Weakly Supervised Representation Learning for Audio-Visual Scene Analysis

Sanjeel Parekh, Slim Essid, Alexey Ozerov, Ngoc Q. K. Duong, Patrick Pérez, Gael Richard

Weakly Supervised Representation Learning for Audio-Visual Scene Analysis

Sanjeel Parekh, Slim Essid, Alexey Ozerov, *Senior Member, IEEE*, Ngoc Q.K. Duong, *Senior Member, IEEE*, Patrick Pérez, and Gaël Richard, *Fellow, IEEE*

***Abstract*—Audio-visual (AV) representation learning is an important task from the perspective of designing machines with the ability to understand complex events. To this end, we propose a novel multimodal framework that instantiates multiple instance learning. Specifically, we develop methods that identify events and localize corresponding AV cues in unconstrained videos. Importantly, this is done using weak labels where only video-level event labels are known without any information about their location in time.**

We show that the learnt representations are useful for performing several tasks such as event/object classification, audio event detection, audio source separation and visual object localization. An important feature of our method is its capacity to learn from unsynchronized audio-visual events. We also demonstrate our framework's ability to separate out the audio source of interest through a novel use of nonnegative matrix factorization. State-of-the-art classification results, with a F1-score of 65.0, are achieved on DCASE 2017 smart cars challenge data with promising generalization to diverse object types such as musical instruments. Visualizations of localized visual regions and audio segments substantiate our system's efficacy, especially when dealing with noisy situations where modality-specific cues appear asynchronously.

***Index Terms*—Multimodal classification, sound event detection, object localization, multiple instance learning, deep learning, audio-visual fusion**

I. Introduction

We are surrounded by events that can be perceived via distinct audio and visual cues. Be it a ringing phone or a car passing by, we instantly identify the audio-visual (AV) components that characterize these events. This remarkable ability helps us understand and interact with our environment. For building machines with such scene understanding capabilities, it is important to design algorithms for learning audio-visual representations from real-world data. This work is a step in that direction, where we aim to learn such representations through weak supervision.

Specifically, we are interested in designing a system that simultaneously tackles multiple related scene understanding tasks which include video event classification, spatial-temporal visual object localization and corresponding audio object enhancement and temporal localization. Obtaining precisely annotated data for doing so is an expensive endeavor, made even more challenging by multimodal considerations. The annotation process is not only error prone and time consuming but also subjective to an extent. Often, event boundaries in audio, extent of video objects or even their presence is ambiguous. Thus, we opt for a weakly-supervised learning approach using data with only video-level event labels, that is labels given for whole video documents without timing information.

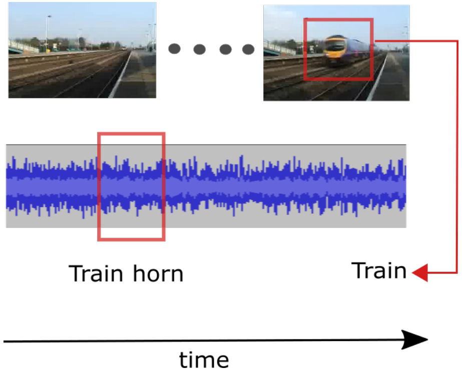

Fig. 1. **Pictorial representation of the problem**: Given a video labeled as "train horn", we would like to: (i) identify the event, (ii) localize both, its visual presence and the temporal segment(s) containing the characteristic sound, and (iii) segregate the characteristic audio cue from the background. Note that the train horn may sound before the train is visible. Our model can deal with such unsynchronized AV events.

To motivate our tasks and method, consider a video labeled as "train horn", depicted in Fig. 1. Assuming that the train is both visible and audible at some time in the video, in addition to identifying the event, we are interested in learning representations that help us answer the following:

- *Where is the visual object or context that distinguishes the event?* In this case it might be the train (object) or tracks, platform (context) *etc.* We are thus aiming for their spatio-temporal localization in the image sequence.
- *When does the sound event occur?* Here it is the train horn. We thus want to temporally localize the audio event.
- *How to enhance the audio object?* Here we are interested in audio source extraction *i.e.* segregating the source of interest from the background sounds.

The variety of noisy situations that one may encounter in unconstrained environments or videos adds to the difficulty of this very challenging problem. Apart from modality-specific noise such as visual clutter, lighting variations and low audio signal-to-noise ratio, in real-world scenarios the appearance of audio and visual elements characterizing the event are often unsynchronized in time. This is to say that the train horn may sound before or after the train is visible, as in previous example. In the extreme, not so rare case, the train may not

S. Parekh, S. Essid and G. Richard are with Telecom Paris, Paris, France
A. Ozerov and N. Duong are with InterDigital, Cesson Sevigne, France
P. Pérez is with Valeo.ai, Paris, France

appear at all. The latter is also commonly referred to as "off–screen" audio [1]. We are interested in designing a system to tackle the aforementioned questions and situations. It is important to mention here that certain challenges are implicit in the typical audio-visual properties of a scene or context. For instance, object sizes, lighting conditions and sound types would all vary depending upon whether we are on a street or in a concert hall. Although we aim to build a system that generalizes well to different contexts, any context-specific tuning for tackling extreme cases is not under consideration in the present study.

Prior research has utilized audio and visual modalities for classification and localization tasks in various contexts. Fusing modality-specific hand-crafted or deep features has been a popular approach for problems such as multimedia event detection and video concept classification [2]–[5]. In particular, these audio, appearance and motion features have been used for learning multimodal codebooks [2] or training modality-specific detectors [3]. Work on fusing intermediate representations for video categorization has also been carried out [4]. On the other hand, AV correlations have been utilized for localization and representation learning in general, through feature space transformation techniques such as canonical correlation analysis (CCA) [6], [7] or deep networks [8]–[12]. While [6], [7] tackle moving sounding object segmentation by using CCA to correlate audio and visual features, [12] aims to extend CCA by learning deep encodings for each modality such that their correlation is maximized. However, a unified multimodal framework for our task, that is learning data representations for simultaneously identifying real world events and extracting the AV cues depicting them, has not been extensively studied in previous works.

Contributions and outline

In this work, we present a complete AV event understanding framework where the modality-specific modules can be trained jointly to perform multiple tasks such as event/object classification, spatio-temporal visual localization, temporal audio localization and source separation. Key attributes and results of our approach are summarized below:

- We report state-of-the-art event classification performance on Detection and Classification of Acoustic Scenes and Events (DCASE) smart cars challenge data [13] and demonstrate usefulness of AV complementatrity. We also show results on Kinetics music instrument dataset [14] to validate our framework's application to diverse object types.
- To highlight flexibility provided by our modular design, we propose several task-specific instantiations. These include changes to allow detection of synchronously appearing AV cues and capability to enhance the audio source of interest. The audio source of interest is defined by the classes on which our system is trained, for *e.g.* a violin in a music concert mixture.
- Additionally, we also show encouraging qualitative visual localization results.

We begin by mentioning connections and distinctions with related works in Section II. This is followed by a description of the proposed framework and its instantiations for tackling classification and localization in Section III. Finally, we discuss our experimental setup in Section IV and validate the usefulness of the learnt representations for event classification, audio event detection, source separation and visual object localization in Section V.

II. Related work

To position our work, we discuss relevant literature that employs weakly supervised learning for visual object localization, audio event detection and source separation. We also delineate several distinctions between the present study and recent multimodal deep learning approaches.

A. Audio scene analysis

Detection and segregation of individual sources in a mixture is central to computational auditory scene analysis [15]. A significant amount of literature exists on supervised audio event detection (AED) [16]–[19]. However, progress with weakly labeled data in the audio domain has been relatively recent. An early work [20] showed the usefulness of multiple instance learning (MIL) techniques to audio using support vector machines (SVM) and neural networks. One of the first to demonstrate that it is indeed possible to perform temporal localization of audio events with reasonable performance using just weak labels.

The introduction of the weakly-labeled audio event detection task in the 2017 DCASE challenge [21][1], along with the release of Google's AudioSet data[2] [22], has led to accelerated progress in the recent past. AudioSet is a large-scale weakly-labeled dataset of audio events collected from YouTube videos. A subset of this data was used for the DCASE 2017 task on large-scale AED for smart cars.[3] Several submissions to the task utilized sophisticated deep architectures with attention units [23], as well as $\max$ and softmax operations [24]. Another recent study introduced a convolutional neural networks (CNN) with global segment-level pooling for dealing with weak labels [25]. It is worth noting that the field is growing rapidly. Concurrent and subsequent studies have greatly exploited the MIL and attention-based learning paradigm [26]–[28]. While we share with these works the high-level goal of weakly-supervised learning, we take a multimodal approach. Moreover, our audio sub-module design, as discussed in the next section, has crucial distinctions with past studies.

Audio source separation research in weakly supervised regime has followed a progress trend similar to the one witnessed in the AED domain. Recent works include several vision–inspired [29] and vision–guided [30]–[32] systems. In particular, nonnegative matrix factorization (NMF) is used within an MIL framework in [32]. Interestingly, the authors use a deep learning based approach for mapping NMF basis vectors to visual objects. Our proposed separation technique goes in this direction with several key differences discussed in Sec. III-D1.

[1] http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/

[2] https://research.google.com/audioset/

[3] http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-large-scale-sound-event-detection

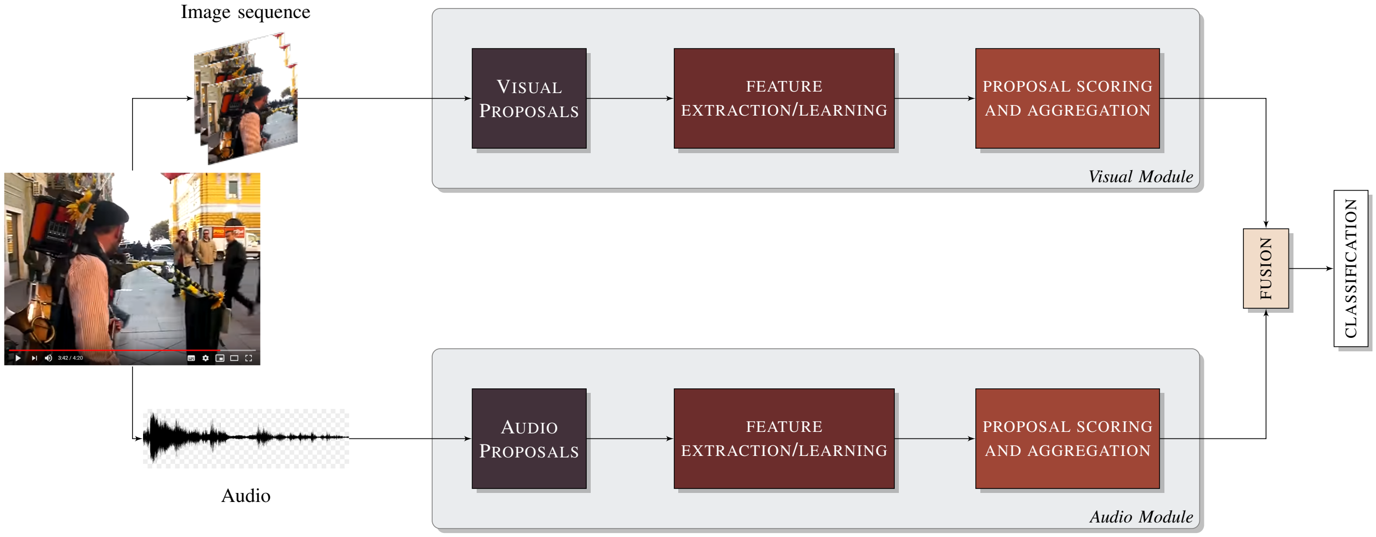
Image sequence
VISUAL PROPOSALS
FEATURE EXTRACTION/LEARNING
PROPOSAL SCORING AND AGGREGATION
Visual Module
FUSION
CLASSIFICATION
AUDIO PROPOSALS
FEATURE EXTRACTION/LEARNING
PROPOSAL SCORING AND AGGREGATION
Audio Module
Audio

Fig. 2. **High level view of the proposed approach**: Given a video captured using a single microphone and camera, we propose the depicted framework for weakly supervised representation learning.

B. *Visual object localization and classification*

There is a long history of works in computer vision applying weakly supervised learning for object localization and classification. MIL techniques have been extensively used for this purpose [33]–[39]. Typically, each image is represented as a set of regions. Positive images contain at least one region from the reference class while negative images contain none. Latent structured output methods, *e.g.*, based on SVMs [40] or conditional random fields (CRFs) [41], address this problem by alternating between object appearance model estimation and region selection. Some works have focused on better initialization and regularization strategies [39], [42], [43] for solving this non-convex optimization problem.

Owing to the exceptional success of CNNs in computer vision, recently, several approaches have looked to build upon CNN architectures for embedding MIL strategies. These include the introduction of operations such as $\max$ pooling over regions [35], global average pooling [38] and their soft versions [44]. Another line of research consists in CNN-based localization over class-agnostic region proposals [36], [37], [45] extracted using a state-of-the-art proposal generation algorithm such as EdgeBoxes [46], Selective Search [47], *etc.* These approaches are supported by the ability to extract fixed size feature maps from CNNs using region-of-interest [48] or spatial pyramid pooling [49]. Our work is related to such techniques. We build upon ideas from the two-stream architecture [37] for classification and localization.

State-of-the-art end-to-end object detection networks such as Faster RCNN [50] and its instance segmentation extension Mask RCNN [51] incorporate proposal generation as part of the system (region proposal network) instead of a separate stage. Nonetheless, these approaches require label annotations for different regions. It is also worth mentioning that some works have extended class-agnostic proposal generation from 2D images to video tube proposals for tasks such as action localization [52] and object detection [53]. However, these involve a computationally expensive pipeline preventing large-scale usage.

C. *Differences with recent AV deep learning studies*

We formulate the problem as a MIL task using class-agnostic proposals from both video frames and audio. This allows us to simultaneously solve the classification and localization problems. Finally, by construction, our framework deals with the difficult case of asynchronous AV events. This is significantly different from recent multimodal deep learning based studies on several counts: Contrary to prior works, where unsupervised representations are learnt through audio–image correlations (temporal co-occurrence), we adopt a weakly-supervised learning approach using event classes. Unlike [8], [9], [11], we focus on localizing discriminative audio and visual components for real-world events.

III. Proposed framework and its instantiations

The tasks under consideration can be naturally formulated as MIL problems [54]. MIL is typically applied to cases where labels are available over bags (sets of instances) instead of individual instances. The task then amounts to jointly selecting appropriate instances and estimating classifier parameters. In our case, a video can be seen as a labeled bag, containing a collection of visual and audio proposals. The term *proposal* refers to image or audio “parts” that may potentially constitute the object of interest. This step is at the core of our approach.

The key idea, as illustrated in Fig. 2, is to extract features from generated proposals and transform them for: (1) scoring each according to their relevance for class labels; (2) aggregating these scores in each modality and fusing them for video-level classification. This not only allows us to train both the sub-modules together through weak-supervision but also enables localization using the proposal relevance scores. Moreover, use of both the modalities with appropriate proposals makes the system robust against noisy scenarios. We present different task-specific variants of this general framework.

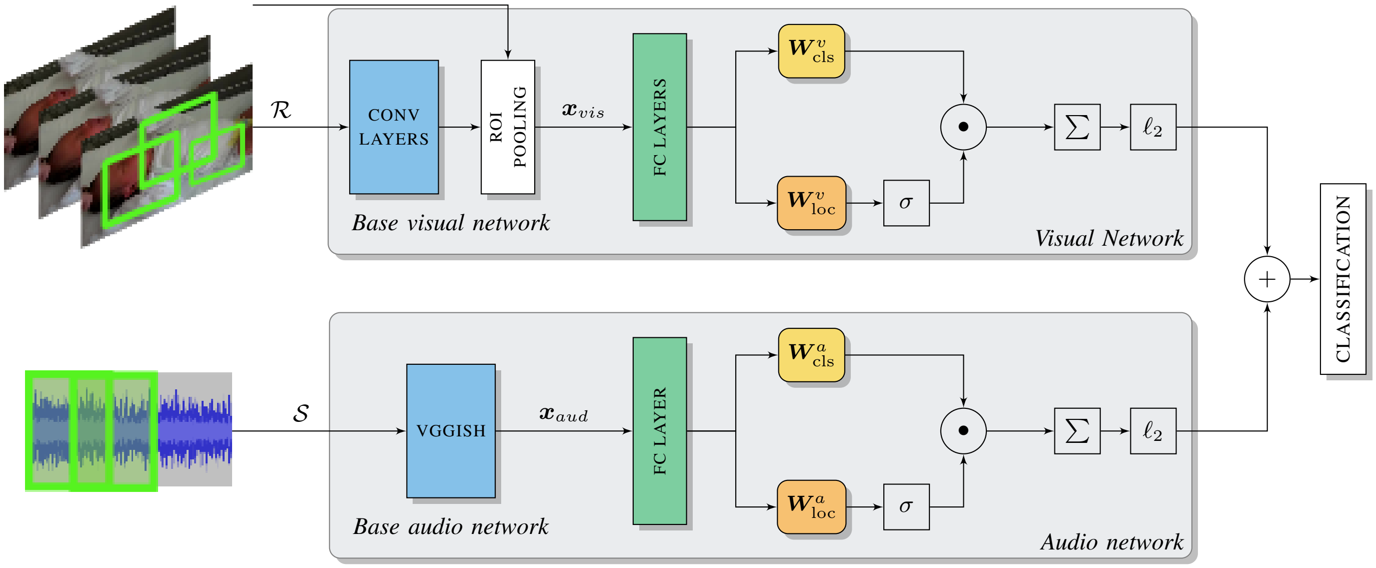

Fig. 3. **Module design**: Given a video, we consider the depicted pipeline for going from audio and visual proposals to localization and classification. Here $\boldsymbol{W}_{\text{cls}}$ and $\boldsymbol{W}_{\text{loc}}$ refer to the fully-connected classification and localization streams respectively; σ denotes softmax operation over proposals for each class, $\odot$ refers to element-wise multiplication; Σ to a summation over proposals and ℓ_2 to a normalization of scores. During training we freeze the weights of blocks denoted in blue.

We now formalize the design of each building block to specifically tackle event classification, visual object and audio event localization. An overview is provided in Fig. 3. A video, V is modeled as:

- a bag of M selected image regions, $\mathcal{R} = \{r_1, r_2, \ldots, r_M\}$, obtained from sub-sampled frames; and
- a set of S audio segments, $\mathcal{A} = \{a_1, a_2, \ldots, a_S\}$.

Given L such training examples, $\mathcal{V} = \{V^{(l)}\}_{l=1}^{L}$, organized into C classes, our goal is to learn a representation to jointly classify and localize image regions and audio segments that characterize a class. Each block from proposal generation to classification is discussed below in detail.

A. *Generating proposals and extracting features*

Visual Proposals. Generating proposals for object containing regions from images is at the heart of various visual object detection algorithms [55], [56]. As our goal is to spatially and temporally localize the most discriminative region pertaining to a class, we choose to apply this technique over sub-sampled video frame sequences. In particular, we sub-sample the extracted frame sequences of each video at a rate of 1 frame per second. This is followed by class-agnostic region proposal generation on the selected frames using EdgeBoxes [46]. This proposal generation method builds upon the insight that the number of contours entirely inside a box is indicative of the likelihood of an object's presence. Its use in our pipeline is motivated by experiments confirming better performance in terms of speed/accuracy tradeoffs over most competing techniques [57]. EdgeBoxes additionally generates a confidence score for each bounding box which reflects the box's "objectness". To reduce the computational load and redundancy, we use this score to select the top few proposals (denoted by M_{img}) from each sampled image and use them for feature extraction. Hence, given a 10 second video, the aforementioned procedure would leave us with a list of $M = 10 \times M_{\text{img}}$ region proposals.

A fixed-length feature vector, $\boldsymbol{x}_{vis}(r_m; V) \in \mathbb{R}^{d_v}$ is obtained from each image region proposal, r_m in V. Here d_v denotes the visual feature vector dimensionality. This computation is done using a convolutional neural network altered with a region-of-interest (RoI) pooling layer. An RoI layer works by computing fixed size feature maps (*e.g.* 6×6 for `caffenet` [58]) from regions of an image using max-pooling [48]. This helps to ensure compatibility between convolutional and fully connected layers of a network when using regions of varying sizes. Moreover, unlike Region-based CNN (RCNN) [56], where each individual region is processed, the shared computation for different regions of the same image using Fast-RCNN implementation [48] leads to faster processing. In Fig. 3 we refer to this Fast-RCNN feature extractor as the base visual network. In practice, feature vectors $\boldsymbol{x}_{vis}(\cdot)$ are extracted after RoI pooling layer and passed through two fully connected layers, which are fine-tuned during training. Typically, standard CNN architectures pre-trained on ImageNet [59] classification are used for the purpose of initializing network weights.

Audio Temporal Segment Proposals. Earlier works on audio indexing have directly used speech features such as mel-frequency cepstral coefficients (MFCCs) even though they are not particularly appropriate to describe general audio sounds [60]. Nowadays, the most popular audio signal representation is the log-Mel spectrogram, as demonstrated by the top-performing systems in the DCASE challenges [13]. For audio, we first represent the raw audio waveform as a log-Mel spectrogram [61]. Each proposal is then obtained by sliding a fixed-length window over the obtained spectrogram along the temporal axis. These are the so called audio temporal segment proposals, also referred to as Temporal Segment Proposals

(TSPs). The dimensions of this window are chosen to be compatible with the audio feature extractor. For our system we set the proposal window length to 960ms with a 50% overlap.

We use a VGG-style deep network known as `vggish` for base audio feature extraction. Inspired by the success of CNNs in visual object recognition Hershey *et al.* [62] introduced this state-of-the-art audio feature extractor as an audio parallel to networks pre-trained on ImageNet for classification. `vggish` has been pre-trained on a preliminary version of YouTube-8M [63] for audio classification based on video tags. It stacks 4 convolutional and 2 fully connected layers to generate a 128 dimensional embedding, $\boldsymbol{x}_{\text{aud}}(a_s; V) \in \mathbb{R}^{128}$ for each input log-Mel spectrogram segment $a_s \in \mathbb{R}^{96\times 64}$ with 64 Mel-bands and 96 temporal frames. Prior to proposal scoring, the generated embedding is passed through a fully-connected layer that is learnt from scratch.

B. *Proposal scoring network and fusion*

So far, we have extracted base features for each proposal in both the modalities and passed them through fully connected layers in their respective modules. Equipped with this transformed representation of each proposal, we use the two-stream architecture proposed by Bilen *et al.* [37] for scoring each of them with respect to the classes. There is one scoring network of the same architecture for each modality as depicted in Fig. 3. Thus, for notational convenience, we generically denote the set of audio or visual proposals for each video by $\mathcal{P}$ and let proposal representations before the scoring network be stacked in a matrix $\mathbf{Z} \in \mathbb{R}^{|\mathcal{P}|\times d}$, where d denotes the dimensionality of the audio/visual proposal representation.

The architecture of this module consists of parallel classification and localization streams. The former classifies each region by passing $\mathbf{Z}$ through a linear fully connected layer with weights $\boldsymbol{W}_{\text{cls}}$, giving a matrix $\mathbf{A} \in \mathbb{R}^{|\mathcal{P}|\times C}$. On the other hand, the localization layer passes the same input through another fully-connected layer with weights $\boldsymbol{W}_{\text{loc}}$. This is followed by a softmax operation over the resulting matrix $\mathbf{B} \in \mathbb{R}^{|\mathcal{P}|\times C}$ in the localization stream. The softmax operation on each element of $\mathbf{B}$ can be written as:

$$[\sigma(\mathbf{B})]_{pc} = \frac{e^{b_{pc}}}{\sum_{p'=1}^{|\mathcal{P}|} e^{b_{p'c}}},\ \forall (p,c) \in (1, |\mathcal{P}|) \times (1, C). \quad (1)$$

This allows the localization layer to choose the most relevant proposals for each class. Subsequently, the classification stream output is weighted by $\sigma(\mathbf{B})$ through element-wise multiplication: $\mathbf{E} = \mathbf{A} \odot \sigma(\mathbf{B})$. Class scores over the video are obtained by summing the resulting weighted scores in $\mathbf{E}$. Concurrent work by [64] discusses a similar MIL module for audio classification.

After performing the above stated operations for both audio and visual sub-modules, in the final step, the global video-level scores are ℓ_2 normalized and added. In preliminary experiments we found this to work better than addition of unnormalized scores. We hypothesize that the system trains better because ℓ_2 normalization ensures that the scores being added are in the same range.

C. *Classification loss and network training*

Given a set of L training videos and labels, $\{(V^{(l)}, \boldsymbol{y}^{(l)})\}_{l=1}^{L}$, we solve a multi-label classification problem. Here $\boldsymbol{y} \in \mathcal{Y} = \{-1, +1\}^C$ with the class presence denoted by +1 and absence by -1. To recall, for each video $V^{(l)}$, the network takes as input a set of image regions $\mathcal{R}^{(l)}$ and audio segments $\mathcal{A}^{(l)}$. After performing the described operations on each modality separately, the ℓ_2 normalized scores are added and represented by $\phi(V^{(l)}; \boldsymbol{w}) \in \mathbb{R}^C$, with all network weights and biases denoted by $\boldsymbol{w}$. All the weights, including and following the fully-connected layer processing stage for both the modalities, are included in $\boldsymbol{w}$. Note that both sub-modules are trained jointly.

The network is trained using the multi-label hinge loss on a batch of size B:

$$L(\boldsymbol{w}) = \frac{1}{CB}\sum_{l=1}^{B}\sum_{c=1}^{C} \max\Big(0, 1 - y_c^{(l)}\phi_c(V^{(l)}; \boldsymbol{w})\Big). \quad (2)$$

To summarize, we have discussed a general instantiation of our framework, capable of processing spatio-temporal visual regions, temporal audio segments for event classification and localizing characteristic proposal in each modality. Dealing with each proposal independent of the time at which it occurs allows tackling AV asynchronicity.

D. *Variants*

In the proposed framework (depiced in Fig. 2) module design can be flexibly modified in a task–specific manner. To demonstrate this, we discuss next two variants that allow performing audio source enhancement and synchronous AV fusion, respectively.

1) Source enhancement variant: Here we propose to design novel audio proposals using NMF with the goal of enhancing the audio source of interest. The primary reason for performing such a decomposition is the hope that each of the resulting spectral patterns would represent a part of just one source. Specifically, using NMF we decompose audio magnitude spectrograms $\mathbf{Q} \in \mathbb{R}_+^{F\times N}$ consisting of F frequency bins and N short-time Fourier transform (STFT) frames, such that,

$$\mathbf{Q} \approx \mathbf{W}\mathbf{H}, \quad (3)$$

where $\mathbf{W} \in \mathbb{R}_+^{F\times K}$and $\mathbf{H} \in \mathbb{R}_+^{K\times N}$ are nonnegative matrices that can be interpreted as the characteristic audio spectral patterns $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ and their temporal activations $\mathbf{H} \in \mathbb{R}_+^{K\times N}$, respectively. Here K is the total number of spectral patterns.

We then apply soft mask based filtering [65] to an audio recording to decompose it into K tracks (also referred to as NMF components) each obtained from $\mathbf{w}_k, \mathbf{h}_k$ for $k \in [1, K]$, where $\mathbf{w}_k$ and $\mathbf{h}_k$ denote spectral pattern and activation vectors corresponding to the k^{th} component, respectively. This is depicted in Fig. 4.

They can now be considered as proposals that may or may not belong to the class of interest. Specifically, we chunk each NMF component into temporal segments, which we call NMF Component proposals or NCPs. We denote the set of NCPs by $\mathcal{D} = \{d_{k,t}\}$, where each element is indexed by the component,

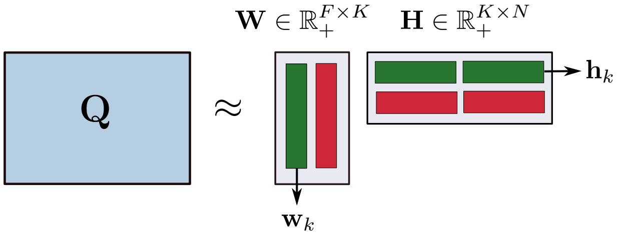
$\mathbf{W} \in \mathbb{R}_+^{F \times K}$
$\mathbf{H} \in \mathbb{R}_+^{K \times N}$
Q
≈
$\mathbf{h}_k$
$\mathbf{w}_k$

Fig. 4. NMF component proposals depiction where spectral patterns, $\mathbf{w}_k$ and corresponding activation vectors, $\mathbf{h}_k$ are shown in the same colour. Furthermore, each part in $\mathbf{h}_k$ refers to a non-overlapping temporal segment.

$k \in [1, K]$ and temporal segment $t \in [1, T]$. The same audio network is used for both TSPs and NCPs. Thus, for each NMF component or track we follow the TSP computation procedure. However, this is done with a non-overlapping window for reducing computational load.

Our system scores each NMF component, $d_{k,t}$ with respect to its relevance for a particular class. These relevance scores can be appropriately aggregated to perform source enhancement. We proceed as follows:

- Denoting by $\beta_{k,t}$ the score for k^{th} component's t^{th} temporal segment, we compute a global score for each component as
$$\alpha_k = \max_{t \in T} \ \beta_{k,t}.$$
It is worth mentioning that other pooling strategies such as mean or weighted rank pooling [44] could also be considered instead of the max operation. However, in our preliminary experiments we found them to yield similar results.
- Next, we apply min-max scaling between [0,1]:
$$\alpha'_k = \frac{\alpha_k - \min(\alpha)}{\max(\alpha) - \min(\alpha)},$$
where $\alpha = (\alpha_1, \ldots, \alpha_K)$ is obtained in the previous step.
- This is followed by soft mask based source and noise spectrogram reconstruction using complex-valued mixture STFT $\mathbf{X}$. Note that we can optionally apply a hard threshold τ on α'_k to choose the top ranked components for the source. This amounts to replacing α'_k by the indicator function $\mathbf{1}[\alpha'_k \geq \tau]$ in the following reconstruction equations:
$$\mathbf{S} = \frac{\sum_{k=1}^{K} \alpha'_k \mathbf{w}_k \mathbf{h}_k}{\mathbf{WH}} \mathbf{X} \tag{4}$$
$$\mathbf{N} = \frac{\sum_{k=1}^{K} (1 - \alpha'_k) \mathbf{w}_k \mathbf{h}_k}{\mathbf{WH}} \mathbf{X} \tag{5}$$
Here $\mathbf{S}$ and $\mathbf{N}$ are the estimates of source of interest and of background noise, respectively. These can be converted back to the time domain using inverse STFT.

It is worth noting two key differences with the approach in [32]: (i) In [32] only the NMF basis vectors are used for training without their corresponding activations. Hence no temporal information is utilized. (ii) Unlike us, they perform a supervised dictionary construction step after training to decompose a test signal.

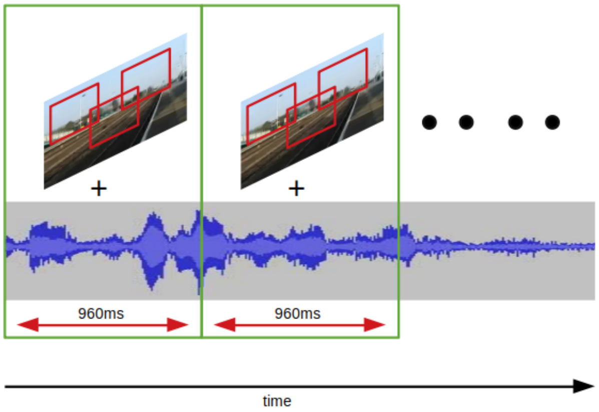

Fig. 5. Synchronized variant - herein audio and visual scores over each temporal segment are aggregated and the best temporal segment is chosen for classification.

2) Synchronous fusion variant: Framework instantiation depicted in Fig. 3 constructs the global score vector for each modality by combining scores over all the proposals, regardless of their temporal index. As noted, such a system is capable of dealing with asynchronous appearance of cues in both the modalities. On the other hand, we could envision a synchronized variant, where we only add scores of visual and audio proposals appearing in the same temporal segment. And construct the global score vector by choosing for each class the best scoring temporal segment. This is illustrated in Fig. 5. This essentially allows us to determine temporal segments where AV cues appear simultaneously. We list below specific changes made to the proposal score computation and fusion module:

1) Firstly, in the localization stream the softmax operation is performed over proposals from each temporal window separately. This amounts to replacing $|\mathcal{P}|$ by $|\mathcal{P}_t|$ in equation (1), where the proposals are indexed by the temporal segment they belong to. For the visual branch this corresponds to region proposals from a frame within the t^{th} temporal segment.
2) Secondly, after obtaining $\mathbf{E}$ *i.e.* the output of the two stream classification, we compute a class score vector for each temporal interval by summing up proposal scores separately over $p \in \mathcal{P}_t$. This gives us a matrix with dimensions $C \times T$ in each modality. Their addition gives us a synchronous AV temporal score.
3) Finally, for each class, the best AV temporal segment is chosen through a $\log - \text{sum} - \exp$ operation. This gives us the class score vector ϕ required for weakly–supervised training using multi–label hinge loss (refer to equation (2)).

IV. Setup and Datasets

A. Setup

All systems except that of [23], including variants, are implemented in Tensorflow. They were trained for 25K iterations using Adam optimizer [66] with a learning rate of 10^{-5} and a batch size of 24. We use the MATLAB implementation of

EdgeBoxes for generating region proposals, obtaining approximately 100 regions per video with $M_{\text{img}} = 10$ and a duration of 10 sec. The implementation is used with default parameter setting. Base visual features, $\boldsymbol{x}_{vis} \in \mathbb{R}^{9216}$ are extracted using `caffenet` [58] with pre-trained ImageNet weights and RoI pooling layer modification [48]. With 6×6 RoI pooling we get a 9216 ($= 256 \times 6 \times 6$) dimensional feature vector. For this, the Fast-RCNN Caffe implementation is used [48]. The fully connected layers, namely fc_6 and fc_7, each with 4096 neurons, are fine-tuned, with 50% dropout during training.

For audio, each recording is resampled to 16 kHz before processing. Log-Mel spectrum over the whole file is computed with a window size of 25ms and 10ms hop length. The resulting spectrum is chunked into segment proposals using a 960–ms window with a 480–ms stride.

For a 10–second recording, this yields 20 segments of size 96×64. We use the official Tensorflow implementation of `vggish`.[4]

B. Datasets

DCASE Smart Cars. We use the recently introduced dataset for the DCASE challenge on large-scale weakly supervised sound event detection for smart cars [21]. This is a subset of Audioset [22] which contains a collection of weakly-annotated unconstrained YouTube videos of vehicle and warning sounds spread over 17 classes. It is categorized as follows (abbreviations used in experiment tables are given in parenthesis that follow each category):

- *Warning sounds*: Train horn (trn-hrn), Air/Truck horn (air-hrn), Car alarm (car-alm), Reversing beeps (rv-bps), Ambulance siren (amb), Police car siren (pol-car), Fire engine/fire truck siren (f-eng), Civil defense siren (civ-def), Screaming (scrm).
- *Vehicle sounds*: Bicycle (bik), Skateboard (skt), Car (car), Car passing by (car-pby), Bus (bus), Truck (trk), Motorcycle (mbik), Train (trn).

This multi-label dataset contains 51,172 training, 488 validation and 1103 testing samples. Despite our best efforts, due to download issues, we were able to fetch 48,719 training, 462 validation and 1030 testing clips. It is worth mentioning that the training data is highly unbalanced with the number of samples for the classes ranging from 175 to 24K. To mitigate the negative effect of this imbalance on training, we introduce some balance by ensuring that each training batch contains at least one sample from some or all of the under-represented classes. Briefly, each batch is generated by first randomly sampling labels from a specific list, followed by fetching examples corresponding to the number of times each label is sampled. This list is generated by ensuring higher but limited presence of classes with more examples. We use a publicly available implementation for this purpose [23].[5]

Kinetics instruments (KI). We also use a subset of the Kinetics dataset [14] that contains 10-s YouTube videos from 15 music instrument classes. From a total of 10,267 videos, we create training and testing sets that contain 9199 and 1023 videos, respectively. KI is a multiclass dataset.

For source enhancement evaluation, we handpicked 45 “clean” instrument recordings, 3 per class. Due to their unconstrained nature, the audio recordings are mostly noisy, *i.e.* videos are either shot with accompanying music/instruments or in acoustic environments containing other background events. In that context, “clean” refers to solo instrument samples with minimal amount of such noise.

V. Experimental Validation and Analysis

In what follows, we thoroughly evaluate the proposed framework’s performance on various scene analysis tasks. In particular, we compare the asynchronous and synchronous variants of our system against several strong baselines for event classification on the DCASE smart cars benchmark. Generalization to diverse object types is shown through results on KI. This is followed by results for temporal localization of the audio event on DCASE. For completeness, we also present experiments on segregating the audio source of interest, as discussed in our prior work [67]. This allows us to demonstrate our system’s capability to perform good source enhancement while training just for weak label classification. This is done by utilizing NMF-based proposals as described in Sec. III-D1. We conclude this section with a discussion of qualitative visual localization examples that show how we deal with extreme noise, including asynchronous AV cues.

A. Event classification

Baselines. To our best knowledge, there is no prior work on deep architectures that perform the task of weakly supervised classification and localization for unsynchronized AV events. Our task and method are substantially different from recently proposed networks like L3 [10], [11] which are trained using synchronous AV pairs on a large collection of videos in a self-supervised manner. However, we designed several strong baselines for comparison and an ablation study. In particular, we compare against the following networks:

1) AV One-Stream Architecture: Applying MIL in a straight-forward manner, we could proceed only with a single stream. That is, we can use the classification stream followed by a max operation for selecting the highest scoring regions and segments for obtaining global video-level scores. As done in [37], we choose to implement this as a multimodal MIL-based baseline. We replace the $\max$ operation by the $\log-\text{sum}-\exp$ operator, its soft approximation. This has been shown to yield better results [34]. The scores on both the streams are ℓ_2 normalized before addition for classification. This essentially amounts to removing from Fig. 3 the localization branches and replacing the summation over proposals with the soft-maximum operation described above. To avoid any confusion, please note that we use the term ‘stream’ to refer to classification and localization parts of the scoring network.

[4] https://github.com/tensorflow/models/tree/master/research/audioset

[5] https://github.com/yongxuUSTC/dcase2017_task4_cvssp/blob/master/data_generator.py

2) Visual-Only (VO) and Audio-Only (AO) Networks: These networks only utilize one of the modalities for classification. However, note that there are still two streams for classification and localization, respectively. For a fair comparison and ablation study we train these networks with ℓ_2 normalization. In addition, for completeness we also implement Bilen *et al.*'s architecture for weakly supervised deep detection networks (WSDDN) with an additional softmax on the classification stream. As the scores are in the range [0,1], we train this particular network with C binary log-loss terms [37]. When discussing results we refer to this system as WSDDN-Type.

3) CVSSP Audio-Only [23]: This state-of-the-art method is the DCASE 2017 challenge winner for the audio event classification sub-task. The system is based on Gated convolutional RNN (CRNN) for better temporal modeling and attention-based localization. They use no external data and training/evaluation is carried out on all the samples. We present results for both their winning fusion system, which combines prediction of various models and Gated-RCNN model trained with log-Mel spectrum.

Results and discussion. We show in Table I the micro-averaged F1 scores for each of the systems described in this paper. The term micro-averaging implies that the F1 score is computed using a global count of total true positives (TP), false negatives (FN) and false positives (FP). Systems (a)-(b) in Table I are the proposed asynchronous and synchronous AV systems respectively and (c)-(f) present variants of (a) which are also treated as baselines, (g)-(h) denote results from CVSSP team [23], winners of the DCASE AED for smart cars audio event tagging task. The proposed systems and their variants are trained with audio temporal segment proposals only. Our proposed two stream multimodal and audio-only systems (a,b,c) outperform all the other approaches by a significant margin. Among the multimodal systems, the two-stream architecture performs much better than the one-stream counter-part, designed with only a classification stream and soft-maximum for region selection. On the other hand, the state-of-the-art CVSSP fusion system, which combines predictions of various models, achieves a better precision than the other methods. It is also worth mentioning that performance of the sync. AV system (b) is lower than the unsynchronized one (a). This is expected as the dataset contains some samples with asynchronously appearing cues. However, the sync. system would still be useful for detecting temporal segments where the AV cues appear together. Several important and interesting observations can be made by looking at these results in conjunction with the class-wise F1–scores reported in Table II.

TABLE I
RESULTS ON DCASE SMART CARS TASK TEST SET. WE REPORT HERE THE MICRO-AVERAGED F1 SCORE, PRECISION AND RECALL VALUES AND COMPARE WITH STATE-OF-THE-ART. WE USE TS, OS AND FS AS ACRONYMS TO REFER TO TWO-STREAM, ONE-STREAM AND FUSION SYSTEM, RESPECTIVELY.

	System	F1	Precision	Recall
(a)	AV TS	**64.2**	59.7	**69.4**
(b)	Sync. AV TS	62.0	57.2	67.6
(c)	AO (Audio-Only) TS	57.3	53.2	62.0
(d)	VO (Visual-Only) TS	47.3	48.5	46.1
(e)	VO TS WSDDN-Type [37]	48.8	47.6	50.1
(f)	AV OS	55.3	50.4	61.2
(g)	CVSSP - FS [23]	55.6	**61.4**	50.8
(h)	CVSSP - Gated-CRNN-logMel [23]	54.2	58.9	50.2

Most importantly, the results emphasize the complementary role of visual and audio sub-modules for this task. To see this, we could categorize the data into two sets: (i) classes with clearly defined AV elements, for instance car, train, motorcycle; (ii) some warning sounds such as, *e.g.*, reverse beeping, screaming, air horn, where the visual object's presence is ambiguous. The class-wise results of the visual-only system are a clear indication of this split. Well-defined visual cues enhance the performance of the proposed multimodal system over audio-only approaches, as video frames carry vital information about the object. On the other hand, in the case of warning sounds, video frames alone are insufficient as evidenced by results for the visual-only system. In this case, the presence of audio assists the system in arriving at the correct prediction. The expected AV complementarity is clearly established through these results.

Note that for some warning sounds the CVSSP method achieves better results. In this regard, we believe better temporal modeling for our audio system could lead to further improvements. In fact, we currently operate with a coarse temporal window of 960ms, which might not be ideal for all audio events. RNNs could also be used for further improvements. We think such improvements are orthogonal and were not the focus of this study. We also observe that results for under-represented classes in the training data such as air horn and reversing beeps are relatively lower. This can possibly be mitigated through data augmentation strategies.

In Table III we report results for the case where all layers of `vggish` are fine-tuned (FT). For this, we remove the FC adaptation layer from the audio network (refer to Fig. 3). It is also worth noting that for these experiments, we reduced the batch size to one due to memory constraints. For DCASE data, which contains approximately 48K training samples, this results in significantly more number of variable updates. Thus, to avoid overfitting, we run the system for 10 epochs and report results with the model that gives the lowest validation error. As expected, fine-tuning `vggish` results in improved performance as the audio features are better adapted to the dataset. We also see competitive performance for instrument classification with KI, where the multimodal system performs better than audio alone. Please note that as KI is a multiclass dataset, we compute the predicted class by taking an argmax over the score vector and report the classification accuracy.

B. Audio temporal localization

We show the sound event detection performance on DCASE smart cars data in Table IV. Following DCASE evaluation

TABLE II
CLASS-WISE COMPARISON ON TEST SET USING F1 SCORES. CLASS ABBREVIATIONS ARE DETAILED IN SEC. IV-B

System	Vehicle Sounds								Warning Sounds								
	bik	bus	car	car-pby	mbik	skt	trn	trk	air-hrn	amb	car-alm	civ-def	f-eng	pol-car	rv-bps	scrm	trn-hrn
AV TS	**75.7**	54.9	75.0	**34.6**	**76.2**	78.6	82.0	**61.5**	40.0	64.7	53.9	80.4	64.4	49.2	36.6	81.1	47.1
Sync. AV TS	65.0	**55.6**	**75.7**	25.6	74.0	**80.5**	**85.1**	57.8	28.4	**65.7**	54.1	82.1	61.3	52.6	39.6	70.6	48.8
AO TS	42.1	38.8	69.8	29.6	68.9	64.9	78.5	44.0	40.4	58.2	53.0	79.6	61.0	51.4	42.9	72.1	46.9
VO TS	72.5	52.0	61.2	15.0	54.1	64.2	73.3	49.7	12.0	33.9	13.5	68.6	46.5	19.8	21.8	44.1	32.1
AV OS	68.2	53.6	74.1	25.6	67.1	74.4	82.8	52.8	28.0	54.7	20.6	76.6	60.4	56.3	18.8	49.4	36.2
CVSSP - FS	40.5	39.7	72.9	27.1	63.5	74.5	79.2	52.3	**63.7**	35.6	**72.9**	**86.4**	**65.7**	**63.8**	**60.3**	**91.2**	**73.6**

TABLE III
RESULTS ON DCASE AND KI WITH FINE TUNED (FT) VGGISH

		DCASE		KI
Systems	F1	Precision	Recall	Accuracy
AV TS - VGGISH FT	65.0	64.9	65.0	84.5
AO TS - VGGISH FT	61.7	61.5	61.9	75.3

protocol, here we report segment–wise aggregated F1 score and error rate (ER) for each system. The official metric, ER, computes total number of substitution, deletion and insertion errors by comparing the ground truth and estimated output using one second long sub–segments [13].

The results for the proposed systems are computed by simply thresholding the two–stream output from the audio sub–module at $\tau = 0$ for the predicted label(s). We note that the results are comparable with the best performing CVSSP system. Note that the winning system for this subtask from Lee *et al.* [68] employs an ensemble method to optimally weigh multiple learned models, using ER as the performance metric to make the final selection. No such fine tuning is performed in our case.

TABLE IV
F1 SCORE AND ERROR RATE FOR SOUND EVENT DETECTION TASK

System	F1	ER
AV TS	51.0	0.76
AO TS	48.5	0.78
AV TS - VGGISH FT	52.3	0.74
AO TS - VGGISH FT	53.0	0.75
CVSSP - FS [23]	51.8	0.73
CVSSP - Gated-CRNN-logMel [23]	47.5	0.78
SNU - Ensemble method [68]	**55.5**	**0.66**

C. Audio source enhancement

Systems. We evaluate audio-visual (V + A) systems with different audio proposal types, namely:

- A (NCP): NMF component proposals,
- A (TSP, NCP): all TSPs and NCPs are put together into the same bag and fed to the audio network.

`vggish` is fine-tuned (as discussed earlier) for the systems listed above to adapt to NCP input.

Baselines. We compare with the following NMF related methods:

- Supervised NMF [69]: We use the class labels to train separate dictionaries of size 100 for each music instrument with stochastic mini-batch updates. At test time, depending on the label, the mixture is projected onto the appropriate dictionary for source reconstruction.
- NMF Mel-Clustering [70]: This blind audio-only method reconstructs source and noise signals by clustering mel-spectra of NMF components. We take help of the example code provided online for implementation in MATLAB [71].

Testing protocol. We corrupt the original audio with background noise corresponding to recordings of environments such as bus, busy street, park, etc. using one audio file per scene from the DCASE 2013 scene classification dataset [72]. The system can be utilized in two modes: *label known* and *label unknown*. For the former, where the source of interest is known, we simply use the proposal ranking given by the corresponding classifier for reconstruction. For the latter, the system's classification output is used to infer the source.

Results and discussion. We report, in Table V, average Source to Distortion Ratio (SDR) [73] over 450 audio mixtures created by mixing each of the 45 clean samples from the dataset with 10 noisy audio scenes. SDR, a popular metric for evaluating source separation systems is given by:

$$\mathrm{SDR} := 10 \log_{10} \frac{s_{\text{target}}}{e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}}, \qquad (6)$$

where s_{target} is the projection of estimated source onto the reference source signal and $e_{\text{interf}}, e_{\text{noise}}, e_{\text{artif}}$ are the interference, noise and artifact error terms, respectively. We refer the reader to [73] for more details. The results look promising but not state-of-the-art. This performance gap can be explained by noting that the audio network is trained for the task of audio event detection and thus does not yield optimal performance for source enhancement. The network focuses on discriminative components, failing to separate some source components from the noise by a larger margin, possibly requiring adaptive thresholding for best results. In other words, as the component scores vary for each example, a single threshold for all cases proves to be sub-optimal. It is worth noting that performance for the proposed systems does not degrade when used in "Label Unknown" mode, indicating that despite incorrect classification the system is able to cluster acoustically similar sounds. Performance of supervised NMF seems to suffer due to training on a noisy dataset. Separation results on in-the-wild YouTube videos are made available on our companion website.[6]

TABLE V
AVERAGE SDR OVER MIXTURES CREATED BY COMBINING CLEAN INSTRUMENT EXAMPLES WITH ENVIRONMENTAL SCENES.

System	Label Known	Label Unknown
Supervised NMF	2.3	–
NMF Mel-Clustering	–	**4.3**
V + A (NCP), soft	3.3	3.3
V + A (NCP), $\tau = 0.1$	**3.8**	3.9
V + A (NCP), $\tau = 0.2$	3.6	3.6
V + A (NCP, TSP), soft	2.1	2.2

D. Qualitative visual localization

In Fig. 6 we present some visual localization results from both, the DCASE and KI dataset. We also present typical failure cases which include focus on discriminative regions, multiple instance grouping and object occlusion. Localization in extreme asynchronous conditions is also discussed in Fig. 7. In the first case **A**, the sound of a car's engine is heard in the first two seconds followed by music. The normalized audio localization heatmap at the bottom displays the scores assigned to each temporal audio segment, s_t by the car classifier. The video frames placed above are roughly aligned with the audio temporal axis to show the video frame at the instant when the car sounds and the point where the visual network localizes. The localization is displayed through a yellow bounding box. To better understand the system's output, we modulate the opacity of the bounding box according to the system's score for it. Higher the score, more visible the bounding box. As expected, we do not observe any yellow edges in the first frame. Clearly, there exists temporal asynchrony, where the system locks onto the car, much later, when it is completely visible. **B** depicts an example, where due to extreme lighting conditions the visual object is not visible. Here too, we localize the audio object and correctly predict the 'motorcycle' class.

For full videos and more such examples we refer the reader to our companion website.[6] Please note that while we have shown encouraging qualitative visual localization performance, quantitative analysis could not be performed due to lack of ground-truth annotation.

VI. CONCLUSION

Building upon ideas from multiple instance learning, we have proposed a modular deep AV scene understanding framework that can be trained jointly to perform several tasks simultaneously. Exploiting our method's modularity, we investigate several instantiations capable of dealing with unsynchronized AV cue appearance, determining synchronous temporal segments and segregating the audio into constituent sources. The latter is made possible through a novel use of NMF decomposition, where, unlike most earlier methods, we only use the given weak labels for training. We report state-of-the-art event classification performance on DCASE 2017 smart cars data along with promising results for spatio-temporal visual localization, audio event detection and source separation. The method generalizes well to diverse object types.

Experiments have also shown that a more accurate audio temporal modeling would be needed to better cope with situations where the visual modality is inefficient. We also note limitations of our model in dealing with very small visual objects (harmonica) and clutter. Since the model relies on generation of good proposals, considering more recent proposal generation strategies [50] could be a promising direction. Furthermore, we believe the presented method could benefit from appropriately incorporating several recent developments in feature and modality fusion [74], [75].

[6] http://bit.ly/2HEJbrl

REFERENCES

[1] J. Woodcock, W. J. Davies, T. J. Cox, and F. Melchior, "Categorization of broadcast audio objects in complex auditory scenes," Journal of the Audio Engineering Society, vol. 64, no. 6, pp. 380–394, 2016.

[2] W. Jiang, C. Cotton, S. F. Chang, D. Ellis, and A. Loui, "Short-term audiovisual atoms for generic video concept classification," in Proceedings of the 17th ACM International Conference on Multimedia. ACM, 2009, pp. 5–14.

[3] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. C. Loui, and J. Luo, "Large-scale multimodal semantic concept detection for consumer video," in Proceedings of the international workshop on Workshop on multimedia information retrieval. ACM, 2007, pp. 255–264.

[4] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang, "Exploiting feature and class relationships in video categorization with regularized deep neural networks," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 2, pp. 352–364, 2018.

[5] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah, "High-level event recognition in unconstrained videos," International journal of multimedia information retrieval, vol. 2, no. 2, pp. 73–101, 2013.

[6] H. Izadinia, I. Saleemi, and M. Shah, "Multimodal analysis for identification and segmentation of moving-sounding objects," IEEE Transactions on Multimedia, vol. 15, no. 2, pp. 378–390, Feb 2013.

[7] E. Kidron, Y. Schechner, and M. Elad, "Pixels that sound," in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1, June 2005, pp. 88–95 vol. 1.

[8] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2405–2413.

[9] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," in Proc. of European Conference on Computer Vision. Springer, 2016, pp. 801–816.

[10] R. Arandjelović and A. Zisserman, "Look, listen and learn," in IEEE International Conference on Computer Vision, 2017.

[11] R. Arandjelovic and A. Zisserman, "Objects that sound," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 435–451.

[12] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in Proc. of International Conference on Machine Learning, 2013, pp. 1247–1255.

[13] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), November 2017, pp. 85–92.

[14] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev et al., "The kinetics human action video dataset," arXiv preprint arXiv:1705.06950, 2017.

[15] A. S. Bregman, Auditory scene analysis: The perceptual organization of sound. MIT press, 1994.

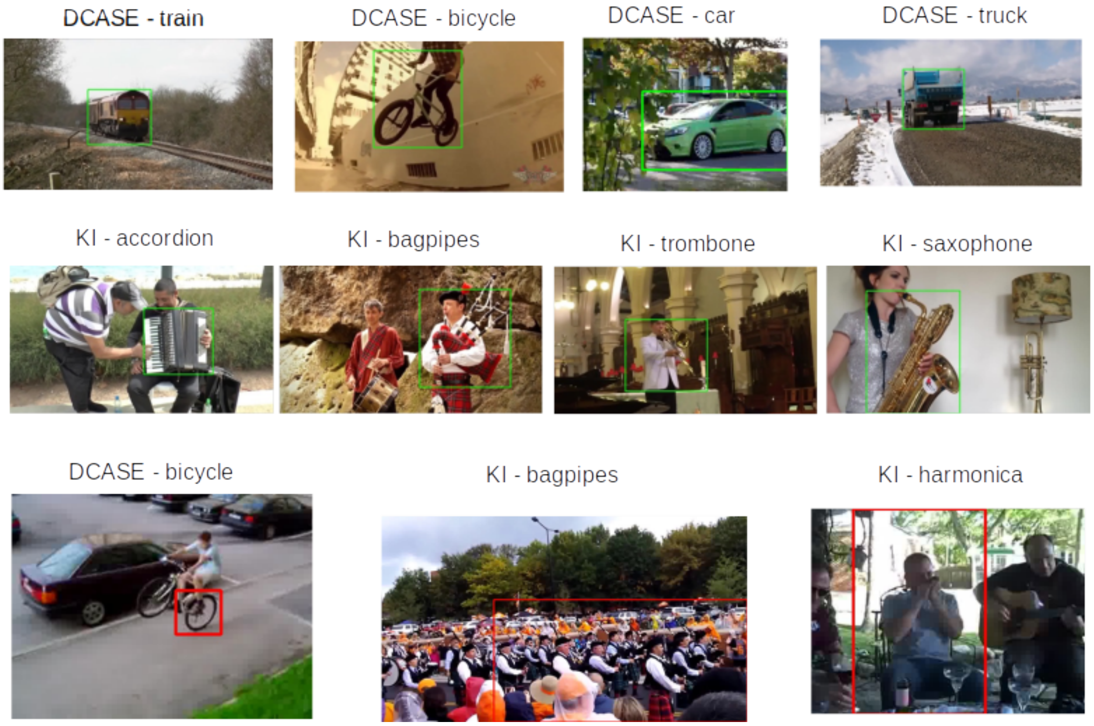

Fig. 6. **Visual localization on DCASE and KI test video frames**. Top row: correct localization for different vehicles (left to right: train, bicycle, car, truck). Middle row: correct localization for different instruments (left to right: accordion, bagpipes, trombone and saxophone). Maximum scoring bounding box shown in green. Bottom row: typical failure cases (in red) such as focus on discriminative regions (bicycle tyre), multiple object instance grouping (bagpipes) and small, occluded object (harmonica). More results on our companion website.[6]

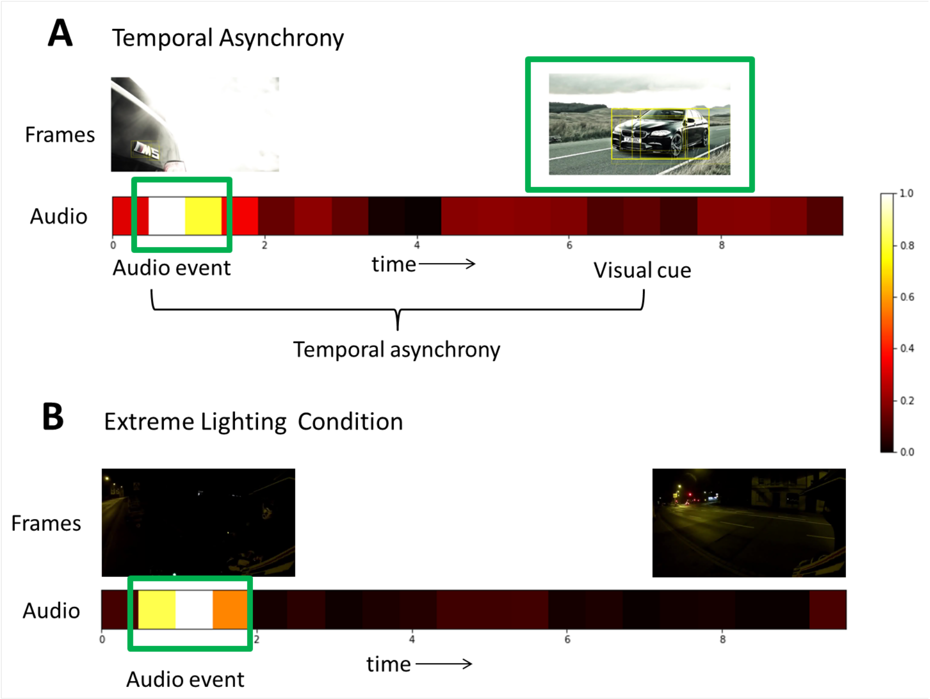

Fig. 7. **Qualitative results for unsynchronized AV events.** For both the cases A and B, the heatmap at the bottom denotes audio localization over segments for the class under consideration. For heatmap display, the audio localization vector has been scaled to lie between [0,1]. The top row depicts video frames roughly aligned to the audio temporal axis. (A) Top: Here we show a video where the visual object of interest appears after the audio event. This is a ‘car’ video from the validation split. The video frames show bounding boxes where edge opacity is controlled by the box’s detection score. In other words, higher score implies better visibility (B) Bottom: This is a case from the evaluation data where due to lighting conditions, the visual object is not visible. However the system correctly localizes in audio and predicts the ‘motorcycle’ class.

[16] A. Mesaros, T. Heittola, O. Dikmen, and T. Virtanen, “Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations,” in ICASSP. IEEE, 2015, pp. 151–155.

[17] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, “Real-world acoustic event detection,” Pattern Recognition Letters, vol. 31, no. 12, pp. 1543–1551, 2010.

[18] S. Adavanne, P. Pertilä, and T. Virtanen, “Sound event detection using spatial features and convolutional recurrent neural network,” in ICASSP. IEEE, 2017, pp. 771–775.

[19] V. Bisot, S. Essid, and G. Richard, “Overlapping sound event detection with supervised nonnegative matrix factorization,” in ICASSP. IEEE, 2017, pp. 31–35.

[20] A. Kumar and B. Raj, “Audio event detection using weakly labeled data,” in Proceedings of the 2016 ACM on Multimedia Conference. ACM, 2016, pp. 1038–1047.

[21] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, “DCASE2017 challenge setup: Tasks, datasets and baseline system,” in Proc. of Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), November 2017, pp. 85–92.

[22] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” in Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017, pp. 776–780.

[23] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, “Surrey-CVSSP system for DCASE2017 challenge task4,” DCASE2017 Challenge, Tech. Rep., September 2017.

[24] J. Salamon, B. McFee, and P. Li, “DCASE 2017 submission: Multiple instance learning for sound event detection,” DCASE2017 Challenge, Tech. Rep., September 2017.

[25] A. Kumar, M. Khadkevich, and C. Fugen, “Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes,” arXiv preprint arXiv:1711.01369, 2017.

[26] Q. Kong, C. Yu, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, “Weakly labelled audioset classification with attention neural networks,” arXiv preprint arXiv:1903.00765, 2019.

[27] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, “Look, listen, and learn more: Design choices for deep audio embeddings,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 3852–3856.

[28] Y. Wang, J. Li, and F. Metze, “A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 31–35.

[29] Q. Kong, Y. Xu, I. Sobieraj, W. Wang, and M. D. Plumbley, “Sound event detection and time–frequency segmentation from weakly labelled data,” IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 27, no. 4, pp. 777–787, 2019.

[30] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, “The sound of pixels,” in ECCV, September 2018.

[31] 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 Trans. Graph., vol. 37, no. 4, pp. 112:1–112:11, Jul. 2018.

[32] R. Gao, R. Feris, and K. Grauman, “Learning to separate object sounds by watching unlabeled video,” in ECCV, September 2018.

[33] C. Zhang, J. C. Platt, and P. A. Viola, “Multiple instance boosting for object detection,” in Advances in neural information processing systems, 2006, pp. 1417–1424.

[34] H. Bilen, M. Pedersoli, and T. Tuytelaars, “Weakly supervised object detection with posterior regularization,” in Proceedings BMVC 2014, 2014, pp. 1–12.

[35] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Is object localization for free?-weakly-supervised learning with convolutional neural networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 685–694.

[36] V. Kantorov, M. Oquab, M. Cho, and I. Laptev, “Contextlocnet: Context-aware deep network models for weakly supervised localization,” in European Conference on Computer Vision. Springer, 2016, pp. 350–365.

[37] H. Bilen and A. Vedaldi, “Weakly supervised deep detection networks,” in CVPR, 2016, pp. 2846–2854.

[38] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on. IEEE, 2016, pp. 2921–2929.

[39] R. G. Cinbis, J. Verbeek, and C. Schmid, “Weakly supervised object localization with multi-fold multiple instance learning,” IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 1, pp. 189–203, 2017.

[40] H. Bilen, V. P. Namboodiri, and L. J. Van Gool, “Object and action classification with latent window parameters,” International Journal of Computer Vision, vol. 106, no. 3, pp. 237–251, 2014.

[41] T. Deselaers, B. Alexe, and V. Ferrari, “Localizing objects while learning their appearance,” in European conference on computer vision. Springer, 2010, pp. 452–466.

[42] M. P. Kumar, B. Packer, and D. Koller, “Self-paced learning for latent variable models,” in Advances in Neural Information Processing Systems, 2010, pp. 1189–1197.

[43] H. O. Song, Y. J. Lee, S. Jegelka, and T. Darrell, “Weakly-supervised discovery of visual pattern configurations,” in Advances in Neural Information Processing Systems, 2014, pp. 1637–1645.

[44] A. Kolesnikov and C. H. Lampert, “Seed, expand and constrain: Three principles for weakly-supervised image segmentation,” in European Conference on Computer Vision. Springer, 2016, pp. 695–711.

[45] G. Gkioxari, R. Girshick, and J. Malik, “Contextual action recognition with r* cnn,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1080–1088.

[46] C. L. Zitnick and P. Dollár, “Edge boxes: Locating object proposals from edges,” in ECCV. Springer, 2014, pp. 391–405.

[47] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, “Selective search for object recognition,” International journal of computer vision, vol. 104, no. 2, pp. 154–171, 2013.

[48] R. Girshick, “Fast R-CNN,” in ICCV. IEEE, 2015, pp. 1440–1448.

[49] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 9, pp. 1904–1916, 2015.

[50] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in Advances in neural information processing systems, 2015, pp. 91–99.

[51] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017, pp. 2980–2988.

[52] J. C. Van Gemert, M. Jain, E. Gati, and C. G. Snoek, “Apt: Action localization proposals from dense trajectories.” in Proc. of BMVC, vol. 2, 2015, p. 4.

[53] D. Oneata, J. Revaud, J. Verbeek, and C. Schmid, “Spatio-temporal object detection proposals,” in European conference on computer vision. Springer, 2014, pp. 737–752.

[54] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, “Solving the multiple instance problem with axis-parallel rectangles,” Artificial intelligence, vol. 89, no. 1-2, pp. 31–71, 1997.

[55] X. Wang, M. Yang, S. Zhu, and Y. Lin, “Regionlets for generic object detection,” in Computer Vision (ICCV), 2013 IEEE International Conference on. IEEE, 2013, pp. 17–24.

[56] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.

[57] J. Hosang, R. Benenson, and B. Schiele, “How good are detection proposals, really?” in 25th British Machine Vision Conference. BMVA Press, 2014, pp. 1–12.

[58] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.

[59] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009, pp. 248–255.

[60] G. Richard, S. Sundaram, and S. Narayanan, “An overview on perceptually motivated audio indexing and classification,” Proceedings of the IEEE, vol. 101, no. 9, pp. 1939–1954, 2013.

[61] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” in Readings in speech recognition. Elsevier, 1990, pp. 65–74.

[62] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold et al., “CNN architectures for large-scale audio classification,” in ICASSP. IEEE, 2017, pp. 131–135.

[63] S. Abu-El-Haija, N. Kothari, J. Lee, A. P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8M: A large-scale video classification benchmark," in arXiv:1609.08675, 2016.

[64] C. Yu, K. S. Barsim, Q. Kong, and B. Yang, "Multi-level attention model for weakly supervised audio classification," arXiv preprint arXiv:1803.02353, 2018.

[65] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," Neural computation, vol. 21, no. 3, pp. 793–830, 2009.

[66] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[67] S. Parekh, A. Ozerov, S. Essid, N. Duong, P. Pérez, and G. Richard, "Identify, locate and separate: Audio-visual object extraction in large video collections using weak supervision," under review at WASPAA 2019, Draft available at https://arxiv.org/abs/1811.04000.

[68] D. Lee, S. Lee, Y. Han, and K. Lee, "Ensemble of convolutional neural networks for weakly-supervised sound event detection using multiple scale input," DCASE2017 Challenge, Tech. Rep., September 2017.

[69] C. Févotte, E. Vincent, and A. Ozerov, "Single-channel audio source separation with NMF: divergences, constraints and algorithms," in Audio Source Separation. Springer, 2018, pp. 1–24.

[70] M. Spiertz and V. Gnann, "Source-filter based clustering for monaural blind source separation," in in Proceedings of International Conference on Digital Audio Effects DAFx'09, 2009.

[71] NMF Mel Clustering Code, http://www.ient.rwth-aachen.de/cms/dafx09/.

[72] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," IEEE Transactions on Multimedia, vol. 17, no. 10, pp. 1733–1746, 2015.

[73] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," IEEE transactions on audio, speech, and language processing, vol. 14, no. 4, pp. 1462–1469, 2006.

[74] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in Advances In Neural Information Processing Systems, 2016, pp. 289–297.

[75] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in ECCV, September 2018.

Sanjeel Parekh received B. Tech (hons.) degree in electronics and communication engineering from LNM Institute of Information Technology, India in 2014 and M.S. in Sound and Music Computing from Universitat Pompeu Fabra (UPF), Spain in 2015. From 2016-19, he worked towards his Ph.D. thesis on audio-visual scene analysis at Technicolor R&D and Télécom Paris, France. His research focuses on developing and applying machine learning techniques to problems in audio and visual domains. Currently, he is with LTCI lab at Télécom Paris, France.

Slim Essid received his state engineering degree from the Éole Nationale d'Ingénieurs de Tunis, Tunisia, in 2001, his M.Sc. (D.E.A.) degree in digital communication systems from the École Nationale Supérieure des Télécommunications, Paris, France, in 2002, his Ph.D. degree from the Université Pierre et Marie Curie (UPMC), Paris, France, in 2005, and his Habilitation à Diriger des Recherches degree from UPMC in 2015. He is a professor in Télécom Paris's Department of Images, Data, and Signals and the head of the Audio Data Analysis and Signal Processing team. His research interests are machine learning for audio and multimodal data analysis. He has been involved in various collaborative French and European research projects, among them Quaero, Networks of Excellence FP6-Kspace, FP7-3DLife, FP7-REVERIE, and FP-7 LASIE. He has published over 100 peer-reviewed conference and journal papers, with more than 100 distinct co-authors. On a regular basis, he serves as a reviewer for various machine-learning, signal processing, audio, and multimedia conferences and journals, e.g., a number of IEEE transactions, and as an expert for research funding agencies.

Alexey Ozerov received the M.Sc. degree in mathematics from the Saint-Petersburg State University,Saint Petersburg, Russia, in 1999, the M.Sc. degree in applied mathematics from the University of Bordeaux 1, Bordeaux, France, in 2003, and the Ph.D. degree in signal processing from the University of Rennes 1, Rennes, France. He was working toward the Ph.D. degree from 2003 to 2006 with the labs of France Telecom R&D and in collaboration with the IRISA institute. From 1999 to 2002, he worked with Terayon Communicational Systems (USA) as an R&D software engineer, first in Saint-Petersburg, and then, in Prague, Czech Republic. He was with Sound and Image Processing Lab, KTH Royal Institute of Technology, Stockholm, Sweden, for one year (2007), with the Department of Images, Data, and Signals, Télécom Paris/LTCI, for one and half year (2008–2009), and with METISS team, IRISA/INRIA, Rennes, France, for two years (2009–2011). He is currently with InterDigital, France. His research interests include various aspects of audio and image/video analysis and processing. Since 2016, he has been a Distinguished Member of the Technicolor Fellowship Network and is currently a Member of the IEEE Signal Processing Society Audio and Acoustic Signal Processing Technical Committee. He is currently an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and received the IEEE Signal Processing Society Best Paper Award in 2014.

Ngoc Q.K. Duong received the B.S. degree from the Posts and Telecommunications Institute of Technology, Hanoi City, Vietnam, in 2004, the M.S. degree in electronic engineering from Paichai University, Daejeon, South Korea, in 2008, and the Ph.D. degree in computer science and control with the French National Institute for Research, Rennes, France, in 2011. From 2004 to 2006, he was with Visco JSC as a System Engineer. He was also a Research Engineer for the acoustic echo/noise cancellation system with Emersys Company, Korea, in 2008. He is currently a Senior Scientist with InterDigital, France. He is the co-author of more than 45 scientific papers and about 30 patent submissions. His research interests include signal processing and machine learning, applied to audio, image, and video. He was the recipient of the several research awards, including the IEEE Signal Processing Society Young Author Best Paper Award, in 2012 and the Bretagne Young Researcher Award, in 2015.

Patrick Pérez is the Scientific Director of Valeo.ai, Paris, France, a Valeo research lab on artificial intelligence for automotive applications. Before joining Valeo, he was a Distinguished Scientist with Technicolor (2009-2018), and a Researcher with Inria (1993-2000, 2004-2009) and with Microsoft Research Cambridge (2000-2004). His research interests include audio/video description, search, and analysis. He is currently on the Editorial Board for the International Journal of Computer Vision.

Gaël Richard received his State Engineering degree from Télécom Paris, France, in 1990, his Ph.D. degree from the University of Paris XI, France, in 1994 in speech synthesis, and his Habilitation à Diriger des Recherches degree from the University of Paris XI in 2001. After receiving his Ph.D. degree, he spent two years at Rutgers University, Piscataway, New Jersey, in the Speech Processing Group of Prof. J. Flanagan, where he explored innovative approaches for speech production. In 2001, he joined Télécom Paris, where he is now a professor in audio signal processing and the head of the Image, Data, and Signal Department. He is co-author of more than 200 papers, His research interests are mainly in the field of speech and audio signal processing and include topics such as signal representations and signal models, source separation, machine-learning methods for audio/music signals, music information retrieval, and multimodal audio processing. He is a Fellow of the IEEE.