# Analysis on Roles of DNNs in End-to-End Acoustic Scene Analysis Framework with Distributed Sound-to-Light Conversion Devices

Yuma Kinoshita and Nobutaka Ono
Tokyo Metropolitan University, Tokyo, Japan

***Abstract*—We conduct an analysis on roles of deep neural networks (DNNs) in an end-to-end acoustic scene analysis framework with distributed sound-to-light conversion devices called *Blinkies*. Blinkies transmit sound information as the intensity of an onboard light-emitting diode (LED). A video camera can then easily collect acoustic information by capturing the LED intensities from multiple Blinkies distributed over a large area. In the end-to-end framework, both sound-to-light conversion and scene analysis processes are performed using two types of DNNs: an encoding network and a scene analysis network. These DNNs are optimized in an end-to-end manner for acoustic scene analysis. Although the efficacy of the end-to-end framework is already confirmed, it is unclear what role each network plays in the framework. In this paper, we examine the role of these networks through a simulation experiment using intermediate signal shuffling. Experimental results suggest that an encoding network does not output scene analysis results, but it achieves node-specific sound-to-light conversion that encodes spatial information of sounds. By using the node-dependent features obtained by encoding networks, the scene analysis network classifies scenes.**

## I. Introduction

Interest in acoustic scene analysis has recently increased, and many workshops and competitions have been held [1], [2]. Acoustic scene analysis is aimed at recognizing activities, such as "cooking," "vacuuming," and "watching TV," or determining what is going on, such as "being on a bus," "being in a park," and "meeting with people," from acoustic information [3]. For analyzing acoustic scenes with high performance, some methods utilize multiple microphones at the same time, that is, a distributed microphone array [4]–[7]. Distributed microphone array techniques are widely used not only for acoustic scene analysis but also for other purposes such as beamforming [8]–[10]. The use of a distributed microphone array enables us to obtain not only spectral information of a large area but also spatial information. In contrast, there are technical challenges in real-time acoustic sensing by using a distributed microphone array, i.e., cable connection with wired communication, network bandwidth limitation through wireless communication, or the synchronization of signals recorded using microphones.

To solve these challenges, we previously developed a sound-to-light conversion device called a *Blinky* shown in Fig. 1 [11]–[15]. A Blinky measured a sound signal using a microphone. In accordance with the sound signal, the Blinky modulated the intensity of an onboard light-emitting diode (LED). Finally, a video camera was used to synchronously capture LED intensities from multiple Blinkies distributed over a large area. To learn the optimal sound-to-light conversion process in Blinkies for acoustic scene analysis, we developed an end-to-end acoustic scene analysis framework with Blinkies [16]. In the end-to-end framework, we train two types of deep neural networks (DNNs) in an end-to-end manner: an encoding network that transforms a sound signal measured via a microphone into a signal to be transmitted by an LED and a scene analysis network that estimates the acoustic scene using captured LED intensities. The literature [16] showed that the end-to-end framework can give us more effective sound-to-light conversion for estimating the acoustic scene than a hand-crafted sound-power-based conversion. However, it is an open question what roles are played by the encoding network and the scene analysis network obtained by end-to-end learning.

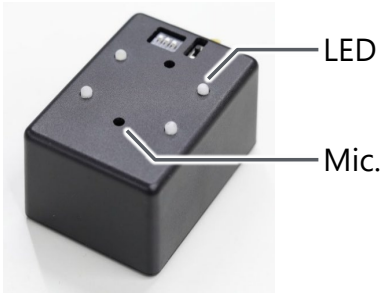



Fig. 1. Distributed sound-to-light conversion device *Blinky*

Because of such a situation, in this paper, we analyze the roles of DNNs in the end-to-end framework by a simulation experiment following a typical acoustic scene classification setup. For pattern recognition based on a sensor network such as a distributed microphone array, there are two main sensor fusion schemes [6], [17]: early fusion and late fusion. In the early fusion, signals acquired by sensor nodes are integrated before recognizing patterns. In the late fusion, pattern recognition is independently performed on each sensor node, then these results are integrated to obtain the final result.

To confirm whether the end-to-end framework performs early or late fusion, in the experiment, intermediate signals between the encoders and the scene analysis network were shuffled, and we investigated the effect of the shuffling on clas-

sification accuracy. If the end-to-end framework perform a late fusion such as a maximum likelihood approach, the shuffling of intermediate signals should not affect the accuracy. If the accuracy is degraded by shuffling, the end-to-end framework may perform early fusion and use the spatial information of the source or Blinkies for scene analysis.

Experimental results of a simulation experiment using the DCASE 2018 Challenge Task 5 dataset showed that the classification accuracy of the end-to-end framework trained without any shuffling was significantly degraded by the shuffling of intermediate signals during a test process. This result suggests that the end-to-end framework performs early fusion, that is, encoding networks trained without shuffling do not output scene analysis results but carry out a node-specific sound-to-light conversion. The node-dependent features obtained by encoding networks would include the spatial and spectral information, and they enable the scene analysis network to classify scenes with high accuracy.

## II. End-to-End Acoustic Scene Analysis Framework with Blinkies

The use of Blinkies enables us to avoid complicated processing, such as synchronization, in the signal acquisition using a distributed microphone array. In this section, we briefly summarize an end-to-end acoustic scene analysis framework with Blinkies.

### A. Overview

Figure 2 shows the proposed end-to-end acoustic scene analysis framework. In the framework, there is an assumption that $M$ Blinkies and a camera are placed at fixed locations. Acoustic scene analysis with Blinkies consists of three parts: sound-to-light conversion in each Blinky, light signal propagation in air, and scene analysis by using captured light signals. Let $N_{\mathrm{sound}}, N_{\mathrm{light}}, N_{\mathrm{frame}}$, and $N_{\mathrm{class}}$ be the length of a sound signal, the length of a light signal, the number of frames of a captured light signal, and the number of classes of acoustic scenes, respectively. The procedure can be written as

$$\boldsymbol{I}_m = \Phi_m(\boldsymbol{x}_m), \tag{1}$$

$$\boldsymbol{p}_m = \Xi_m(\boldsymbol{I}_m), \tag{2}$$

$$\hat{\boldsymbol{y}} = \Psi(\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_M), \tag{3}$$

where $\boldsymbol{x}_m$ is an $N_{\mathrm{sound}}$-dimensional vector that indicates an acoustic signal recorded by a microphone on the $m$-th Blinky, $\boldsymbol{I}_m$ is an $N_{\mathrm{light}}$-dimensional vector that indicates a light signal emitted by the $m$-th Blinky, $\boldsymbol{p}_m$ is an $N_{\mathrm{frame}}$-dimensional vector that indicates a video signal at the $m$-th Blinky captured by a camera, and $\hat{\boldsymbol{y}}$ is an $N_{\mathrm{class}}$-dimensional vector that indicates a predicted scene label. Functions $\Phi_m : \mathbb{R}^{N_{\mathrm{sound}}} \to \mathbb{R}^{N_{\mathrm{light}}}, \Xi_m : \mathbb{R}^{N_{\mathrm{light}}} \to \mathbb{R}^{N_{\mathrm{frame}}}$, and $\Psi : \mathbb{R}^{N_{\mathrm{frame}} \times M} \to [0,1]^{N_{\mathrm{class}}}$ denote sound-to-light conversion on the $m$-th Blinky, light signal propagation between the $m$-th Blinky and a camera, and scene analysis processes, respectively. We use two types of DNNs for $\Phi_m(\cdot)$ and $\Psi(\cdot)$: an encoding network and a scene analysis network. The former converts recorded signals into signals that can be effectively transmitted and are appropriate for scene analysis, and the latter performs scene analysis. To train these DNNs in an end-to-end manner, the light propagation $\Xi_m(\cdot)$ is modeled as differentiable physical layers.

### B. Encoding and scene analysis networks

The sound-to-light conversion $\Phi_m(\cdot)$ in the $m$-th Blinky is performed by using an encoding network. The encoding network is a 1D convolutional neural network (CNN) [18], and it downsamples microphone signals $\boldsymbol{x}_m$ using strided convolution layers. The downsampling rate is set in accordance with Blinky's audio buffer size.

For the scene analysis $\Psi(\cdot)$, a scene analysis network having a simple VGG-like architecture with 1D convolution layers [19] is used. Similarly to the encoding network, downsampling layers in this network are replaced with 1D strided convolution layers. and the resulting feature map is transformed by a global average pooling layer into a vector. The vector is fed into a linear layer to obtain the final scene analysis results.

### C. Differentiable physical layers

Light signal propagation $\Xi_m(\cdot)$ between the $m$-th Blinky and a camera is modeled as two differentiable physical layers: a light propagation layer and a camera response layer. The layers enable DNNs to consider physical phenomena.

*1) Light propagation layer:* LED light from Blinkies propagates in air, and a video camera captures it. The LED light intensity at the camera is affected by attenuation $a$ depending on the angle and distance between each LED and the video camera. In addition to this attenuation, ambient light is added to the light intensity as a positive bias $b$.

For these reasons, a light propagation layer models the signal transmission between a Blinky and a camera using attenuation $a$, bias $b$, and noise $\boldsymbol{\epsilon}$ as

$$\boldsymbol{I}_m = a\boldsymbol{x}_m + b\mathbf{1} + \boldsymbol{\epsilon}, \tag{4}$$

where $\mathbf{1}$ is a vector whose elements are one and whose size is the same as that of $\boldsymbol{x}_m$. We assume that attenuation $a$ is inversely proportional to the square of the distance between a Blinky and a camera, and $\boldsymbol{\epsilon}$ follows a normal distribution. $b$ can be calculated from the pixel value when the corresponding LED is not lit.

*2) Camera response layer:* An imaging sensor on a camera captures light, and the light is integrated over the time, which depends on the frame rate $F_{s,m}$ of the camera. A camera response layer is a model of the integration on a camera sensor. This integration can be interpreted as a sampling operation with low-pass filtering. For this reason, the camera response layer resamples an input signal $\boldsymbol{I}_m$ to the camera frame rate $F_{s,m}$ using

$$\boldsymbol{p}_m = \mathrm{resample}(\boldsymbol{I}_m), \tag{5}$$

where $\mathrm{resample}(\cdot)$ indicates the resample operation. Since most cameras have a frame rate of 30 fps, we set $F_{s,m}$ to

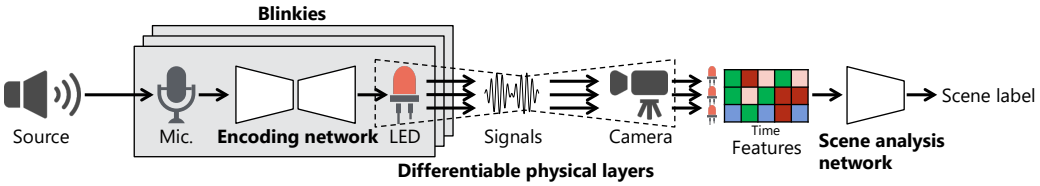



Fig. 2. End-to-end acoustic scene analysis framework with distributed sound-to-light conversion devices (Blinkies)

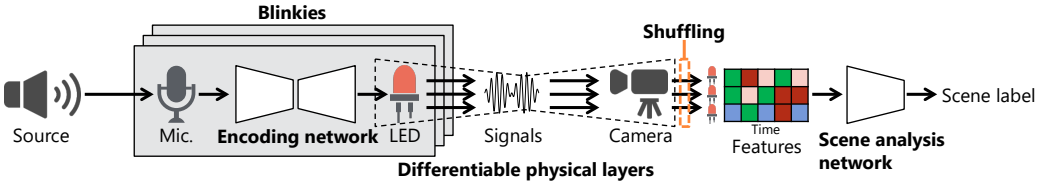



Fig. 3. Intermediate signal shuffling for studying end-to-end acoustic scene analysis framework with Blinkies

30 Hz in this work. Note that the nonlinear transform by a camera such as *gamma correction* can be avoided by using raw video frames. Hence, we do not consider such nonlinear transform in the camera response layer.

## III. Methodology

The efficacy of training both encoding networks and the scene analysis network is already confirmed in the literature [16]. However, it is unclear what roles are played by the encoding networks and the scene analysis network obtained by end-to-end learning. In particular, which does the end-to-end framework perform, early fusion or late fusion? For understanding the roles of the networks, we evaluate the scene classification accuracy of the end-to-end framework with shuffling intermediate signals between the encoding networks and the scene analysis network.

### A. Shuffling intermediate signals

Throughout this paper, we use a term "shuffle" in the meaning of giving a random permutation. Figure 3 illustrates how to shuffle intermediate signals. We shuffle the Blinky signals $\boldsymbol{p}_m$ given by Eqs. (1) and (2) and estimate the scene based on them. It can be written as

$$\tilde{\boldsymbol{y}} = \Psi(\boldsymbol{p}_{\sigma(1)}, \boldsymbol{p}_{\sigma(2)}, \cdots, \boldsymbol{p}_{\sigma(M)}), \tag{6}$$

where $\sigma(\cdot)$ denotes a random permutation of $(1, 2, \cdots, M)$ and $\tilde{\boldsymbol{y}}$ is a prediction under the shuffling.

This shuffling makes it difficult for the scene analysis network to utilize spatial information. For example, even when the first and second Blinkies are set close to the TV and kitchen, respectively, the scene analysis network does not utilize the information explicitly since it does not know which channels their signals appear. However, we have to note that this shuffling is different from the spatially shuffling of the Blinky position. During the end-to-end training, each Blinky can acquire a different encoder. Then, when we spatially permutate Blinkies, the relationship between observed sounds and encoders would change. While in the case of the shuffling in this paper, the relationship maintains.

In a simulation experiment in Sec. IV, we train the end-to-end framework with the shuffling in Eq. (6) under the following three conditions:

- Train both encoding and scene analysis networks without shuffling, i.e., conditions (a), (b), (c), and (d) in Table I,
- Train both encoding and scene analysis networks with shuffling, i.e., conditions (e) and (d) in Table I,
- Train both encoding and scene analysis networks without shuffling, and then train only the scene analysis network with shuffling, i.e., conditions (g) and (h) in Table I (fine-tuning).

After that, we test resulting three frameworks with and without shuffling.

### B. Relationship between shuffling and sensor fusion

There is a relationship between shuffling and sensor fusion. In the case of early fusion, each encoding network performs a specific sound-to-light conversion for each Blinky node. The scene analysis network integrates signals from encoding networks and recognizes scenes. In this case, the classification accuracy of the framework trained without shuffling will be degraded by shuffling during testing, and fine-tuning scene analysis network will not work. In addition, the classification accuracy of the framework trained with shuffling and tested without shuffling will be lower than that of the framework trained and tested without shuffling because spatial information of the source or Blinkies can be used in the former case, in addition to spectral information.

In the case of late fusion, in contrast, each encoding network independently recognizes scenes and these results are integrated in the scene analysis network to obtain the final result. For this reason, shuffling of intermediate signals should not affect the accuracy or accuracy degradation due to shuffling during testing can be avoided by fine tuning the scene analysis network. Since the spatial information cannot be used in the case of late fusion, the classification accuracy

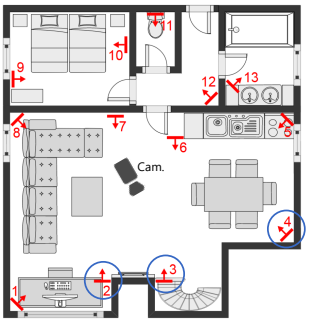



Fig. 4. Arrangement of microphone arrays [22]

of the framework trained with shuffling and tested without shuffling will be almost the same as that of the framework trained and tested without shuffling.

For this reason, we analyze the end-to-end framework by shuffling intermediate signals in the next section.

## IV. Simulation

We simulated the end-to-end framework that is intended to use Blinkies and evaluated it by an acoustic scene analysis experiment with the DCASE 2018 Challenge Task 5 development dataset [20], [21].

### A. Simulation Conditions

The DCASE 2018 Challenge Task 5 dataset is a derivative of the SINS dataset [22]. It contains a continuous recording of one person living in a vacation home for one week and scene labels for classifying sound clips from the recoding into nine scenes. Figure 4 shows the arrangement of the 13 microphone arrays used to construct the SINS dataset. Although the DCASE 2018 Challenge Task 5 dataset consists of a development dataset and an evaluation dataset, we used only the development dataset. This is because the evaluation dataset has no information on which microphone recorded each clip in the evaluation dataset. For this reason, we divided the development dataset into three subsets for training, validation, and testing. This partitioning was performed in accordance with a list for cross-validation provided with the dataset, and each subset contains sound clips from each of the nine scenes.

Sound clips in the DCASE 2018 Challenge Task 5 development dataset were recorded with four microphones called Nodes 1–4 (see Fig. 4). Their length and sampling frequency are unified to 10 s and 16 kHz, respectively. In these sound clips, we utilized clips recorded by Nodes 2, 3, and 4 for this simulation, because the number of clips recorded by Node 1 is different from those recorded by the other nodes.

We prepared three encoding networks and fed clips recorded by Nodes 2, 3, and 4 into the networks. Signals transformed by the networks and propagated through the differentiable physical layers were concatenated and fed into the scene analysis network, where we assumed that a camera was located at the center of the living room, as shown in Fig. 4. Under this assumption, the distances between the camera and Nodes 2, 3, and 4 were set to 1.13, 1, and 1.62, respectively, with the distance between the camera and Node 3 being 1. These networks were trained with 200 epochs using the training subset and well-known cross-entropy loss. Here, the Adam optimizer [23] was utilized for optimization, where the parameters in Adam were set as $\alpha = 0.001, \beta_1 = 0.9$, and $\beta_2 = 0.999$. The learning rate $\alpha$ was multiplied by 1/10 when the number of epochs reached 100 and 150. The method by He et al. [24] was used for initializing the network. The validation subset was used to check for the overlearning of the networks.

In addition to the end-to-end framework, i.e., the CNN-based encoding network + physical layers + VGG 1D (CNN / VGG 1D in Table I), we also evaluated a non-end-to-end framework with Blinkies using a sound-power-based sound-to-light conversion, i.e., power calculation + physical layers + VGG 1D (Power / VGG 1D in Table I).

### B. Results

Table I shows the classification accuracy for the test subset. From the table, we can confirm that condition (c) achieved a higher accuracy than condition (a). This result indicates that training both encoding and scene analysis network in an end-to-end manner is effective to obtain better sound-to-light conversion for acoustic scene analysis than the hand-crafted sound-power-based conversion.

Comparing condition (c) with condition (d), we can see that the end-to-end framework trained without shuffling is more sensitive to shuffling during testing. As a result, the accuracy of the end-to-end framework was significantly decreased, and this accuracy was lower than that of the non-end-to-end framework with shuffling during testing [see condition (b)]. For this reason, we think that each encoder trained without shuffling learned sound-to-light conversion specific to each node. In addition, this result suggests that the mismatch of microphone positions during training and testing may be discriminated from the scene analysis results.

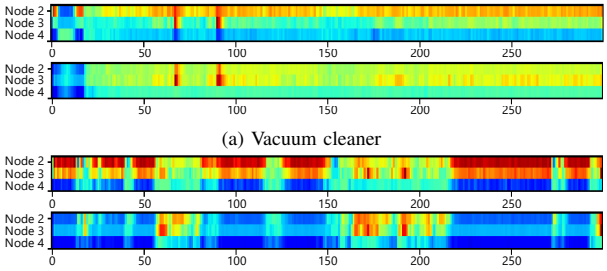
Shuffling during training made DNNs robust against shuffling as shown in conditions (e) and (f), but the accuracy was not as good as in condition (c) without any shuffling. This result indicates that spatial information of sounds and/or nodes are useful for acoustic scene analysis.

Fine-tuning of the scene analysis network with shuffling was not effective in improving accuracy as shown in conditions (g) and (h). This means that encoding networks did already perform scene classification, but that the scene analysis network perform classification by appropriately processing the node-dependent features obtained by the encoding networks.

Figure 5 shows examples of feature maps, i.e., outputs from camera response layers, obtained by "CNN / VGG 1D" trained with and without shuffling, where Fig. 5 (a) shows feature maps for a sound clip labeled "vacuum cleaner" and Fig. 5 (b) shows those for a sound clip labeled "social activity." As shown in this figure, the feature maps obtained by the framework trained without shuffling were different from those of the framework trained with shuffling. In the

TABLE I
TOTAL AND CLASS-WISE ACCURACY. END-TO-END ACOUSTIC SCENE ANALYSIS FRAMEWORK (CNN / VGG 1D) WAS TRAINED UNDER THREE SHUFFLING CONDITIONS: WITHOUT SHUFFLING, WITH SHUFFLING, AND FINE-TUNING. AFTER THAT, RESULTING FRAMEWORKS WERE TESTED WITH/WITHOUT SHUFFLING.

| Condition | (a) | (b) | (c) | (d) | (e) | (f) | (g) | (h) |
|---|---|---|---|---|---|---|---|---|
| Framework | Power / VGG1D | | CNN / VGG1D | | | | | |
| Shuffling (Train) | without | without | without | without | with | with | fine tuning | fine tuning |
| Shuffling (Test) | without | with | without | with | without | with | without | with |
| Total accuracy | 0.8038 | 0.7165 | 0.9394 | 0.5347 | 0.8617 | 0.8511 | 0.5729 | 0.4004 |
| Absence | 0.7329 | 0.7506 | 0.9484 | 0.2933 | 0.9349 | 0.9358 | 0.4413 | 0.1691 |
| Cooking | 0.6168 | 0.1433 | 0.9283 | 0.3925 | 0.5919 | 0.5296 | 0.7477 | 0.5202 |
| Dishwashing | 0.4607 | 0.1573 | 0.7865 | 0.1798 | 0.0787 | 0.0787 | 0.2472 | 0.0449 |
| Eating | 0.5245 | 0.1189 | 0.8182 | 0.1049 | 0.3357 | 0.4685 | 0.3566 | 0.0280 |
| Other | 0.7154 | 0.6000 | 0.6538 | 0.5154 | 0.4538 | 0.4615 | 0.7000 | 0.7692 |
| Social activity | 0.7375 | 0.7375 | 0.9575 | 0.2741 | 0.9575 | 0.8842 | 0.7336 | 0.2857 |
| Vacuum cleaner | 0.6333 | 0.1000 | 1.0000 | 0.2667 | 0.3500 | 0.0333 | 0.0167 | 0.0000 |
| Watching TV | 0.8995 | 0.9344 | 0.9484 | 0.9563 | 0.9703 | 0.9694 | 0.6643 | 0.7893 |
| Working | 0.9260 | 0.7815 | 0.9753 | 0.5587 | 0.9294 | 0.9209 | 0.5986 | 0.2993 |




(a) Vacuum cleaner

(b) Social activily

Fig. 5. Examples of feature maps. Top of each subfigure shows feature map of "CNN / VGG 1D" trained without shuffling [condition (c) in Table I] and bottom of each subfigure shows feature map of "CNN / VGG 1D" trained with shuffling [condition (d) in Table I]. The horizontal axis shows the discrete time index (video frame index).

case of vacuum cleaner, the sound of a vacuum cleaner was heard continuously and the framework trained with shuffling almost always produced high feature values for all three nodes. Note that amplitude of the signal from Node 4 become smaller than those of others since the signal is highly affected by attenuation. In contrast, feature values obtained by the framework without shuffling changed complexly. In the case of social activity, a man talked with a woman and this trend is more pronounced. In addition, the framework trained without shuffling yielded a lower feature value for Node 4 when it provided a higher feature value for Nodes 2 and 3. For these reasons, it is considered that the framework trained without shuffling acquired suitable sound-to-light conversion for each node, and the spatial or spectral information of the sound source was encoded in the complex changing feature pattern.

## V. CONCLUSION

In this paper, we conducted an study on the end-to-end scene analysis framework intended to use Blinkies to examine roles of the encoding networks and the scene analysis network by an simulation experiment using the DCASE 2018 Challenge Task 5 dataset. In the experiment, we trained and tested the networks with/without shuffling intermediate signals between the encoders and the scene analysis network, and we investigated the effect of the shuffling on the classification accuracy. Experimental results suggest that encoding networks trained without shuffling did not output scene analysis results but they achieved node-specific sound-to-light conversion that encodes spatial information of sounds, and the scene analysis network classifies scene by using node-dependent features obtained by encoding networks. It is also suggested that the mismatch of microphone positions during training and testing may be discriminated from scene analysis results.

In future work, we will conduct experiments using eight microphones in the SINS Database in order to study the effects of the number of Blinkies. In addition, we will develop a novel method to detect and compensate the mismatch of microphone positions during training and testing in accordance with scene analysis results.

## ACKNOWLEDGMENT

This work was supported by JST CREST Grant Number JPMJCR19A3 and JSPS KAKENHI Grant Number JP20H00613

## REFERENCES

[1] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "CLEAR Evaluation of Acoustic Event Detection and Classification Systems," in *Multimodal Technologies for Perception of Humans*, Springer Berlin Heidelberg, 2007, pp. 311–322.

[2] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2013, pp. 1–4. [Online].

[3] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic Scene Classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, May 2015.

[4] P. Giannoulis, A. Brutti, M. Matassoni, A. Abad, A. Katsamanis, M. Matos, G. Potamianos, and P. Maragos, "Multi-room speech activity detection using a distributed microphone network in domestic environments," in *Proceedings of European Signal Processing Conference*, Aug. 2015, pp. 1271–1275.

[5] J. Kürby, R. Grzeszick, A. Plinge, and G. A. Fink, "Bag-of-Features Acoustic Event Detection for Sensor Networks," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop*, Sep. 2016, pp. 55–59.

[6] K. Imoto and N. Ono, "Spatial Cepstrum as a Spatial Feature Using a Distributed Microphone Array for Acoustic Scene Analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1335–1343, Jun. 2017.

[7] K. Imoto, "Acoustic Scene Classification Using Multichannel Observation with Partially Missing Channels," in *Proceedings of European Signal Processing Conference*, Aug. 2021.

[8] Z. Yang, S. Guan, and X.-L. Zhang, "Deep Ad-hoc Beamforming Based on Speaker Extraction for Target-Dependent Speech Separation," *arXiv preprint arXiv:2021.00403*, Dec. 2020. [Online]. Available: http://arxiv.org/abs/2012.00403

[9] J. Chen and X.-L. Zhang, "Scaling Sparsemax Based Channel Selection for Speech Recognition with ad-hoc Microphone Arrays," in *Proc. Interspeech*, ISCA, Aug. 2021, pp. 291–295.

[10] C. Liang, J. Chen, S. Guan, and X.-L. Zhang, "Attention-based multi-channel speaker verification with ad-hoc microphone arrays," *arXiv preprint arXiv:2107.00178*, Jun. 2021. [Online]. Available: http://arxiv.org/abs/2107.00178

[11] R. Scheibler, D. Horiike, and N. Ono, "Blinkies: Sound-to-light conversion sensors and their application to speech enhancement and sound source localization," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, IEEE, Nov. 2018, pp. 1899–1904.

[12] R. Scheibler and N. Ono, "Multi-modal Blind Source Separation with Microphones and Blinkies," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2019, pp. 366–370.

[13] D. Horiike, R. Scheibler, Y. Wakabayashi, and N. Ono, "Blinkformer: Light-aided beamforming for multiple targets enhancement," in *Proceedings of IEEE International Workshop on Multimedia Signal Processing*. Sep. 2019, pp. 1–6.

[14] R. Scheibler and N. Ono, "Blinkies: Open Source Sound-to-Light Conversion Sensors for Large-Scale Acoustic Sensing and Applications," *IEEE Access*, vol. 8, pp. 67 603–67 616, 2020.

[15] D. Horiike, R. Scheibler, Y. Kinoshita, Y. Wakabayashi, and N. Ono, "Energy-Based Multiple Source Localization with Blinkies," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Dec. 2020, pp. 443–448.

[16] Y. Kinoshita and N. Ono, "End-to-End Training for Acoustic Scene Analysis with Distributed Sound-to-Light Conversion Devices," in *Proceedings of European Signal Processing Conference*, Aug. 2021.

[17] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of ACM MULTIMEDIA*, Nov. 2005, pp. 399–402.

[18] O. Ronneberger, P.Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9351. Springer, Nov. 2015, pp. 234–241.

[19] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint arXiv:1409.1556*, Sep. 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[20] G. Dekkers, P. Karsmakers, and L. Vuegen, "Monitoring of domestic activities based on multi-channel acoustics," 2018. [Online]. Available: http://dcase.community/challenge2018/task-monitoring-domestic-activities

[21] G. Dekkers and P. Karsmakers, "DCASE 2018, Task 5: Monitoring of domestic activities based on multi-channel acoustics - Development dataset," 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1247102

[22] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, T. van Waterschoot, B. Vanrumste, M. Verhelst, and P. Karsmakers, "The {SINS} Database for Detection of Daily Activities in a Home Environment Using an Acoustic Sensor Network," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop*, 2017, pp. 32–36.

[23] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, pp. 1–15, Dec. 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *Proceedings of IEEE International Conference on Computer Vision*, Santiago, Dec. 2015, pp. 1026–1034.