Visually Indicated Sound Generation by Perceptually Optimized Classification

Kan Chen[1], Chuanxi Zhang[1(✉)], Chen Fang[2], Zhaowen Wang[2], Trung Bui[2], and Ram Nevatia[1]

[1] University of Southern California, Los Angeles, USA
{kanchen,chuanxiz,nevatia}@usc.edu
[2] Adobe Research, San Jose, USA
{cfang,zhawang,bui}@adobe.com

Abstract. Visually indicated sound generation aims to predict visually consistent sound from the video content. Previous methods addressed this problem by creating a single generative model that ignores the distinctive characteristics of various sound categories. Nowadays, state-of-the-art sound classification networks are available to capture semantic-level information in audio modality, which can also serve for the purpose of visually indicated sound generation. In this paper, we explore generating fine-grained sound from a variety of sound classes, and leverage pre-trained sound classification networks to improve the audio generation quality. We propose a novel Perceptually Optimized Classification based Audio generation Network (POCAN), which generates sound conditioned on the sound class predicted from visual information. Additionally, a perceptual loss is calculated via a pre-trained sound classification network to align the semantic information between the generated sound and its ground truth during training. Experiments show that POCAN achieves significantly better results in visually indicated sound generation task on two datasets.

Keywords: Visually indicated sound generation · Perceptual loss

1 Introduction

When we observe visual events in the world, such as a stick hitting a metal object, or a car racing or a helicopter flying, we can immediately imagine and associate some sounds with these events. The objective of our paper is to synthesize realistic sound that correspond to the visual content in a silent video

K. Chen and C. Zhang—Equal contribution. Project page: www.github.com/kanchen-usc/VIG.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-11024-6_43) contains supplementary material, which is available to authorized users.

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11134, pp. 560–574, 2019.
https://doi.org/10.1007/978-3-030-11024-6_43

(*i.e.*, visually indicated sound generation). This ability is useful for many real applications, such as sound/video editing automation, enhanced experience of immersion in virtual reality and assistance for people with visual impairments.

Visually indicated sound generation is a challenging problem that involves parsing visual information and converting it into sound in audio modality. A number of methods have been suggested in recent work such as [1–3], which adopt a Convolutional Neural Network (CNN) to encode visual features and a Long Short Term Memory Network (LSTM) [4] or a Generative Adversarial Network (GAN) to generate sound. One common characteristic in these approaches is that they consider visually indicated sounds to belong to variations of a single class even though the sounds for different activities can be quite different. For example, in Fig. 1, the sound of hitting "iron cabinet" lasts longer than hitting "water"; besides, spectrograms of these two sounds show different distributions: the sound of hitting "iron cabinet" contains more high-frequency components than the sound of hitting "water".

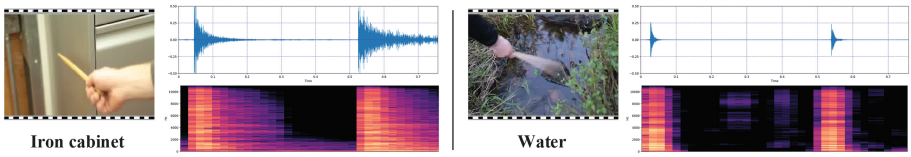

Fig. 1. Difference between sound of hitting "iron cabinet" and "water" in sound wave and spectrogram. It is hard for a generic model to handle all kinds of sound generation.

To address the significant variations, we introduce the concept of sound classes where each type of action generates sounds belonging to a specific class and then use class predictions to generate more finely tuned sounds. We average sound clips of same class to create a base sample. Given visual features, our audio generation model predicts sound class and transforms the predicted sound class's base sample to visually consistent sound. Furthermore, we leverage a state-of-the-art sound classification network to compute a *perceptual loss* [5] during training; this loss aims to align the predicted sound's semantic characteristics with ground truth in the feature space of the pre-trained sound classification network.

In implementation, we propose a novel Perceptually Optimized Classification based Audio generation Network (POCAN). POCAN adopts a CNN+LSTM structure to encode visual features, predict sound class and regress sound spectrograms. The generated sound wave is calculated as the Inverse Short Time Fourier Transform (ISTFT) [6] of the predicted spectrogram, which is the sum of predicted regression parameters and a base sample corresponding to the predicted sound class. During training, a pre-trained SoundNet [7] is deployed to compute the perceptual loss as the feature difference between the predicted sound and its ground truth. Analogous perceptual loss has been used for image generation [5] but is novel to audio generation, to the best of our knowledge.

We evaluate POCAN on the popular Greatest Hits Dataset [1]. Besides, we collected visual frames and evaluate POCAN on a subset of AudioSet [8]. Quantitative evaluations are conducted on sound classification and retrieval tasks which have also been used in [1,3] and have shown to have a high correlation with subjective evaluations. In both of these tests, POCAN outperforms state-of-the-art methods by a large margin. Besides, we provide some generated sound samples in the supplementary material for qualitative evaluation.

Our contributions are three-fold: (1) We propose to generate visually indicated sound considering different sound classes; (2) We leverage pre-trained SoundNet and apply a perceptual loss to refine POCAN during training; (3) We collect a visually indicated sound generation dataset and plan to release it upon publication.

In the following paper, we first discuss related work in Sect. 2. More details of POCAN and collected dataset are provided in Sects. 3 and 4 respectively. Finally we compare POCAN with other approaches in Sect. 5.

2 Related Work

Learning Visual-Audio Correlation by Video Self-supervision. Most videos contain synchronized visual and audio information, which provide self-supervision to learn the visual-audio correlation. Owens *et al.* [1] propose to learn visual features supervised by audio information, and achieve good performance in object detection task. On the other hand, Aytar *et al.* [7] deploy a deep convolutional network to learn efficient sound representations under visual supervision. Harwath *et al.* [9] apply visual supervision to predict similarity scores for input images and spoken audio spectrum. Arandjelovic *et al.* [10] propose a deep neural network to learn the visual-audio correlation, and achieve state-of-the-art performances on both visual and sound recognition tasks. POCAN also learns the audio-visual correlation and generates sound from visual frames under video self-supervision.

Mapping from visual signals to sound requires generating/retrieving reasonable sound clips based on visual information. Owens *et al.* [1] adopt a CNN+LSTM structure to regress cochleagram [11] of visually indicated sound based on video frames. Chen *et al.* [2] propose a GAN structure to generate sound of musical instruments conditioned on visual modality. Recently, Zhou *et al.* [3] adopt a SampleRNN [12] structure to generate visually indicated sound raw wave from visual features for each single sound class. Inspired by these models, POCAN adopts a CNN+LSTM structure to generate sound spectrograms.

Perceptual optimization has been successfully applied in image generation task [13–15]. Mahendran *et al.* [16] invert CNN features by minimizing a feature reconstruction loss to understand the visual information captured by different network layers. Based on this, Dosovitskiy *et al.* [17] propose to invert CNN features to image via a per-pixel reconstruction loss. Johnson *et al.* [5] apply a feature reconstruction loss to generate images, which achieves better perfor-

mance. Inspired by this success, we adopt a perceptual loss in audio modality to further boost sound generation.

Multimodal learning aims to learn the relationship between different modalities. Significant progress has been observed in visual and language modality learning, which includes Image Retrieval [18] and Visual Question Answering (VQA) [19,20]. Recently, Chen *et al.* [21,22] introduce regression and attention mechanisms in the grounding task. Based on recent progress in video detection and classification [23–25], Hendricks *et al.* [26] address the problem of temporal localization using natural language. In this paper, we focus on the learning of audio and visual modalities.

3 Method

POCAN is composed of two parts: classification based audio generation and perceptual optimization, which is shown in Fig. 2. In this work, we focus on generating visually indicated sound with fixed time length. We first present the framework of POCAN in Sect. 3.1, followed by the details of classification based audio generation and perceptual optimization in Sects. 3.2 and 3.3 respectively. Finally, we illustrate how to train POCAN and generate sound wave in Sect. 3.4.

3.1 Framework

The goal of POCAN is to generate a sound wave y given the corresponding video clip's frame sequence $\{x\}$. We do not generate raw sound wave directly from visual information; instead, we predict the spectrogram $\mathbf{s}$ of sound clip y, which can be converted back to a wave form via an Inverse Short Term Fourier Transform (ISTFT) [6]. To achieve this, fine-grained audio generation part predicts sound class probability distribution $\boldsymbol{p}$ as well as spectrogram regression parameters $\mathbf{d}$ based on visual features. According to the predicted distribution $\boldsymbol{p}$, the most probable sound class's base sample is selected, and the synthesized sound spectrogram $\mathbf{s}'$ is the addition of base sample and the predicted regression parameters. To capture semantic characteristics of real sound y, synthesized spectrogram $\mathbf{s}'$ is converted to wave form $\hat{y}$ via ISTFT. A perceptual loss $\mathcal{L}_p$ is then calculated by comparing the difference between SoundNet features of y and $\hat{y}$. The objective for POCAN is:

$$\arg\min_{\theta} \sum_{x} \mathcal{L}_{cls}(\boldsymbol{p}, c) + \lambda \mathcal{L}_{reg}(\mathbf{s}', \mathbf{s}) + \mu \mathcal{L}_p(\hat{y}, y) \tag{1}$$

where θ denotes the POCAN's parameters to be optimized. λ, μ are hyperparameters. c is the class label of sound clip y. $\mathcal{L}_{cls}$ is the loss for sound class prediction. $\mathcal{L}_{reg}$ is a regression loss for synthesizing sound spectrogram $\mathbf{s}'$. $\mathcal{L}_p$ is a perceptual loss for capturing semantic characteristics from real sound clip y.

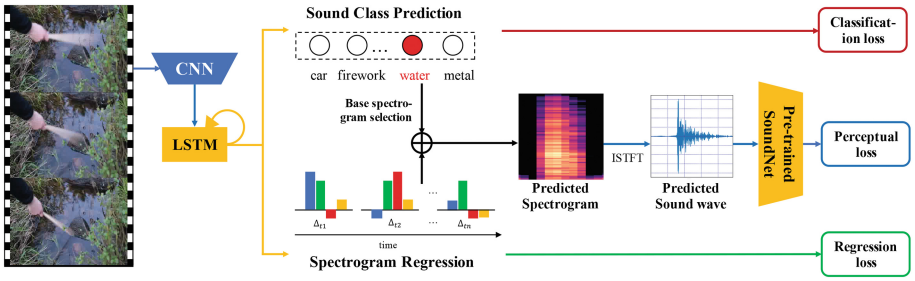

Fig. 2. Framework of Perceptually Optimized Classification based Audio generation Network (POCAN). Video frames are first processed by a CNN and then fed into a LSTM. To generate sound clips, POCAN predicts sound classes and regresses LSTM's hidden states into spectrograms and then transform the predicted spectrograms into sound waveforms. To increase the quality of generated sound, a pre-trained SoundNet [7] is applied to calculate perceptual loss during the training stage.

3.2 Classification Based Audio Generation Network

For visual input, each video frame x_i is encoded as a visual feature vector $\mathbf{x}_i \in \mathbb{R}^{d_v}$ by a pre-trained CNN [27]. d_v represents the dimension of visual feature vectors. To encode the temporal information in video frames, we feed these video features $\{\mathbf{x}_t\}$ into a LSTM [4], where the encoding procedure can be written as

$$\begin{aligned}
\boldsymbol{i}_t &= \sigma(\mathbf{W}_{vi}\mathbf{x}_t + \mathbf{W}_{hi}\boldsymbol{h}_{t-1} + \boldsymbol{b}_i) \\
\boldsymbol{f}_t &= \sigma(\mathbf{W}_{vf}\mathbf{x}_t + \mathbf{W}_{hf}\boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \\
\boldsymbol{o}_t &= \sigma(\mathbf{W}_{vo}\mathbf{x}_t + \mathbf{W}_{ho}\boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \\
\boldsymbol{g}_t &= \phi(\mathbf{W}_{vg}\mathbf{x}_t + \mathbf{W}_{hg}\boldsymbol{h}_{t-1} + \boldsymbol{b}_g) \\
\boldsymbol{c}_t &= \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \boldsymbol{g}_t \\
\boldsymbol{h}_t &= \boldsymbol{o}_t \odot \phi(\boldsymbol{c}_t)
\end{aligned} \tag{2}$$

where ϕ is the hyperbolic tangent function and $\odot$ represents the element-wise production between two vectors. The encoded features are represented as LSTM hidden states $\{\mathbf{h}_i\} \in \mathbb{R}^{d_h}$. d_h is the dimension of hidden space.

To predict sound class and regression parameters, we project the hidden states $\{\mathbf{h}_i\}$ into an audio space:

$$\boldsymbol{f}_i = \mathbf{W}\mathbf{h}_i + \mathbf{b} \tag{3}$$

where $\mathbf{W} \in \mathbb{R}^{(d_c+d_s)\times d_h}$, $\mathbf{b} \in \mathbb{R}^{(d_c+d_s)}$ are training parameters to be optimized. d_c denotes the number of sound classes, d_s is the feature dimension of sound spectrogram. The first d_c elements in $\boldsymbol{f}_i$ represent logits of sound class probability

prediction for frame x_i, while the rest elements record regression parameters of spectrogram. The classification loss $\mathcal{L}_{cls}$ is:

$$\mathcal{L}_{cls}(\boldsymbol{p}, c) = -\log \boldsymbol{p}[c],\ \boldsymbol{p} = \frac{1}{t_s}\sum_{i=1}^{t_s}\sigma(\boldsymbol{f}_i[0 : d_c - 1]) \tag{4}$$

where σ is a softmax function. t_s is the length of sequence $\{\boldsymbol{f}_i\}$, which is the same as the time length of spectrogram to be generated.

To synthesize spectrogram, we average different sound spectrograms according to their classes in the training set. Each class j then has an averaged spectrogram $\mathbf{A}_j \in \mathbb{R}^{d_s \times t_s}$ as a base sample. The synthesized sound spectrogram is calculated as:

$$\mathbf{s}' = \mathbf{d} + \mathbf{A}_{j^*},\quad \text{s.t.}\quad j^* = \arg\max_i\{\boldsymbol{p}[i]\} \tag{5}$$

where regression parameters $\mathbf{d} \in \mathbb{R}^{d_s \times t_s}$ are generated by stacking vectors $\{\boldsymbol{f}_i[d_c : d_c + d_s - 1]\}$. After obtaining fine-grained sound spectrogram, the regression loss $\mathcal{L}_{reg}$ is calculated by a smooth L1 regression loss function $g(.)$:

$$\mathcal{L}_{reg} = ||g(\mathbf{s}' - \mathbf{s})||_1,\ g(x) = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & |x| \geq 1 \end{cases} \tag{6}$$

3.3 Perceptual Optimization

To further improve the realism of the generated sound, we leverage state-of-the-art sound classification networks to capture different sound's semantic characteristics. Specifically, we adopt a pre-trained SoundNet [7], freeze its parameters and apply it to encode sound features for both real sound y and synthesized sound $\hat{y}$ during training. The synthesized sound wave $\hat{y}$ is generated from an ISTFT operation from predicted spectrogram $\mathbf{s}'$. The perceptual loss is then calculated by comparing features from real sound and synthesized sound:

$$\mathcal{L}_p = ||g(\phi(\hat{y}) - \phi(y))||_1 \tag{7}$$

where $\phi(.)$ denotes the feed-forward feature extraction process of SoundNet [7]. Other notations are the same as Eq. 6.

3.4 Training and Sound Wave Generation

Due to different sampling rates in visual and audio modalities (audio signal sample rate is much higher than the video frame rate), the length of visual feature sequence is shorter than the time length of audio's spectrogram. We uniformly replicate visual features in each time step so that visual sequence's length is the same as audio's spectrogram. The parameters to be optimized include parameters in LSTM and projection parameters in Eq. 3. POCAN is trained end-to-end using the Adam [28] algorithm.

Following [1] and [3], we generate and evaluate sound wave in two ways. First is directly converting synthesized sound spectrogram into sound wave via ISTFT;

we denote this generated sound as *raw sound*, which is useful for evaluating what information is captured by the audio features. Second, for the task of generating plausible visually indicated sound to human ears, we use the generated *raw sound* as a query and retrieve the nearest neighbor in the training set according to the similarities between spectrogram features, and set the retrieved real sound wave as our generation result. The similarity between two spectrogram features $\mathbf{s}_1$ and $\mathbf{s}_2$ is calculated as:

$$\mathtt{sim}(\mathbf{s}_1, \mathbf{s}_2) = \left\langle \frac{1}{t_s}\sum_{i=1}^{t_s} \mathbf{s}_1[i], \frac{1}{t_s}\sum_{i=1}^{t_s} \mathbf{s}_2[i] \right\rangle \tag{8}$$

where $\langle ., . \rangle$ represents cosine distance. We denote this second type of sound as *exemplar sound*, which is used for retrieval and human evaluation.

Table 1. Number training and testing samples in visually indicated sound generation (VIG) dataset

Class	Dog bark	Cattle	Sheep bleat	Chicken	Church bell
# Train	1124	739	1117	1085	1095
# Test	59	60	60	86	61
Class	Helicopter	Fire alarm	Hammer	Gunshot	Fireworks
# Train	1111	854	435	1001	1109
# Test	58	60	60	171	60
Class	Thunder-storm	Car racing	Rail transport	Splash water	Spray
# Train	1109	1098	1012	862	1119
# Test	62	72	167	58	60

4 Datasets

We evaluate POCAN on two datasets: Greatest Hits Dataset [1] and a manually annotated subset from AudioSet [8].

Greatest Hits Dataset (GHD) [1] contains 977 videos from indoor (64%) and outdoor (36%) scenes. There are 733 videos (21436 clips) and 244 videos (7008 clips) in this dataset for training and testing respectively. There are 17 sound classes in GHD ($d_c = 17$). Each labeled video lasts 0.5s with a single class label.

Visually Indicated sound Generation dataset (VIG) is a subset from AudioSet [8]. AudioSet [8] is a large-scale dataset of manually annotated audio events. There are 2,084,320 human labeled 10-s sound clips in 632 audio event classes from Youtube videos. Among them, we manually select 16,024 high quality sound clips in 15 classes which have strong correlation with visual frames ($d_c = 15$). Each sound clip belongs to one video. The number of training and test clips for each class is shown in Table 1. Some examples of video clips and corresponding sound clips are visualized in Fig. 3.

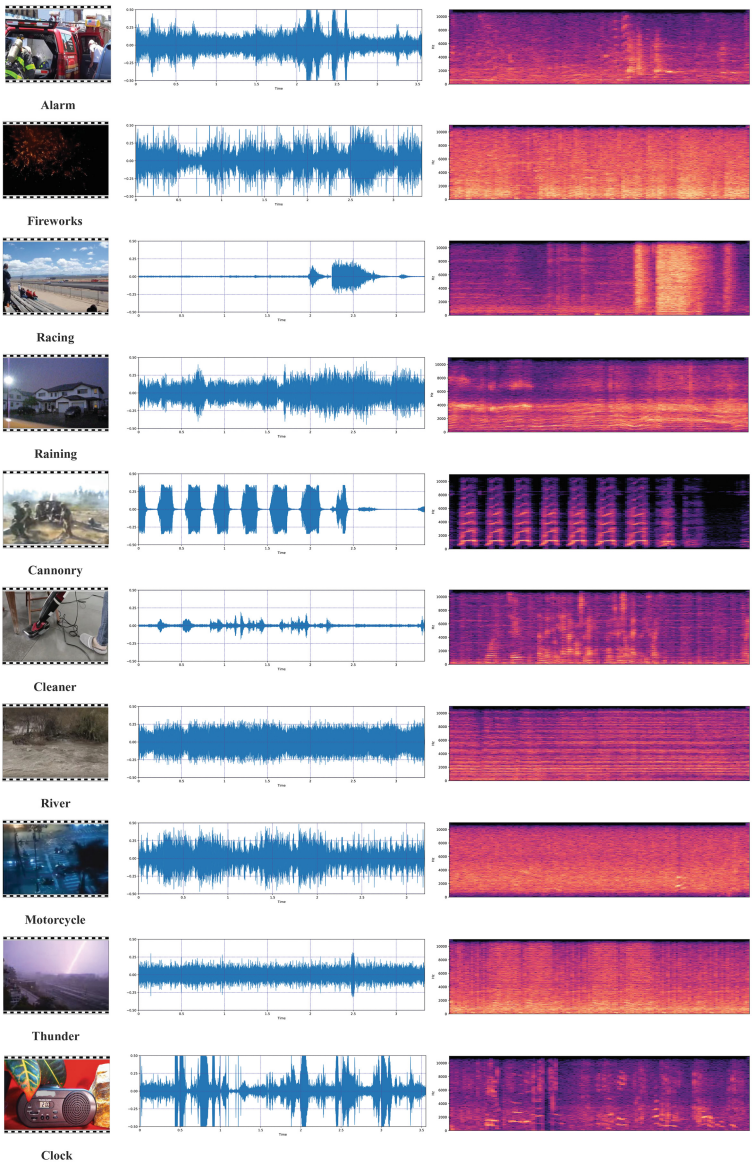

Fig. 3. Some examples of video clips and corresponding sound clips of VIG dataset

5 Experiments

POCAN is evaluated on GHD and VIG for visually indicated sound generation task. Since no public implementation is available for state-of-the-art method on GHD[1], we re-implemented the method in [1], but there may be differences from the authors' implementation.

5.1 Experiment Setup

We introduce the details of feature representation, model initialization, evaluation metric and compared method in this subsection.

Audio Feature Representation. To calculate regression loss in Eq. 6, we compute spectrogram of each sound wave via a Short Time Fourier Transform (STFT) [6] operation. We use Hann window [29] of size 2048 to encode and decode sound spectrograms. The feature dimension is 1025 ($d_s = 1025$). We use sample rate of 22.05 kHz and 8 kHz for sound wave on GHD and VIG respectively. The time length for spectrograms is 22 ($t_s = 22$) on GHD and 157 on VIG ($t_s = 157$). We denote spectrogram feature as "spec". For fair comparison with [1], we also extract cochleagram [11] for each sound clip, which is denoted as "coch", the feature dimension is 42 ($d_s = 42$). For perceptual loss in Eq. 7, we apply a pre-trained SoundNet [7] and extract its `conv7` features for each sound clip's real and predicted sound wave during training. The feature dimension is 1024.

Visual Feature Representation. We apply a 200-layer ResNet [27] pre-trained on ImageNet [30] to extract visual feature for each frame in a video clip. The feature dimension is 2048 ($d_v = 2048$). We denote these features as "res". For fair comparison with [1], we also apply an AlexNet [31] pre-trained on ImageNet [30] to extract visual features, with dimension $d_v = 4096$. We denote these features as "alex".

Model Initialization. During training, we set the batch size as 40. Hyperparameters λ, μ are set to be 50, 100 during training respectively. The dimension of hidden states of LSTM is 128 ($d_h = 128$). We apply Xavier method [32] to initialize training parameters in POCAN.

Evaluation Metric. We choose Recall at top K (R@K) as the evaluation metric. For generating each *exemplar sound*, we check the top K retrieved samples from the training set in the ranking list of each test sample. If there exists a retrieved training sample having the same sound class as the test sample, we consider it as a successful retrieval. R@K measures the success ratio of all test samples in the top K retrieval results. Besides, to compare with [1], we also train a 5-layer sound neural network classifier from real sound in the training set, and feed generated *raw sound* to check the classification results.

[1] Project page of Greatest Hits Dataset (GHD): http://vis.csail.mit.edu.

Compared Approach. We choose [1] as the compared method, which achieves state-of-the-art performance on GHD. We also re-implemented and evaluated [1] on VIG. We are unable to compare with [3] as the code and dataset is not publicly available at this time.

5.2 Performance on GHD

We evaluate different models' *exemplar sound* for R@K and *raw sound* for classification task on GHD respectively.

Comparison in R@K. Following the settings of [1], we adopt AlexNet features for visual modality and cochleagram for audio modality. The performance of [1] (alex + coch) is shown in Table 2. By replacing audio features with spectrogram, we observe a slight improvement in R@K (0.44% in R@1). Fixing spectrogram features, we then replace AlexNet features with ResNet features (Owens *et al.* [1] (res + spec)), and observe a further improvement of 2.19%, 2.72%, 7.55% in R@1, R@5, R@10 respectively.

Based on this feature combination (res + spec), we evaluate the Classification based Audio generation Network (CAN) in POCAN. In Table 2, we observe a significant improvement of 12.79%, 10.15%, 6.38% in R@1, R@5, R@10 respectively. This indicates CAN generates better sound so that the most similar retrieved samples contain more characteristics of the test sample's sound class. We further evaluate the full POCAN model on GHD and observe that POCAN achieves new state-of-the-art performance in similar sound retrieval task, with 15.68%, 13.70% and 7.15% increase over the method of [1] (res + spec) on R@1, R@5, R@10 respectively.

Table 2. Different models' performances of R@K on GHD (K = 1, 5, 10)

Model	K = 1	K = 5	K = 10
Owens *et al.* [1] (alex + coch)	0.1471	0.3982	0.4896
Owens *et al.* [1] (alex + spec)	0.1515	0.4077	0.5083
Owens *et al.* [1] (res + spec)	0.1734	0.4349	0.5838
CAN (res + spec)	0.3013	0.5364	0.6476
POCAN (res + spec)	**0.3302**	**0.5719**	**0.6553**

Sound Classification. Similar to [1], we evaluate whether generated sound contains semantic information of the sound class. We train a 5-layer neural network to classify different sounds. Each layer is a 1D convolution layer followed by a rectified linear unit (ReLU) non-linear activation function. This network is trained by real sound clips from the training set of GHD. In test stage, we generate the *raw sounds* from different models, and feed them into the pre-trained classifier. We calculate average classification accuracy for test set of GHD as the

Table 3. Classification accuracy of different model's generated sound by a pre-trained 5-layer neural network classifier.

Model	Accuracy (%)
Owens *et al.* [1] (res + spec)	20.11
CAN (res + spec)	35.46
POCAN (res + spec)	**36.32**
Real sound clips	51.34

metric. The classifier's performance as well as different models' sound classification accuracies are provided in Table 3. It is worth noticing that our neural network classifier achieves 51.34% classification accuracy, while a pre-trained SVM mentioned in [1] achieves 45.8%, which indicates that our classifier is better at classifying sound.

We observe that fine-grained generation part achieves better performance than [1] in sound classification task, This indicates that *raw sound* generated by fine-grained generation part provides more sound class information than that of [1], which is easier for classifier to recognize. We further apply the perceptual loss and evaluate sound generated by POCAN. Our classifier reports the highest classification accuracy of 36.32%, which is 16.11% higher than that of [1]. Besides, we draw the confusion matrix for sound classification results of [1] and POCAN in Fig. 4. From confusion matrices, we find POCAN's sound achieves consistently better performance over [1] in all sound categories. The sound classes with obvious improvement include tile, water and cloth.

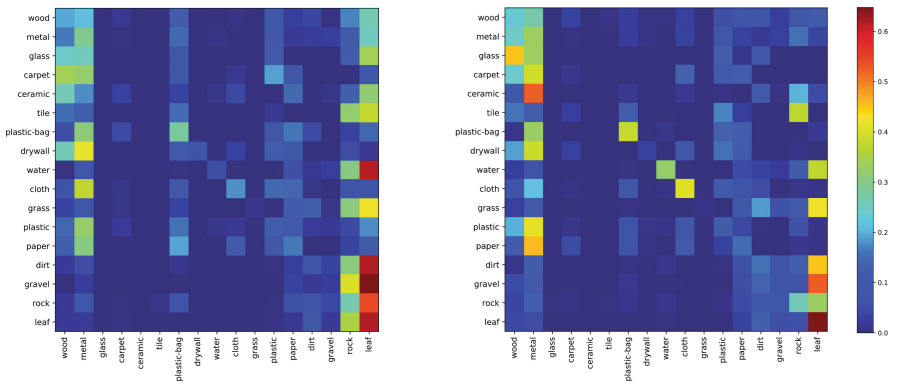

Fig. 4. Comparison of confusion matrices of sound classification results by a pre-trained 5-layer neural network classifier. Each row is the confusion made for a single sound class. Left and right figure is confusion matrix of sound generated by [1] and POCAN respectively.

Table 4. Different models' performances of R@K on VIG (K = 1, 5, 10)

Model	K = 1	K = 5	K = 10
Owens *et al.* [1] (res + spec)	0.0997	0.2888	0.4640
CAN (res + spec)	0.1180	0.3469	0.4709
POCAN (res + spec)	**0.1223**	**0.3625**	**0.4802**

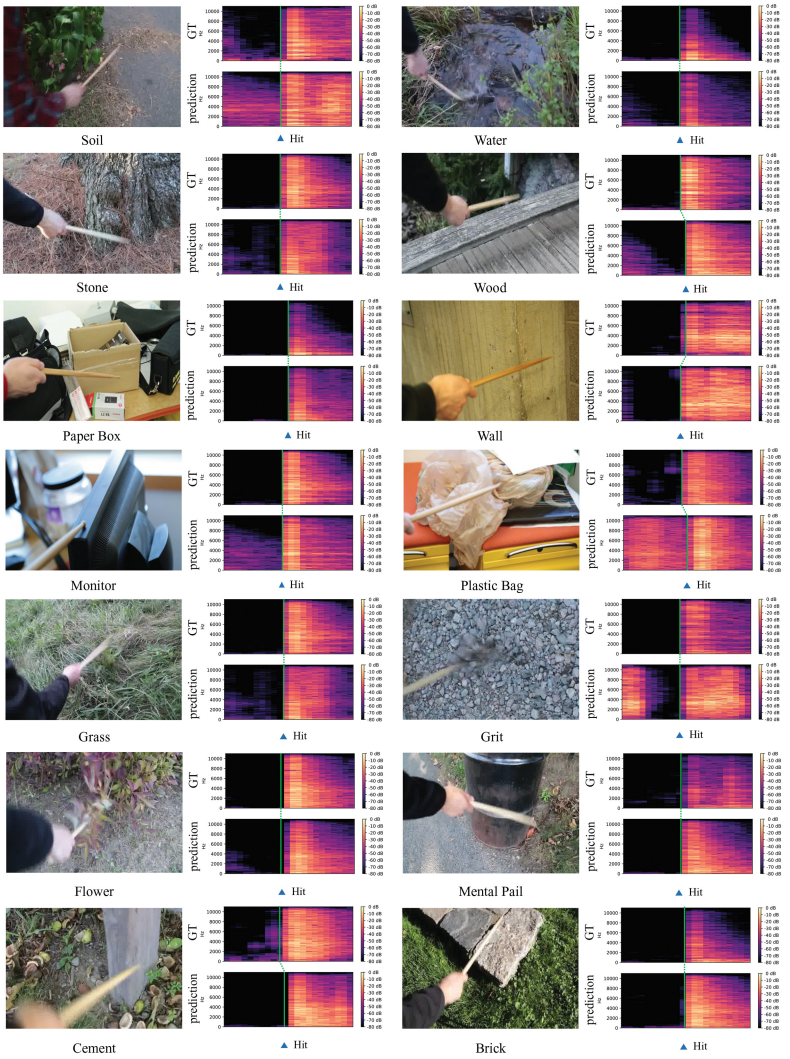

Fig. 5. Spectrograms of ground truth sound (GT) and retrieved *exemplar sound* by POCAN on GHD dataset. For each sample, we label some moments when actions happen in GT and *exemplar sound*.

5.3 Performance on VIG

Comparison in R@K. Based on the feature combination of ResNet as visual features ("res") and spectrogram as audio features ("spec"), we evaluate different models' R@K on VIG. In Table 4, by adopting CAN, we observe an improvement of 1.83%, 5.81% and 0.69% in R@1, R@5 and R@10 respectively. After applying the perceptual loss, POCAN achieves the state-of-the-art performance,

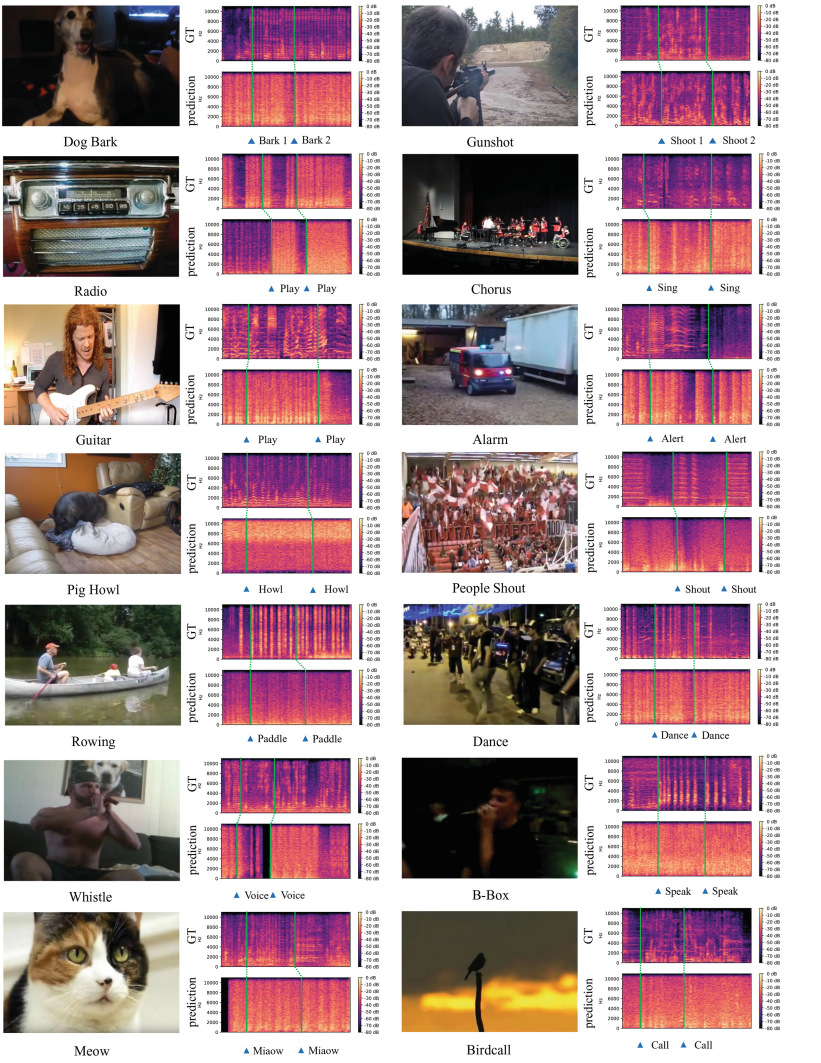

Fig. 6. Spectrograms of ground truth sound (GT) and retrieved *exemplar sound* by POCAN on VIG dataset. For each sample, we label some moments when actions happen in GT and *exemplar sound*.

with 2.26%, 7.37% and 1.62% increase in R@1, R@5 and R@10 over [1] respectively. We notice that the room for improvement on VIG is still big. This may be because the time length of sound clips in VIG is 10 s, while it is only 0.5 s on GHD. In this case, the sequences of both audio and visual features become 20 times longer, which brings extra difficulty for a system to generate reasonable sound.

5.4 Qualitative Evaluation

For qualitative evaluation, we visualize some spectrograms of *exemplar sound* generated by POCAN as well as its corresponding ground truth in Figs. 5 and 6. For each sample, we label its sound class and some time points when action happens in that clip. We observe the pattern of *exemplar sound* is similar to ground truth sound, and the occurrence of sound events are temporally close. However, POCAN also retrieves less similar samples which contain more actions or noise (*e.g.*, first result in row 2 of Fig. 6). The project page is in http://www.github.com/kanchen-usc/VIG, with demo video available online.

6 Conclusion

We proposed a novel Perceptually Optimized Classification based Audio generation Network (POCAN) which aims to produce visually indicated sound conditioned on video frames. Compared to previous methods, we consider sound class information and adopt a perceptual loss during training stage. To evaluate POCAN, we collect a visually indicated sound generation dataset from AudioSet [8]. Experiments show that POCAN provides significant improvement in visually indicated sound generation task on two datasets.

References

1. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: CVPR (2016)
2. Chen, L., Srivastava, S., Duan, Z., Xu, C.: Deep cross-modal audio-visual generation. In: ACM MM Workshop (2017)
3. Zhou, Y., Wang, Z., Fang, C., Bui, T., Berg, T.L.: Visual to sound: generating natural sound for videos in the wild. CoRR (2017)
4. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**, 1735–1780 (1997)
5. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
6. Welch, P.: The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans. Audio Electroacoust. **15**, 70–73 (1967)

7. Aytar, Y., Vondrick, C., Torralba, A.: SoundNet: learning sound representations from unlabeled video. In: NIPS (2016)
8. Gemmeke, J.F., et al.: Audio set: an ontology and human-labeled dataset for audio events. In: ICASSP (2017)
9. Harwath, D., Torralba, A., Glass, J.: Unsupervised learning of spoken language with visual context. In: NIPS (2016)
10. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: ICCV (2017)
11. Muthusamy, Y.K., Cole, R.A., Slaney, M.: Speaker-independent vowel recognition: spectrograms versus cochleagrams. In: ICASSP (1990)
12. Mehri, S., et al.: SampleRNN: an unconditional end-to-end neural audio generation model. In: ICLR (2016)
13. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. In: CVPR (2014)
14. Szegedy, C., et al.: Intriguing properties of neural networks. In: ICLR (2014)
15. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H.: Understanding neural networks through deep visualization. In: CVPR (2015)
16. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: CVPR (2015)
17. Dosovitskiy, A., Brox, T.: Inverting visual representations with convolutional networks. In: CVPR (2016)
18. Chen, K., Bui, T., Fang, C., Wang, Z., Nevatia, R.: AMC: attention guided multimodal correlation learning for image search. In: CVPR (2017)
19. Antol, S., et al.: VQA: visual question answering. In: ICCV (2015)
20. Chen, K., Wang, J., Chen, L.C., Gao, H., Xu, W., Nevatia, R.: ABC-CNN: an attention based convolutional neural network for visual question answering. In: CVPRW (2016)
21. Chen, K., Kovvuri, R., Gao, J., Nevatia, R.: MSRC: multimodal spatial regression with semantic context for phrase grounding. IJMIR **7**, 17–28 (2018)
22. Chen, K., Kovvuri, R., Nevatia, R.: Query-guided regression network with context policy for phrase grounding. In: ICCV (2017)
23. Gao, J., Chen, K., Nevatia, R.: CTAP: complementary temporal action proposal generation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11206, pp. 70–85. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01216-8_5
24. Myers, G.K., et al.: The 2014 SESAME multimedia event detection and recounting system. In: Proceedings of TRECVID (2014)
25. Gao, J., Yang, Z., Sun, C., Chen, K., Nevatia, R.: TURN TAP: temporal unit regression network for temporal action proposals (2017)
26. Hendricks, L.A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: ICCV (2017)
27. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
28. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2015)
29. Harris, F.J.: On the use of windows for harmonic analysis with the discrete Fourier transform. Proc. IEEE **66**, 51–83 (1978)
30. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR (2009)
31. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NIPS (2012)
32. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: AISTATS (2010)