TOWARDS GENERATING AMBISONICS USING AUDIO-VISUAL CUE FOR VIRTUAL REALITY

Aakanksha Rana, Cagri Ozcinar*, and Aljosa Smolic*

V-SENSE, School of Computer Science and Statistics, Trinity College Dublin, Ireland.

ABSTRACT

Ambisonics *i.e.*, a full-sphere surround sound, is quintessential with 360° visual content to provide a realistic virtual reality (VR) experience. While 360° visual content capture gained a tremendous boost recently, the estimation of corresponding spatial sound is still challenging due to the required sound-field microphones or information about the sound-source locations. In this paper, we introduce a novel problem of generating Ambisonics in 360° videos using the audio-visual cue. With this aim, firstly, a novel 360° audio-visual video dataset of 265 videos is introduced with annotated sound-source locations. Secondly, a pipeline is designed for an automatic Ambisonic estimation problem. Benefiting from the deep learning based audio-visual feature-embedding and prediction modules, our pipeline estimates the 3D sound-source locations and further use such locations to encode to the B-format. To benchmark our dataset and pipeline, we additionally propose evaluation criteria to investigate the performance using different 360° input representations. Our results demonstrate the efficacy of the proposed pipeline and open up a new area of research in 360° audio-visual analysis for future investigations.

***Index Terms*—** Virtual Reality, 360° video, Spatial sound, Ambisonics, Multi-model, Deep learning.

1. INTRODUCTION

Recent advancements in virtual reality (VR) technologies have paved the way of capturing and sharing omnidirectional videos (ODVs) over social media. ODV, also known as 360° video, provides the visual representation of the 360° surrounding of the captured scene. This emerging representation can be navigated with three degrees of freedom by rotating and changing the viewing direction of VR devices (*e.g.*, tablet, laptop, head-mounted display). Users can nowadays, easily capture the 360° content with the help of affordable 360° cameras available in the market (*e.g.*, Ricoh Theta [1] and Samsung Gear 360 [2]) and share their ODVs over social networks to engage viewers deeply.

Essentially, creating realistic VR experiences requires the ODVs to be captured with their spatial audios. The spatial aspect of sound plays a significant role in informing the viewers about the location of objects in the 360° environment, providing an immersive multimedia experience. In practice, however, existing affordable 360° cameras can capture the visual scene either with mono or stereo audio signals. As a consequence, such audio-visual content is incapable of creating a magical sense of "being there" in the VR environment.

Recent user studies have amplified the need for spatial audio to achieve presence in VR films [3–5]. A spatial audio signal (also referred to as 3D audio or 360 audio) is considered as a powerful way of directing viewers' attention [6, 7], however, requiring expensive sound-field microphones, professional sound recording and production tools [8, 9].

The lack of spatial audio captured by affordable 360° cameras poses an interesting alternative audio-visual research topic in multimedia signal processing community – *given ODV with a mono/stereo audio signal, can we create the spatial audio to be used for VR video applications?*.

In this work, we introduce a novel problem of generating Ambisonics from the mono/stereo audio signal based on the audio-visual cue. Ambisonics is an effective way of representing the spatial audio and providing 3D sound for VR applications [10]. The Ambisonic technology is based on the spherical harmonic decomposition of the sound field and can encode the wave equation in the spherical coordinate system (r, Φ, θ). In this context, r is the distance to the source point from the center of coordinates, $\Phi \in (-\pi, \pi]$ and $\theta \in [-\pi/2, \pi/2]$. This direction of the sound can be encoded into four different channels (W, X, Y, Z) of the B-format [11], which is the basis for the first order Ambisonics. Hence, the location of a sound source on the sphere is required to be known to generate Ambisonics.

To tackle this problem, in this paper, we establish a well-annotated dataset and design a 4-stage pipeline to estimate the locations of sound sources. The dataset contains 265 ODVs with different audio-visual scenarios, such as round-table discussions, presentations, meetings, documentaries. In addition, we design a 4-stage pipeline to estimate the locations of sound sources on the sphere to generate the Ambisonics, where the 4-stages are namely, representation, feature embedding, prediction, and encoding. Our proposed pipeline adopts the deep learning based audio-visual feature embeddings and prediction techniques to facilitate the 3D localization of sound source using different ODV input representations. Also, we benchmark our dataset by proposing evaluation criteria to investigate the performance of our pipeline.

In a nutshell, the main contributions of this paper are threefold. First, we address the problem of automatic spatial audio estimation based on audio-visual cue as a first work. Second, we establish the first 360° Audio-Visual Dataset (360AVD) which contains 265 video clips with a well-annotated ground-truth providing the sound direction and location. Third, we propose evaluation criteria and perform preliminary quantitative and qualitative analysis using two widely used ODV projection techniques and state-of-the-art feature embedding and prediction algorithms to estimate the location of the sound source. We expect that releasing this dataset and addressing this novel research question will foster further research in multimedia signal processing area.

The rest of this paper is organized as follows. Section 2 presents the related work on audio-visual machine learning and generating spatial audio. The proposed audio-visual dataset and the proposed pipeline are described in Section 3 and 4. Section 5 presents the

*These authors contributed equally to this work.

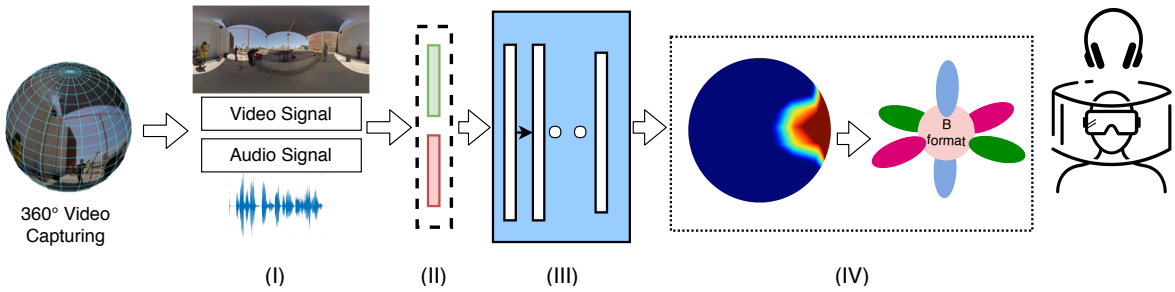


Fig. 1: Generation of a full-sphere surround sound environment in ODVs using audio-visual embeddings. The four modules of the pipeline are (I) Representation, (II) Feature embedding, (III) Prediction and (IV) Ambisonics encoding.

experiments with the used metrics. Our conclusions are drawn in Section 6.

2. RELATED WORK

Recent works showed that the location of a sound source could be estimated based on audio-visual signals. For instance, Owens *et al.* [12] modeled the visual and auditory signals and predicted the sound source pixel location on a given traditional 2D video. Similarly, Tian *et al.* [13] proposed audio-guided visual attention mechanism to explore audio-visual correlation with a target to predict event localization in traditional 2D videos. Also, Ephrat *et al.* [14] presented a deep network-based model that incorporate audio-visual cue to extract each speaker sound from a mixture of sounds. Audio-visual salient event detection was studied in [15] based on visual, audio and text modalities. The work showed that the performance of visual saliency estimation could be improved by incorporating audio and text. Coleman *et al.* [16] showed that an object-based audio capturing could achieve a convincing 3D audio experience over headphones. Furthermore, in [17], source separation system was presented for high order Ambisonics recording. A multi-channel spatial filter was derived based on the long short-term memory recurrent neural network with an assumption of known the directions of arrival of the directional sound sources.

However, to the best of our knowledge, no research on Ambisonics generation based on the audio-visual cue of ODV content currently exists.

3. DATASET

Dataset is fundamental in predicting the sound location on ODV. However, to the best of our knowledge, there is no publicly available dataset suitable for our objective. To this end, we present the first 360AVD which contains 265 video clips with a well-annotated ground-truth providing the sound direction and location. The dataset has been prepared using publicly available YouTube 360° unlabeled videos. Each clip in the dataset is of 10 seconds where all the audio sources are manually annotated per second. To annotate the audio source locations, we used Microsoft's visual object tagging tool[1] which supports labeling of multiple pixel locations for each second of a given video content. The dataset contains a different range of categories: presentation, documentary, debates and casual discussions. Sample scenes from the proposed dataset are presented in Figure 2. Our proposed dataset with our source codes are available at `https://github.com/V-Sense/360AudioVisual`.

Fig. 2: Sample scenes from the proposed 360AVD.

4. PROPOSED SYSTEM

Figure 1 illustrates the proposed pipeline for the generation of a full-sphere surround sound environment in ODV. The proposed system consists of four main modules, namely (I) input representations, (II) feature embedding, (III) prediction models, and (IV) Ambisonics encoding. At the first module, our aim is to investigate the impact of different sphere-to-planar projections for ODV. At the second module, we jointly estimate features for audio and video signals. Then, the combined information from the video and audio signals is fed into the prediction module to predict the 3D sound source location of a given ODV. Finally, we generate first-order Ambisonics by including the direction of the sound and encoding the estimated multi-channel audio based on the B-format.

We detail each module of the pipeline in the following subsections.

[1]Visual Object Tagging Tool: An electron app for building end to end Object Detection Models from Images and Videos: `https://github.com/Microsoft/VoTT`

4.1. Input Representations

Fig. 3: Visual representation of equirectangular and cubemap projections for the captured 360° video.

We investigate the performance of two widely used ODV representations, equirectangular and cubemap (or cubical), as shown in Figure 3. The first, an equirectangular projection, is the most straightforward format that represents a spherical object on a 2D planar surface. The second, a cubic projection, is a collection of six cube faces which are utilized to fill the whole sphere around the viewer. The first projection, however, contains less geometrical distortions than the second one.

4.2. Feature Embedding

Combining visual [18] and audio embedding using deep learning techniques have been recently studied in the literature [12, 13] for several end-to-end application-based scenarios such as source separation, action recognition, and audio-visual alignment. Most of the works are designed for traditional 2D video with a mono/stereo audio signal. However, such models combining the visual and sound information have not been studied for 360° video content.

Motivated from [12, 13], we formulate the feature embedding and prediction modules of the proposed Ambisonics generation pipeline. For the feature embedding module, we firstly employ the pre-trained VGG-19 [19] network to compute the visual features using the selected 360° visual representation *i.e.*, equirectangular or cubical projections. For each second, similar to [13], we compute the feature maps from 15 frames and average them to obtain one feature map, v, of the dimension of $7 \times 7 \times 512$. For each second of audio signal, we simultaneously extract the 128 dimensional audio representation, a, using a pre-trained VGG-like network [20]. Then, we finally feed the extracted feature embedding to the following prediction module to obtain the sound source probability maps, as shown in Figure 1.

4.3. Prediction Module

To predict the location of the sound source, we first adopt the sub-networks of the proposed deep networks, Tian *et al.* [13] and Owens *et al.* [12], originally designed for traditional 2D videos. We then alter them for our task at hand *i.e.*, to predict the 3D volumetric maps. Both models are recently published and their models are publicly available. We used the middle-layers from these pre-trained models to obtain the sound source location maps. The two prediction modules of our pipeline are named as a self-supervised module (SsM) and attention module (Att).

SsM module: It is an adaptation of the fusion sub-network proposed in [12] with three convolutional layers. To predict the sound localization map S_p^{SsM}, we concatenate the feature embedding and feed it to the convolutional layers, and finally apply the spherical mapping function f over the probability estimation map given as:

$$S_p^{SsM} = f(\sigma(\mathcal{L}^T conv_l)), \tag{1}$$

where σ is the sigmoid function, $\mathcal{L}$ is the affine layer, $conv_l$ is the last convolutional layer and function f maps $(x, y) \rightarrow (\theta, \phi)$ coordinates.

Att module: It is inspired from attention mechanism detailed in [13], adaptively learns to locate the visible regions in each second of the video from where the sound originates. To predict 3D sound for each second, localization maps S_p^{Att} is mathematically defined as:

$$S_p^{Att} = f(softmax(\omega \cdot \rho(l_v) + l_a)), \tag{2}$$

where the ρ_r is a hyperbolic tangent function, ω is a weighting parameter and l_a and l_v are the audio-visual transformation layers. Altogether, the attention weight vector computed using the multi-layer perception (MLP) like formulation as detailed in [13], is finally transformed by applying function f.

4.4. Ambisonics Encoding

After the sound localization maps S is estimated, we encode localized sound sources to the B-format. For this, the location of the i-th sound source is first estimated as follows:

$$\tilde{\Phi}_i, \tilde{\theta}_i = \mathcal{C} \tag{3}$$

where $\mathcal{C}$ is the set of a sound source location based on 3D coordinates, $\{C\}_{i=1}^{N}$, N is the number of sound sources, $\tilde{\Phi}_i$ and $\tilde{\theta}_i$ are the predicted spherical locations of the i-th sound source. The center of sound source C_i is the mean location of distribution of 3D point in the i-th sound source probability volumes S. The spherical volumes S are obtained by an absolute threshold, *i.e.*, for all coordinates where $S(x, y, z) \leq \epsilon$, values are equated to 0. Hence, we encode the B-format as follows:

$$\begin{aligned} W(t) &= \sum_{i}^{N} s_i(t)/\sqrt{2}, \\ X(t) &= \sum_{i}^{N} s_i(t) cos\tilde{\Phi}_i cos\tilde{\theta}_i, \\ Y(t) &= \sum_{i}^{N} s_i(t) sin\tilde{\Phi}_i cos\tilde{\theta}_i, \\ Z(t) &= \sum_{i}^{N} s_i(t) sin\tilde{\theta}_i, \end{aligned} \tag{4}$$

where $s_i(t)$ is the i-th sound signal of a given ODV, and the set of four audio channels (W, X, Y, Z) form the estimated Ambisonics. The non-directional sound pressure level is represented as W, and three other channels, (X, Y, Z), are described as the position of the sound: front-to-back (X), side-to-side (Y), and up-to-down (Z).

5. EXPERIMENTS

This section describes the proposed metrics and preliminary results obtained from our proposed pipeline for Ambisonics generation.

5.1. Metrics

To evaluate the performance of the predicted sound source location quantitatively, we introduce two metrics, namely, 360 Sound Source Distance (360-SSD) and 360 overlap error (360-OvErr).

Models	360-SSD			360-OvErr		
	ϵ=0.6	0.5	0.4	0.6	0.5	0.4
SsM-Cubical	**0.71 ± 0.04**	0.72 ± 0.08	0.74 ± 0.06	**0.71 ± 0.06**	0.77 ± 0.05	0.82 ± 0.04
SsM-EquiR	0.75 ± 0.06	0.77 ± 0.09	0.79 ± 0.07	0.78 ± 0.07	0.84 ± 0.06	0.88 ± 0.08
Att-Cubical	0.72 ± 0.05	0.73 ± 0.05	0.74 ± 0.04	0.72 ± 0.05	0.74 ± 0.08	0.78 ± 0.08
Att-EquiR	0.76 ± 0.04	0.77 ± 0.08	0.78 ± 0.06	0.84 ± 0.06	0.85 ± 0.06	0.86 ± 0.06

Table 1: Quantitative Results on 360AVD Dataset. The scores are averaged on 265 ODVs for all models.

360-SSD: It estimates the Euclidean distance between the centre of the predicted i-th sound source, $C_i^p(x, y, z)$, and the centre of ground truth i-th sound source, $C_i^g(x, y, z)$, in the Cartesian coordinate system *i.e.*, $||C_i^g - C_i^p||$. The center of sound source C_i is defined in Section 4.3. For 360-SSD, all distances are normalized and the probability spheres have radius 0.5.

360-OvErr: This metric is based on the ratio of an intersection of the predicted and ground truth probability volumes to the union and, mathematically given as $1 - \frac{S_p \cap S_g}{S_g \cup S_p}$, where S_p is the predicted probability volumes and S_g belongs to the ground truth. This measure can be seen as a 3D variant of single object localization error proposed in [21].

5.2. Implementation

Feature embedding and prediction. We base on the VGG-19 [19] and VGG-like [20] network for feature embedding module and computing audio and visual features are trained on Imagnet [19] and Audioset [22] dataset. For our SsM module, we adopt the convolutional layers from [12] which are trained large scale dataset of 750,000 videos. Similarly, we adopt the transformation layers l_a and l_v in Eq. 2 from the model proposed in [13] trained on a large-scale AVE dataset. The followed training paradigm is similar to [12, 13].

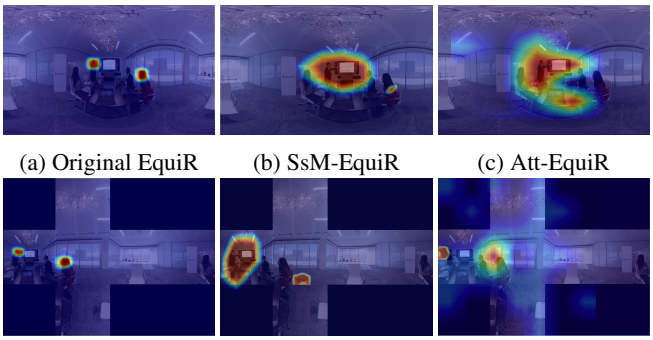
Ambisonic encoding. To encode Ambisonics, we used the Facebook 360 encoder tool from the Facebook Spatial Workstation [23]. Each predicted location for each sound was added to the location channels of B-format. Afterward, the MP4Box [24] was used to wrap the ODV and multi-channel audio together within an MP4 header file.

5.3. Results

In this section, we carried out the performance evaluation study of the proposed Ambisonics generation pipeline using both well-known representations and state-of-the-art prediction models over the proposed 360AVD dataset.

We first evaluated the performance of the predicted probability volumes, quantitatively, by using the proposed metrics: 360-SSD and 360-OvErr. For both metrics, less score stands for better performance. In Table 1, we present the results averaged over 265 ODVs with cubical and equirectangular formats and different settings of ϵ. Using both the metrics, we observe that SsM model-based prediction module with cubical ODV input representation performs the best in terms of localizing the exact sound source 3D location as well as the region. Att model-based prediction module competes closely with the former for localizing the 3D sound source.

For both models, we observe that cubical ODV representation provides better localization than the equirectangular format. This partly accounts to less distortions present in the cubical format, which in turn is favorable for distinctive feature embedding and prediction. On the other hand, the overall higher 360-OvErr with equirectangular format demonstrate the influence of the distortions where it leads to larger detected regions. This can be additionally seen in Fig. 4 where the qualitative performance of both models are illustrated with cubical and equirectangular representation. Additional results, *e.g.*, ODV with generated Ambisonics, are available on our Github page.



(a) Original EquiR (b) SsM-EquiR (c) Att-EquiR

(d) Original-Cubical (e) SsM-Cubical (f) Att-Cubical

Fig. 4: Qualitative Results: Row I shows the original 360° video frame with (a) overlay-ed ground truth, predicted results from (b) SsM and (c) Att modules in equirectangular representation. Row II shows the same frame with (d) overlay-ed ground truth, predicted results from (e) SsM and (f) Att modules in cubical representation.

6. CONCLUSION

This paper introduces a novel research problem of automatic Ambisonics generation from the mono/stereo audio signal based on audio-visual cue. For this aim, we propose a pipeline to predict the sound source location in a 3D space and time. The proposed pipeline contains four stages, representation, feature embedding, prediction, and Ambisonics encoding. To investigate the performance of each module, we introduce the first audio-visual dataset of 265 omnidirectional videos consisting of various single to multiple speech scenarios and evaluation metrics. Our initial analysis shows that the cubical representation of omnidirectional video with the self-supervised deep learning prediction algorithm performs better performance. All obtained results suggest that the problem of accurate sound source location estimation using audio-visual cue for Ambisonics remains open with a quite large room of improvement. The future work will consider developing optimal modules for our end-to-end pipeline for Ambisonics generations.

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number 15/RP/2776. We gratefully acknowledge the support of NVIDIA Corporation with the donated GPU used for this research. We gratefully acknowledge Dr. Enda Bates for all the insightful discussions and help.

References

[1] "RICOH THETA," https://theta360.com/en/, Accessed: 2018-10-27.

[2] "Samsung Gear 360," https://www.samsung.com/global/galaxy/gear-360/, Accessed: 2018-10-27.

[3] C. O Fearghail, C. Ozcinar, S. Knorr, and A. Smolic, "Director's cut - analysis of aspects of interactive storytelling for vr films," in *International Conference for Interactive Digital Storytelling (ICIDS) 2018*, 2018.

[4] S. Knorr, C. Ozcinar, C. O Fearghail, and A. Smolic, "Director's cut - a combined dataset for visual attention analysis in cinematic vr content," in *The 15th ACM SIGGRAPH European Conference on Visual Media Production (CVMP)*, 2018.

[5] C. Ozcinar and A. Smolic, "Visual attention in omnidirectional video for virtual reality applications," in *10th International Conference on Quality of Multimedia Experience (QoMEX)*, 2018.

[6] N. O'Dwyer, N. Johnson, R. Pagés, J. Ondřej, K. Amplianitis, E. Bates, D. Monaghan, and A. Smolić, "Beckett in vr: Exploring narrative using free viewpoint video," in *ACM SIGGRAPH 2018 Posters*, New York, NY, USA, 2018, SIGGRAPH '18, pp. 2:1–2:2, ACM.

[7] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein, "Saliency in vr: How do people explore virtual environments?," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 4, pp. 1633–1642, April 2018.

[8] M. Frank, F. Zotter, and A. Sontacchi, "Producing 3d audio in ambisonics," in *Audio Engineering Society Conference: 57th International Conference: The Future of Audio Entertainment Technology–Cinema, Television and the Internet*. Audio Engineering Society, 2015.

[9] H. Lim, C. Ozcinar, A. P. Hill, and A. Kondoz, "Sound localisation for 3d multimedia streaming," in *Proceedings of the 12th Western Pacific Acoustics Conference (WESPAC)*, 2015.

[10] E. Bates, "Comparing ambisonic microphones [blog]," https://endabates.wordpress.com/2017/06/19/comparing-ambisonic-microphones/, Accessed: 2018-10-27.

[11] M. A. Gerzon, "Ambisonics in multichannel broadcasting and video," *Journal of the Audio Engineering Society. Audio Engineering Society*, vol. 33, no. 11, pp. 859–871, 1985.

[12] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," *European Conference on Computer Vision (ECCV)*, 2018.

[13] T. Yapeng, S. Jing, Bochen L., D. Zhiyao, and X. Chenliang, "Audio-visual event localization in unconstrained videos," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[14] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics*, July 2018.

[15] P. Koutras, A. Zlatintsi, E. Iosif, A. Katsamanis, P. Maragos, and A. Potamianos, "Predicting audio-visual salient events based on visual, audio and text modalities for movie summarization," in *IEEE International Conference on Image Processing (ICIP)*, Sept 2015, pp. 4361–4365.

[16] P. Coleman, A. Franck, J Francombe, Q Liu, T. de Campos, R. J. Hughes, D. Menzies, M F S Gálvez, Y. Tang, J. Woodcock, P. J. B. Jackson, F. Melchior, C. Pike, F. M. Fazi, T. J. Cox, and A. Hilton, "An Audio-Visual system for Object-Based audio: From recording to listening," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 1919–1931, Aug. 2018.

[17] L. Perotin, R. Serizel, E. Vincent, and A. Gurin, "Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 36–40.

[18] A. Rana, J. Zepeda, and P. Pérez, "Feature Learning for the Image Retrieval Task," in *Computer Vision - FSLCV, ACCV 2014 - Singapore, November 1-2, 2014*, 2014, pp. 152–165.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Machine Learning*, 2015.

[20] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. Channing Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. W. Wilson, "CNN architectures for large-scale audio classification," *CoRR*, vol. abs/1609.09430, 2016.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, Dec. 2015.

[22] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017.

[23] "The Facebook 360 Spatial Workstation," https://facebook360.fb.com/spatial-workstation/, Accessed: 2018-10-27.

[24] J. Le Feuvre, C. Concolato, J.-C. Dufourd, R. Bouqueau, and J.-C. Moissinac, "Experimenting with multimedia advances using GPAC," in *Proceedings of the 19th ACM International Conference on Multimedia*, New York, NY, USA, 2011, MM '11, pp. 715–718, ACM.