake feature distribution. The fourth term represents the similarity between the *real* image v and *imagined* image from audio recording $De(G(P(v)))$.

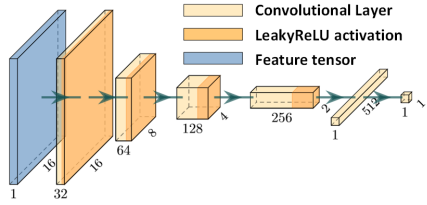

Fig. 5. Discriminator architectures. The input feature is 16×16.

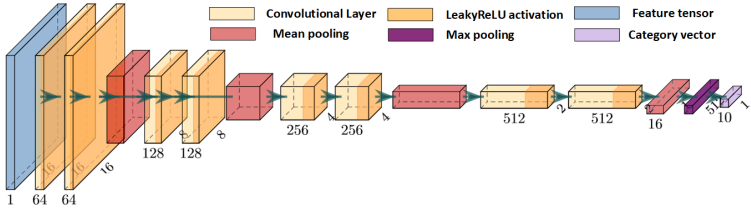

Fig. 6. Classifier architecture. The input feature dimension is 16×16, and the output is the category of sound recordings (the dimension is 10×1 for DCASE and 50×1 for ESC).

Our full objective with AV-10 dataset is:

$$L_Y = L_{E,De} + L_{D,G} \tag{3}$$

We should optimize the following loss function:

$$E^*, De^*, G^*, D^* = \arg\min_{G} \max_{D} \min_{De+E} (L_Y) \tag{4}$$

where $*$ is the optimized mapping.

3.4 Classifier network

Different from the previous four networks trained with video dataset, the classifier is trained on an audio dataset where the previous four networks are fixed. The goal of the classifier network is to predict the category of the audio recording in the audio dataset, such as DCASE dataset. The architecture is shown in Fig. 6, where the input is the *imagined* features. After eight convolution layers, four mean-pooling and a max-pooling layer, the output dimension is the number of the category with a cross-entropy training loss.

4 AV-10 Dataset

For audio scenes classification, DCASE dataset [26] and ESC-50 dataset [30] are widely used. The DCASE dataset is the subset of the TAU urban acoustic

Fig. 7. AV-10 Dataset: Sample frames from our video dataset. For visualization purposes, categories are named based on DCASE dataset. In AV-10 dataset, each image would be given 5 categories of Place 365 and 5 categories of Audio Set with their credibility.

scenes 2019 dataset [14] which consists of recordings from 10 acoustic scenes. For each scenes class, the recordings are done in the 6 to 12 cities by the Soundman OKM II Klassik/studio A3, binaural microphone, a Zoom F8 audio recorder, Galaxy S7, Iphone SE and Gopro H5 using 48kHz sampling rate and 24-bit resolution. The original recording was split into 10 seconds sequences. Since the DCASE considers the various scene, recording equipment and cities, it is mainly considered to evaluate our network. For further evaluating our network in a different scene, ESC-50 with 50 sound scenes with 2,000 sequences is used.

For video dataset, Sub-URMP [6], INIS [6], Youtube 8M [1] and SoundNet [4] datasets are widely used. The Sub-URMP dataset has the paired images and audio recording extracted from 107 videos of 13 kinds of instruments in the University of Rochester Musical Performance (URMP) dataset [23]. The INIS dataset contains ImageNet [9] images of five musical instruments and each image is paired with a short sound clip of a solo performance of the corresponding instrument. Since the Sub-URMP and INIS dataset only contains musical videos, they are not used in our network. Youtube 8M dataset is a large (350k hours) and imbalanced. However, it does not provide the original videos and only provides frame-level and video-level features as TensorFlow record files. Furthermore, most of Youtube video includes the background music making the classification task more challenging. Apart from that, Youtube 8M does not provide the original images. Therefore, the original Youtube 8M is not used. The SoundNet dataset is a set of videos from Flickr and content two million videos (about 20k hours). Compared to the Youtube 8M, the length of recording is short, and the recording is not edited. However, the SoundNet dataset does not provide the labels information. The recordings, which are not matched to the DCASE dataset and ESC-50 dataset, increase the difficulty of training GAN network. To the best of our knowledge, there is no existing dataset that we can

directly work on. Therefore, we compose a novel dataset, called AV-10, which is a subset of Youtube 8M. The training, validation and testing dataset have 24000 sequences (about 66 hours), 16000 sequences (about 44 hours) and 16000 sequences (about 44 hours). A sample subset of the video frames are illustrated in Fig. 7. In AV-10 dataset, the image resolution is 640×640. Audio recordings with two channels converted from videos, are resampled to 32 kHz which contains most energy. Compared to the original dataset, we make five improvements:

- AV-10 provides the download URL of each original video files. Since the Youtube 8M provides the Youtube ID for each video, we find the original URL of the most of videos and download them. Compared to Youtube 8M datset, the videos of AV-10 dataset allow the researcher to choose visual or audio features as they need.
- AV-10 only contains ten categories, corresponding to the categories of DCASE dataset. The congruent relationship between the categories of AV-10 and DCASE is shown in Table 1. The categories of AV-10 dataset are mentioned in 2^{nd} column.
- We curate the videos into several non-overlapping clips (10 seconds for DCASE, five seconds for ESC-50 and 1 second to provide the respective timing information). We ignore the first five-second clips at the beginning of each video which usually includes an intro.
- Each category in AV-10 is associated with up-to five visual scene labels as determined by a scene classification network. As each video, in Youtube 8M, is associated with multiple scene labels (an average of 3.01), some part of the clip might be unrelated, which affects the performance of the model in our early research. In order to effectively select appropriate YouTube videos, a VGG 16 model is pre-trained on the Places 365 dataset [33]. AV-10 dataset only considers the Youtube videos that are classified as the related labels (3^{th} column), corresponding to the categories of DCASE dataset, as shown in Table 1. This gives us a smaller and more relevant video subset (about 472 hours) for our audio-classification task.
- Each category in AV-10 is associated with up-to five acoustic scene labels as determined by a acoustic scene classification network. Some Youtube videos are modified by adding unrelated background music, making the generation difficult. It is an impossible task to reconstruct *imagined* images and classify the scene by the background music. A 9-layer CNN model [18] is pre-trained on the Audio Set dataset [11]. AV-10 dataset only considers the Youtube videos that are classified as the related categories (4^{th} column), corresponding to the categories of DCASE dataset, as shown in Table 1.

Since the same categories in the different dataset have different definitions, the categories among these four datasets do not have a direct one-to-one correspondence in Table 1. For example, 'Traffic' of Youtube 8M is an abstract category and is matched to three categories, 'road', 'highway' and 'street' of Place 365. Apart from that, some categories, such as 'Public square' of DCASE, do not appear in other dataset and are replaced by several related categories such as 'Times Square' of Youtube 8M and 'downtown' of Place 365. Please note

Table 1. Congruent relationship of the categories of Youtube 8M, DCASE, Place 365, Audio Set and AV-10 datasets. The categories of AV-10 dataset are named consistently as that of D-CASE.

Youtube 8M	DCASE/AV-10	Place 365	Audio Set
Airport terminal	Airport	airport_terminal	Aircraft, Rail transport, Subway metro
Bus	Bus	bus_interior, car_interior	Truck, Engine, Bus, Car, Train
New York City Subway	Metro	subway_station, train, car	Train wheels squealing, Subway metro
Train station	Metro station	train_interior, subway_station	Train wheels squealing, Rail transport, Train
National park	Park	park, garden, orchard	Outside rural or natural, Insect, Bird
Times Square	Public square,	picnic_area, downtown	Engine, Car, Vehicle, Outside rural or natural
Traffic	Street traffic,	road, highway, street	Motorcycle, Engine, Car,Vehicle
Shopping mall	Shopping mall	shopping_mall,	Crowd, Cheering, Children shouting
Road	Street pedestrian	road, crosswalk, boardwalk, highway	Engine, Car, Vehicle
Tram	Tram	tram, train, car	Engine, Subway metro, Vehicle, Train

that all category information is not used during the training of our model, but only for creating AV-10 dataset. The codes selecting the category with the two pre-trained models on Place 365 and Audio Set are available to download with the AV-10. We hope this helps the researchers in their future work.

To save the computation cost, all frames to train AVGAN are resized to 128×128 pixels. For the audio part of AV-10, DCASE and ESC-50, the audio waveform is resampled at 16k Hz. The spectrograms are computed by taking STFT with 16 FFT frequency bands. The training and test sets of DCASE include 14.4K and 5.8K audio recordings. Our network is implemented in TensorFlow and optimized by ADAM with $\beta_1 = 0.5$, $\beta_2 = 0.0002$, and the learning rate of 0.001 and the batch size of 16 for 9000 epochs. The network was trained for 223 hours on 4 NVIDIA 2080 Ti.

5 Experimentation result

We test our model on AV-10 for image reconstruction and the DCASE datasets with 10 categories [26] and ESC-50 dataset with 50 categories [30] for acoustic

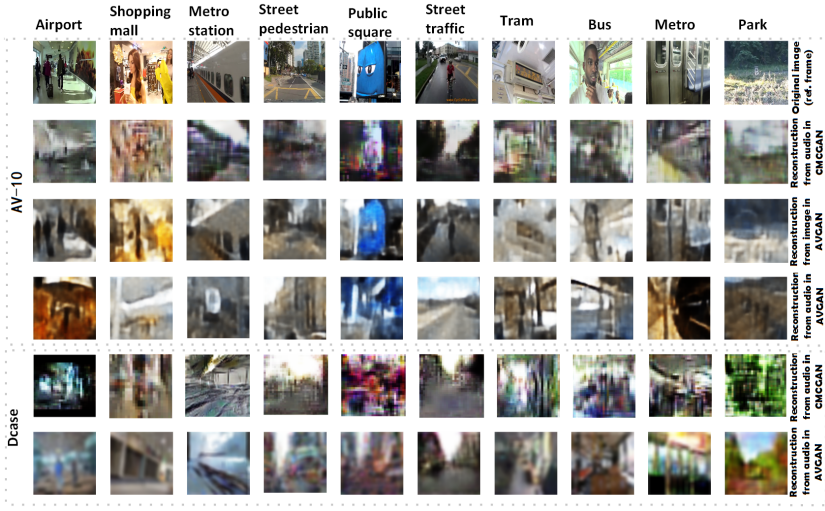

Fig. 8. Results of CMCGAN and our AVGAN on the AV-10 and DCASE: The top row describes the actual reference image corresponding to the scene in AV10 dataset. The second, third and fourth rows, describe the results obtained from CMCGAN, reconstruction from AVGAN using real features and reconstruction of AVGAN using imagined features respectively. For DCASE, we only show the results of CMCGAN and AVGAN with the *imagined* feature.

scene classification. Our goal is to classify acoustic scenes and to reconstruct images resembling the true scene images only with audio recording.

5.1 Image reconstruction

The reconstruction results of the CMCGAN [13] and our proposed method, AVGAN, on AV-10 and DCASE are shown in Fig. 8. For each example of AV-10, we show the true images for reference (1^{st} row), the reconstructed image of the CMCGAN (2^{nd} row), the reconstructed image of our proposed method from the image (3^{rd} row) and the reconstructed image of our proposed method from the audio (4^{th} row). Since the input of the CMCGAN and our AVGAN is the audio recording, the reconstructed images are fuzzy and have few differences to the reference image. The main difference is that some scene details, such as the people in the airport and the logo on the metro, are ignored since the high-frequency information is dropped by the VAE based encoder and decoder. Note, the goal of the reconstructed image is to classify the audio recording, not to recreate objects in a scene. For examples of DCASE, we only show the reconstructed images of CMCGAN (5^{th} row) and our proposed method (6^{th} row), since the true image is unknown. Compared to the results on AV-10, the results on DCASE become more abstract, since the density distribution of AV-10 and DCASE may be different. For example, the class label 'Park' means national park for AV-10

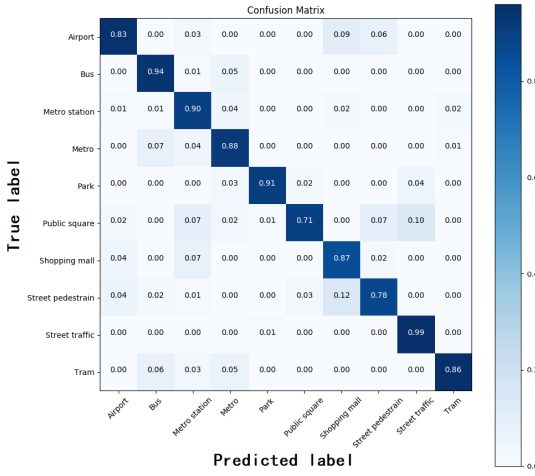

Fig. 9. Normalized confusion matrix of AVGAN for DCASE dataset.

and city park for DCASE. Compared to the results of CMCGAN, our method is more meaningful and clear, especially for the categories (public square, park) with the high variation. CMCGAN uses a generator to reconstruct the images directly. However, due to the high variance of the training data, the generated images are noisy. Since we use an intermediate representation, our generator does not suffer from mode collapse and AVGAN generates clearer images than CMCGAN.

5.2 Acoustic Scene Classification

Tables 2 and 3 show the accuracy of the different methods tested on DCASE and ESC-50 dataset respectively. As the length of audio recording is different in DCASE and ESC-50, their classifiers are separately trained. The input of the generator is ten seconds for DCASE while it is the five seconds for ESC-50.

On both benchmarks, we convincingly beat the previous state-of-the-art methods. For DCASE, although our method only improves the accuracy by 1% for the seen cities, the accuracy for unseen cities is improved by a large margin of 11%, since our proposed is trained on YouTube videos recorded in different cities. The results of our AVGAN and SDCNN both indicate that the GAN networks can improve the accuracy of the acoustic scene classification. For evaluating the effect of the AV-10 dataset, the accuracy of SDCNN trained on labelled AV-10 and DCASE (SDCNN + AV-10) is 87.1 for seen cities and 81.2 for unseen cities, which is lower than that of AV-GAN. The confusion matrix across all folds on DCASE is reported in Fig. 9. Since recordings of public square and street traffic are similar, the accuracy of AVGAN on these two categories are lower than others.

Table 2. Acoustic Scene Classification at the seen cities and unseen cities on DCASE.

Method	Seen Cities	Unseen Cities
SoundNet [4]	78.7	77.3
CCNN [31]	83.8	76.1
FACN [21]	84.9	78.1
SDCNN [5]	86.7	77.9
SDCNN + AV-10	87.1	81.2
AVGAN	**88.5**	**86.2**

Table 3. Acoustic Scene Classification at the seen cities and unseen cities on ESC-50.

Method	Accuracy
SVM-MFCC [30]	39.6
Autoencoder [4]	39.9
Random Forest [30]	44.3
Piczak ConvNet [29]	64.5
SoundNet [4]	74.2
AV learning [2]	79.3
Human [30]	81.3
Transfer learning [22]	83.5
AVGAN (ours)	**83.7**

For ESC-50, we have 6 % higher accuracy than human performance and have a similar performance to the transfer learning method [22], which is pre-trained on an audio dataset [11]. These results show that a large number of video data can improve the performance of classification. Note that since the size of the AV-10 dataset is huge, we only train our network with a subset for efficiency, so it is possible that further gains can be achieved by using all the available training data.

6 Conclusion

We have presented an audio-visual generative adversarial network (AVGAN) for reconstructing images from audio recordings and classifying acoustic scenes. Compared to the state-of-the-art audio-visual cross-modal learning methods, AVGAN does not to generate high-dimensional images with audio recordings, but rather generates a low-dimensional intermediate *imagined* features, mapping audio and images, which address the mode collapse problem. We have demonstrated that our method can reconstruct meaningful *imagined* visual scenes with the *imagined* features of audio recording. For audio scene classification, the experiments on DCASE dataset and ESC-50 dataset show we can achieve higher accuracy than the state-of-the-art methods, especially in unseen cities since AVGAN. In order to evaluate, a novel large video dataset, AV-10 dataset, is provided.

Compared to other video datasets, it has been cleaned and labelled and links to original recordings and the timing information will be provided. Although our method can achieve compelling results in acoustic scene classification and image reconstruction, the *imagined* images are a little fuzzy, since the encoder and decoder are modified from VQ-VAE where the high-frequency information is lost. Our future research will focus on how to reconstruct a scene with detailed geometric features.

References

1. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675 (2016)
2. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the Conference on Computer Vision. pp. 609–617 (2017)
3. Athanasiadis, C., Hortal, E., Asteriadis, S.: Audio-visual domain adaptation using conditional semi-supervised generative adversarial networks. Neurocomputing (2019)
4. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: Advances in Neural Information Processing Systems. pp. 892–900 (2016)
5. Chen, H., Liu, Z., Liu, Z., Zhang, P., Yan, Y.: Integrating the data augmentation scheme with various classifiers for acoustic scene modeling. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE) (2019)
6. Chen, L., Srivastava, S., Duan, Z., Xu, C.: Deep cross-modal audio-visual generation. In: Proceedings of the Conference on Thematic Workshops of ACM Multimedia 2017. pp. 349–357. ACM (2017)
7. Choi, K., Fazekas, G., Sandler, M.: Automatic tagging using deep convolutional neural networks. arXiv preprint arXiv:1606.00298 (2016)
8. Chu, S., Narayanan, S., Kuo, C.C.J., Mataric, M.J.: Where am I? Scene recognition for mobile robots using audio features. In: Proceedings of the Conference on multimedia and expo. pp. 885–888. IEEE (2006)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the Conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
10. Dimitrov, S., Britz, J., Brandherm, B., Frey, J.: Analyzing sounds of home environment for device recognition. In: Proceedings of the European Conference on Ambient Intelligence. pp. 1–16. Springer (2014)
11. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780. IEEE (2017)
12. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein GANs. In: Advances in neural information processing systems. pp. 5767–5777 (2017)
13. Hao, W., Zhang, Z., Guan, H.: Cmcgan: A uniform framework for cross-modal visual-audio mutual generation. In: Proceedings of the Conference on Artificial Intelligence (2018)
14. Heittola, T., Mesaros, A., Virtanen, T.: Tau urban acoustic scenes 2019, development dataset (2019)
15. Imai, S.: Cepstral analysis synthesis on the mel frequency scale. In: Proceedings of the Conference on Acoustics, Speech, and Signal Processing. vol. 8, pp. 93–96. IEEE (1983)
16. Intraub, H.: The representation of visual scenes. Trends in Cognitive Sciences **1**(6), 217–222 (1997)
17. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)

18. Kong, Q., Cao, Y., Iqbal, T., Xu, Y., Wang, W., Plumbley, M.D.: Cross-task learning for audio tagging, sound event detection and spatial localization: DCASE 2019 baseline systems. arXiv preprint arXiv:1904.03476 (2019)
19. Kong, Q., Iqbal, T., Xu, Y., Wang, W., Plumbley, M.D.: DCASE 2018 challenge baseline with convolutional neural networks. arXiv preprint arXiv:1808.00773 (2018)
20. Kong, Q., Sobieraj, I., Wang, W., Plumbley, M.: Deep neural network baseline for DCASE challenge 2016. Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE) (2016)
21. Koutini, K., Eghbal-zadeh, H., Widmer, G.: CP-JKU submissions to DCASE'19: Acoustic scene classification and audio tagging with receptive-field-regularized cnns. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE) (2019)
22. Kumar, A., Khadkevich, M., Fügen, C.: Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes. In: Proceedings of the Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 326–330. IEEE (2018)
23. Li, B., Liu, X., Dinesh, K., Duan, Z., Sharma, G.: Creating a musical performance dataset for multimodal music analysis: Challenges, insights, and applications. IEEE Trans. Multimedia, submitted. Available: https://arxiv. org/abs/1612.08727 (2016)
24. Liu, Y., Wang, W., Chambers, J., Kilic, V., Hilton, A.: Particle flow SMC-PHD filter for audio-visual multi-speaker tracking. In: International Conference on Latent Variable Analysis and Signal Separation. pp. 344–353. Springer (2017)
25. Manning, C.D., Manning, C.D., Schütze, H.: Foundations of statistical natural language processing. MIT press (1999)
26. Mesaros, A., Heittola, T., Virtanen, T.: Acoustic scene classification in DCASE 2019 challenge: closed and open set classification and data mismatch setups. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (2019)
27. Oh, T.H., Dekel, T., Kim, C., Mosseri, I., Freeman, W.T., Rubinstein, M., Matusik, W.: Speech2face: Learning the face behind a voice. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. pp. 7539–7548 (2019)
28. van den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. In: Advances in Neural Information Processing Systems. pp. 6306–6315 (2017)
29. Piczak, K.J.: Environmental sound classification with convolutional neural networks. In: IEEE International Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–6. IEEE (2015)
30. Piczak, K.J.: ESC: Dataset for environmental sound classification. In: Proceedings of the ACM international conference on Multimedia. pp. 1015–1018. ACM (2015)
31. Seo, H., Park, J., Park, Y.: Acoustic scene classification using various pre-processed features and convolutional neural networks. In: Proceedings of the Conference on Acoustics, Speech and Signal Processing (ICASSP) (2019)
32. Sutskever, I., Jozefowicz, R., Gregor, K., Rezende, D., Lillicrap, T., Vinyals, O.: Towards principled unsupervised learning. arXiv preprint arXiv:1511.06440 (2015)
33. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)