SEEING THROUGH SOUNDS: PREDICTING VISUAL SEMANTIC SEGMENTATION RESULTS FROM MULTICHANNEL AUDIO SIGNALS

Go Irie[1] *Mirela Ostrek*[2] *Haochen Wang*[3] *Hirokazu Kameoka*[1]
Akisato Kimura[1] *Takahito Kawanishi*[1] *Kunio Kashino*[1]
[1]NTT Corporation [2]The University of Zagreb [3]The University of British Columbia

ABSTRACT

Sounds provide us with vast amounts of information about surrounding objects and can even remind us visual images of them. Is it possible to implement this noteworthy human ability on machines? In this paper, we study a new task that consists of predicting image recognition results in the form of semantic segmentation with given multichannel audio signals. Our approach uses a convolutional neural network that is designed to directly output semantic segmentation results by taking audio features as its inputs. A bilinear feature fusion scheme is incorporated that efficiently models underlying higher-order interactions between audio and visual sources. Experimental evaluations with both synthetic and real sound datasets show that our approach can recover the desired segmented images reasonably well.

Index Terms— cross-modal analysis, semantic segmentation, convolutional neural network, multichannel audio

1. INTRODUCTION

Human beings develop a deeper understanding of their surrounding environments by combining their senses. They know what kinds of sounds are likely to be generated by certain sources, and they can infer what kind of object is making a given sound and its location. For example, if we suddenly hear a barking sound, we can produce a mental image of a dog somewhere nearby, without looking at it. This remarkable ability called auditory scene analysis [1] has been proven to be particularly useful in our daily lives [2] and allows us to perform cross-modal mappings between natural sounds and their physical sources as in hearing a dog generates a visual image of a dog. This may pose one interesting question – *Is it possible to equip the machines with such a noteworthy ability?*

In this paper, we attempt to develop a method for predicting what objects are where in a scene from audio information alone, i.e., without actually looking at the scene. An overview of the problem considered in this paper is illustrated in Fig. 1. Suppose there are a few sound sources, e.g., people, standing in a room, and their voices and appearance are synchronously captured by a microphone array and a camera, respectively. Our task is to predict the semantic segmentation result, i.e., the pixel-level object classification result of the camera image similar to that shown in the bottom right part of the figure, solely from the recorded multichannel audio signals.

Some very recent studies have focused on cross-modal analysis between audio and visual information, such as visual feature learning from sounds [3, 4], sound prediction from silent videos [5], and cross-modal content generation [6]. A paper that is more relevant to ours is [7], which proposes learning audio features based on image-level object classification results. Another such paper is [8], which considers audio source localization from unlabeled videos. In contrast to these previous studies, we consider the new task of *predicting pixel-level semantic segmentation results from multichannel audio signals*. To the best of our knowledge, this topic has never before been addressed. This is a more challenging task that requires the joint handling of semantic and geometric correspondences between audio and visual sources at the pixel-level. It may provide the community with the opportunity to consider a new multiplex problem featuring audio and image processing that covers several central research topics including scene understanding, 3D geometry, object localization, and event detection.

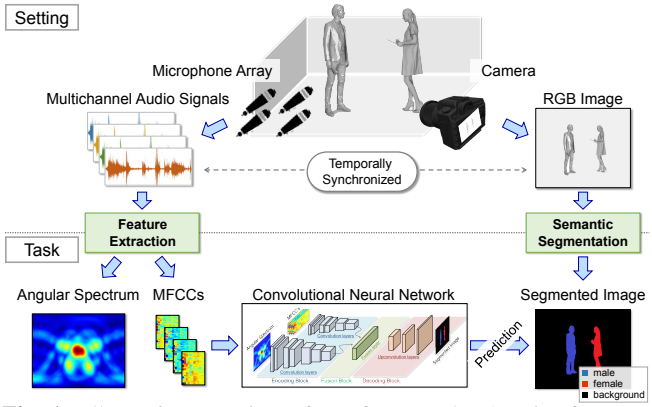

Fig. 1. Illustrative overview of our framework. A pair of temporally synchronized multichannel audio signals and an RGB image of a scene are captured by a microphone array and a camera, respectively. Our task is to predict the semantic segmentation result of the image from an angular spectrum and MFCCs extracted from the audio signals. To this end, our approach uses a convolutional neural network (CNN) with bilinear feature fusion.

Human beings probably gain the ability to handle such a complex cross-modal association by learning through experience [9]. Inspired by this, we approach the problem by considering machine learning using convolutional neural networks (CNNs). Our CNN architecture is designed to employ multichannel audio features as its inputs and to directly output the predicted segmentation result. We incorporate a bilinear feature fusion scheme in our network to efficiently model higher-order interactions between audio and visual information. Experiments on both synthetic and real sound datasets show that our model can recover desirable segmentation results reasonably well.

2. METHOD

The approach we propose in this paper first extracts audio features from multichannel audio signals, and then uses a CNN to predict the desirable segmentation results from the features. In this paper, we consistently consider a case where a microphone array consists of four microphones.

2.1. Audio Features

One straightforward choice would be to directly feed raw waveforms or spectrograms into a CNN as attempted before in [7] for audio

feature learning. However, the multichannel audio that we use is far richer than the monaural sound assumed in their work, hence, it is more important to extract the information needed to make learning stable.

Since semantic information and spatial information are both needed to recover the semantic segmentation results, the desired audio features should provide information about the spatial positions and categories of the sound sources (e.g., *male*, *female*, *dogs*, etc.) at the same time. We therefore extract the angular spectrum (AS) and the mel frequency cepstral coefficients (MFCC), which are often used for sound source localization and audio event detection, respectively. Specifically, we use the classical GCC-PHAT method [10] to estimate the two-dimensional AS in the form of a normalized 2D array of azimuth and elevation angles. As regarding the MFCCs, we first compute d-dimensional coefficients for w windows from each of four channels and then concatenate them into the form of a tensor with a size of $w \times d \times 4$ (channels). We use $d = 12$ and $w = 124$ in our method. The resulting AS and MFCCs can both be seen as 2D images, and thus can be fed into a CNN with 2D convolution layers.

2.2. CNN Architecture

Our CNN architecture is basically inspired by a fully convolutional network (FCN) for image semantic segmentation [11]. An FCN typically consists of a sequence of convolution and upconvolution layers for obtaining the direct mapping from an RGB image to its segmented form. However, such a simple architecture is insufficient in our case, which requires the integration of the two audio features, i.e., AS and MFCCs, and their translation to a segmented image, which is spatially inconsistent with the audio features.

We therefore designed our network to have the following three blocks, which we call the Encoding Block, Fusion Block, and Decoding Block. (i) The Encoding Block has two separate sequences of convolution layers for the two features, which makes it possible to preserve their meaningful information while reducing their spatial size, (ii) the Fusion Block fuses the two streams into a single feature map, and (iii) the Decoding Block is a sequence of upconvolution layers for upsampling the feature map to recover the desired segmented image. The overall architecture is illustrated in Fig. 2.

Bilinear feature fusion. The Fusion Block is especially important in our task for aligning audio-visual features. One simple solution would be to employ concatenation after aligning the dimensions of the two features by using fully-connected (FC) layers. However, such a straightforward application of FC layers tends to be costly and prone to overfitting. To avoid this problem, we propose the adoption of bilinear feature fusion [12, 13, 14], which has proved effective for modeling the inherent interactions behind spatially unaligned features with fewer parameters.

Denote the outputs of the Encoding Block for AS and MFCCs by $\mathbf{a} \in \mathbb{R}^{D_a}$ and $\mathbf{m} \in \mathbb{R}^{D_b}$, respectively[1], and the output of the Fusion Block by $\mathbf{v} \in \mathbb{R}^{D_v}$. The basic assumption is that the fused vector $\mathbf{v}$ can be computed by using a bilinear projection of $\mathbf{a}$ and $\mathbf{m}$.

$$\mathbf{v} = (\mathcal{T} \times_1 \mathbf{a}) \times_2 \mathbf{m}, \tag{1}$$

where $\times_i$ means a mode-i product, and $\mathcal{T}$ is a 3-way tensor of size $D_a \times D_m \times D_v$ which has to be learned through a training process. Although such a "full" tensor model is sufficiently flexible to model higher-order interactions between dimensions over the feature vectors, its huge number of parameters often makes this approach prohibitive. $\mathcal{T}$ can be decomposed by Tucker decomposition into

$$\mathcal{T} = ((\mathcal{T}_c \times_1 W_a) \times_2 W_m) \times_3 W_v, \tag{2}$$

where $\mathcal{T}_c \in \mathbb{R}^{d_a \times d_m \times d_v}$ is a core tensor that captures the principal interactions between the vectors, and $W_a \in \mathbb{R}^{D_a \times d_a}$, $W_m \in \mathbb{R}^{D_m \times d_m}$, and $W_v \in \mathbb{R}^{D_v \times d_v}$ are factor matrices for $\mathbf{a}$, $\mathbf{m}$, and $\mathbf{v}$, respectively. Typically $d_{\{a,m,v\}} \ll D_{\{a,m,v\}}$. Plugging this into Eq. (1), we obtain

$$\mathbf{v} = (((\mathcal{T}_c \times_1 (\mathbf{a}^\top W_a)) \times_2 (\mathbf{m}^\top W_m)) \times_3 W_v, \tag{3}$$

where the resulting number of parameters is significantly reduced to $d_a d_m d_v + D_a d_a + D_m d_m + D_v d_v$. The number of parameters can be reduced even further by assuming that the slice matrices inside the core tensor $\mathcal{T}_c$ are of low-rank [14]. If the rank is R, Eq. (3) turns into

$$\mathbf{v} = \left(\sum_{r=1}^{R} (\mathbf{a}^\top W_a M_r) * (\mathbf{m}^\top W_m N_r) \right) W_v \tag{4}$$

where $M_r \in \mathbb{R}^{d_a \times d_v}$ and $N_r \in \mathbb{R}^{d_m \times d_v}$ are compositions of r-th slice of $\mathcal{T}_c$. Our Fusion Block uses Eq. (4) to compute $\mathbf{v}$ and it is fed to the following Decoding Block. Note that Eq. (4) is differentiable with respect to all the unknowns, so they can be trained through backpropagation.

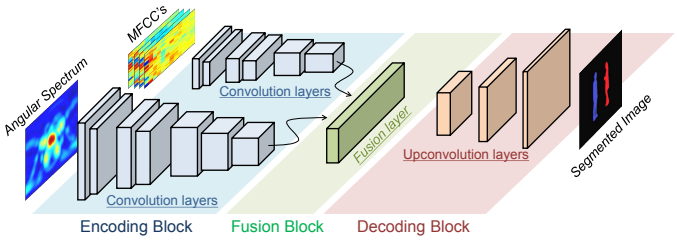

Fig. 2. Our CNN architecture.

Implementation details. The AS and MFCCs are first separately fed to the Encoding Block of our CNN. The AS stream consists of seven Conv-ReLU (ReLU after 2D convolution) layers with a kernel size of 3×3 and a stride of 2×2. The MFCC stream has six Conv-ReLU layers, where the first half have a kernel size of 3×3 and a stride of 2×1 and the second half have a kernel size of 3×3 and a stride of 2×2. The number of channels is gradually increased from 64 to 256 for both. The outputs from both streams have a dimension of 1024 after flattening (hence $D_a = D_m = 1024$). They are then fed into the Fusion Block. In the Fusion Block, the size of the core tensor $\mathcal{T}_c$ and the factor matrices W_a, W_m and W_v are set at $d_a = d_m = d_v = 32$, and $R = 10$. Given these numbers, the resulting output of the Fusion Block can be reshaped into a single 3-way tensor with a size of $6 \times 4 \times 128$ (hence $D_v = 3072$). The Decoding Block consists of three upconvolution layers and a softmax function at the top. Inspired by the generator used for deep convolutional generative adversarial networks (DCGANs), rather than FCN, they are designed to have relatively richer depths. Specifically, the configurations of the three layers are composed of a stride of 2, kernel sizes of 16, 8 and 4, and 128, 64, and C channels in that exact order, where C is the number of object categories (i.e., sound source categories) to be classified. The resolution of the final image is 360×480.

The model is implemented using Tensorflow. All the training is accomplished with Adam and takes 100 epochs from scratch. The learning rate is set at 1×10^{-4}, $\beta_1 = 0.5$, and $\beta_2 = 0.999$. The loss function used is cross-entropy loss.

[1] Although each is naturally a 3-way tensor as is, they can be readily reshaped into a vector by flattening.

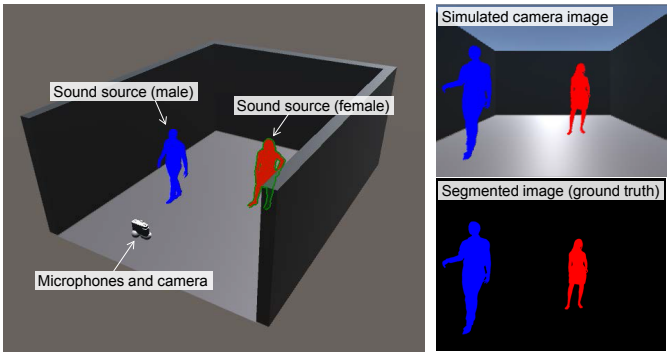

Fig. 3. Synthetic data generator and example of generated ground truth images.

Table 1. Results with synthetic dataset.

	pixel acc	mean acc	mean IU	f.w. IU
Spec	0.946	0.420	0.396	0.907
MFCC Only	0.940	0.483	0.438	0.904
AS Only	0.960	0.592	0.535	0.938
FC Fusion	0.966	0.671	0.596	0.943
Ours	**0.967**	**0.716**	**0.623**	**0.946**

3. EXPERIMENTS

To facilitate an evaluation of various sound source settings, we first evaluate our method using synthetic datasets with a reverberant room environment. Then we demonstrate the performance on a real sound dataset recorded with a simple setup.

3.1. Evaluation with Synthetic Dataset

We first explain how we generate our synthetic dataset and then report our quantitative and qualitative results.

Dataset Generation. We generated our dataset by using a data generator to simulate audio mixtures and camera images in a reverberant 3D room (Fig. 3). We assume situations where objects (sound sources) in five categories, i.e., *male*, *female*, *dog*, *cat*, and *pig*, are in a rectangular room 4.5 m wide $\times$ 6.0 m long $\times$ 2.5 m high, where our microphones and camera are mounted on a wall. Up to three object models (such as those shown in Fig. 3) from the five categories are positioned randomly in the room and produce sounds simultaneously. More specifically, the spatial coordinates of their positions are sampled from a uniform distribution with a constraint that requires the minimum distance between two arbitrary objects to exceed 30 cm. Ten types of different object models are prepared for each category. Given the positions, the audio signals to be captured by each microphone are synthetically generated by using the Roomsimove toolbox[2]. The sound sources of the human voices and animal sounds are selected randomly from TIMIT and (the corresponding sound classes of) ESC-50[3] [15], respectively, and assigned to the source positions of the object models (i.e., mouth positions). All the sources are resampled to 16 kHz. The ground truth semantic segmentation result is generated by assuming a normal perspective projection. The 2D camera image is first simulated by applying a perspective transformation to the 3D coordinate of the room, and a segmented image is then obtained by extracting the regions of the object models from the image. Example images are also given in Fig. 3.

The resulting synthetic dataset comprises $30,000$ pairs of multichannel audio signals and the corresponding ground truth images.

[2]http://bass-db.gforge.inria.fr/bss_locate/

[3]https://github.com/karoldvl/ESC-50

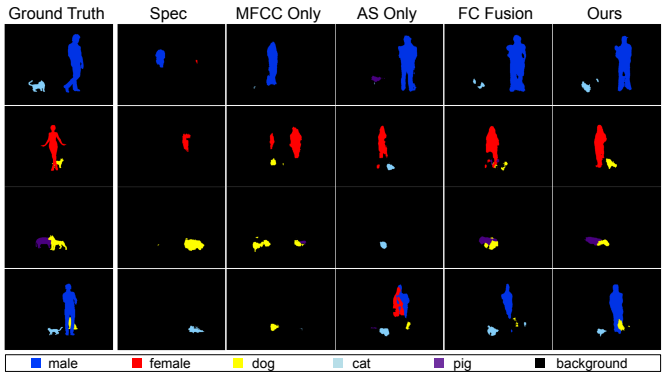

Fig. 4. Qualitative results with synthetic dataset. Colors correspond to categories of objects (sound sources).

We used $25,000$ for training and $5,000$ for the test.

Results. In addition to our method (**Ours**), we also evaluate several variants for comparative analysis. First, to clarify the effectiveness of two handcrafted features compared to raw signals, we evaluate a model **Spec** that uses complex spectrograms of four channels as the input to the CNN, instead of AS and MFCC. Second, to analyze the complementarity of AS and MFCC, we prepare two separate models of **AS Only** and **MFCC Only**, respectively, that use only of the two features. Third, to show the effectiveness of the bilinear feature fusion in Ours, we also evaluate a network that uses a fully-connected layer (followed by concatenation) for feature fusion (**FC Fusion**). For the sake of fairness, all the networks are designed to have approximately the same number of trainable parameters and depths. We evaluate the performance of the models with respect to four common metrics for semantic segmentation [11]: pixel accuracy (pixel acc), mean accuracy (mean acc), mean intersection-over-union (mean IU) and frequency weighted IU (f.w. IU). All of them take values in $[0.0, 1.0]$, and higher is better.

The quantitative results are shown in Table 1. First, all the methods yield prediction accuracies that are clearly better than chance levels. This confirms that our framework can recover the visual semantic segmentation results from multichannel audio. Second, AS Only and MFCC Only perform better than Spec. This may suggest that it may be difficult to find a meaningful mapping from such a highly redundant feature representation for the pixel-level prediction. Third, Ours and FC Fusion outperform MFCC Only and AS Only. This proves that the features are complementary in our task. Fourth, Ours is always competitive with or slightly better than FC Fusion especially in terms of mean acc and mean IU, which suggests that Ours can efficiently model the interactions behind the audio-visual information with the same number of parameters.

Fig. 4 shows some examples of ground truth and predicted segmentation results obtained with the compared methods. The results provided by Ours are visually closer to the corresponding ground truth images when they are compared with those obtained by the other methods. MFCC Only tends to return correct categories but incorrect positions even for relatively simple cases such as those in the first and second rows in Fig. 4 due to the absence of location information. Conversely, AS can almost always correctly predict positions but returns incorrect categories. Ours and FC Fusion clearly outperform AS Only and MFCC Only, which suggests that the two features are both important and complementary. The results predicted by FC Fusion are rather fragmented compared with those predicted by Ours, which may be due to overfitting. Ours can still return smooth and reasonable predictions even for a more complex example in the fourth row where a male is standing in front of a dog.

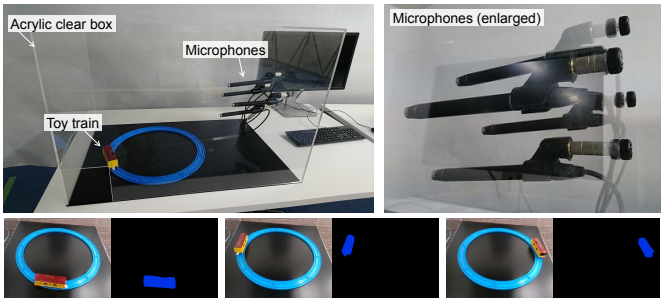

Fig. 5. Experimental setup for real sound experiments (top). Examples of RGB images captured by a camera and corresponding ground truth segmented images (bottom).

Table 2. Results with real sound dataset.

	pixel acc	mean acc	mean IU	f.w. IU
Spec	0.979	0.697	0.684	0.965
MFCC Only	0.993	0.915	0.874	0.987
AS Only	0.984	0.811	0.767	0.973
FC Fusion	0.993	0.918	0.883	0.988
Ours	**0.998**	**0.974**	**0.957**	**0.996**

3.2. Evaluation with Real Sound Dataset

The aim of this experiment is to analyze the applicability of our method to real sounds.

Dataset Generation. We use a structured setup as shown in Fig. 5 to generate our dataset. Four omnidirectional DPA ST4006A microphones and a top view camera synchronously record the sounds (running noise) and visuals of a small toy train running on circumferentially connected track. To control the reverb and noises, the entire setup is covered with a clear acrylic box 60 cm wide $\times$ 90 cm long $\times$ 60 cm high. We record 20 mins of audio signals and corresponding ground truth images and use half the results for training and the other half for the test. Other protocols are the same as those used in Sec. 3. Examples of the ground truth segmented images obtained with this setup are shown in the bottom row of Fig. 5.

Results. The quantitative results are shown in Table 2. The overall tendencies are similar to those found with the synthetic dataset. Most of the methods yield reasonable prediction accuracies. This confirms the applicability of our framework to real sounds. Ours is always competitive with or better than the other methods, which proves the effectiveness of our method in this setup.

Fig. 6 shows a few examples of the predicted results. As we can see from the results in the first and second rows, Ours and FC Fusion return more accurate prediction results than the others. Spec is unable to predict the positions of the train in most of the examples. From these observations, we can conclude that the combination of MFCC and AS is effective even with real sounds. A comparison of Ours with FC Fusion shows that the silhouettes predicted by FC Fusion tend to be degenerated, while Ours recovers a shape closer to that in the ground truth images. This emphasizes that the bilinear fusion scheme is more robust and can efficiently determine the semantic and geometric alignments over the audio-visual information.

4. CONCLUSION

We focused on the task of predicting image semantic segmentation results from multichannel audio signals, which has never before been addressed. This task requires us to associate both the semantic and geometric aspects of audio and visual information at the pixel level. To achieve this goal, we based our approach on a combination of CNN-based machine learning and handcrafted features extracted by using signal processing techniques. Experiments with both synthetic and real sound datasets proved that our approach could yield reasonable prediction accuracies. With these results, this paper will provide an opportunity for the community to consider a novel multiplex audio-visual processing problem. Conducting evaluations on more diverse sound source categories in more unstructured setups would constitute interesting future directions for this research.

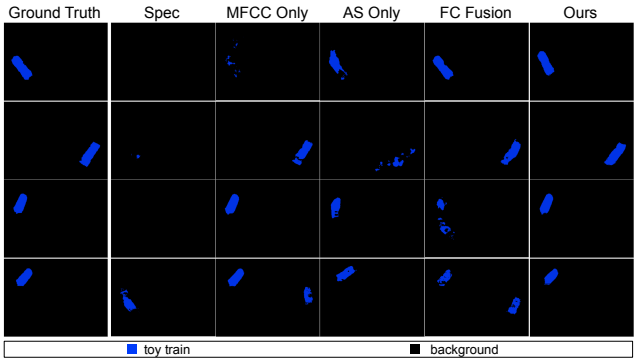

Fig. 6. Qualitative results with real sound dataset.

5. REFERENCES

[1] A.S. Bregman, *Auditory Scene Analysis*, MIT Press Cambridge, 1990.

[2] W.W. Gaver, "What in the world do we hear?: An ecological approach to auditory event perception," *Ecological Psychology*, vol. 5, pp. 1–29, 1993.

[3] A. Owens, J. Wu, J.H. McDermott, W.T. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," in *Proc. ECCV*, 2016.

[4] R. Arandjelović and A. Zisserman, "Look, listen and learn," in *Proc. ICCV*, 2017.

[5] A. Owens, P. Isola, J. McDermott, A. Torralba, E.H. Adelson, and W.T. Freeman, "Visually indicated sounds," in *Proc. CVPR*, 2016.

[6] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in *Proc. ACM Multimedia Workshop*, 2017.

[7] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning ssound representations from unlabeled video," in *Proc. NIPS*, 2016.

[8] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I.S. Kweon, "Learning to localize sound source in visual scenes," in *Proc. CVPR*, 2018.

[9] D.A. Bulkin and J.M. Groh, "Seeing sounds: Visual and auditory interactions in the brain," *Current Opinion in Neurobiology*, vol. 16, pp. 415–419, 2006.

[10] C. Knapp and G. Carter, "The generalized cross-correlation method for estimation of time delay," *IEEE TASSP*, vol. 24, pp. 320–327, 1976.

[11] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE TPAMI*, vol. 39, pp. 640–651, 2016.

[12] C. Feichtenhofer, K. Simonyan, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. CVPR*, 2016.

[13] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnns for fine-grained visual recognition," in *Proc. BMVC*, 2017.

[14] H. Ben-younes, R. Cadene, M. Cord, and N. Thome, "Mutan: Multimodal Tucker fusion for visual question answering," in *Proc. ICCV*, 2017.

[15] K.J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. ACM Multimedia*, 2015.