URBAN SOUND & SIGHT: DATASET AND BENCHMARK FOR AUDIO-VISUAL URBAN SCENE UNDERSTANDING

Magdalena Fuentes[1*], *Bea Steers*[1*], *Pablo Zinemanas*[2], *Martín Rocamora*[3], *Luca Bondi*[4], *Julia Wilkins*[1], *Qianyi Shi*[1], *Yao Hou*[1], *Samarjit Das*[4], *Xavier Serra*[2], *Juan Pablo Bello*[1]

[1] New York University, New York, NY
[2] Universitat Pompeu Fabra, Barcelona, Spain
[3] Universidad de la República, Montevideo, Uruguay
[4] Bosch Research, Pittsburgh, PA, USA
[*]Equal contribution

ABSTRACT

Automatic audio-visual urban traffic understanding is a growing area of research with many potential applications of value to industry, academia, and the public sector. Yet, the lack of well-curated resources for training and evaluating models to research in this area hinders their development. To address this we present a curated audio-visual dataset, Urban Sound & Sight (Urbansas), developed for investigating the detection and localization of sounding vehicles in the wild. Urbansas consists of 12 hours of unlabeled data along with 3 hours of manually annotated data, including bounding boxes with classes and unique id of vehicles, and strong audio labels featuring vehicle types and indicating off-screen sounds. We discuss the challenges presented by the dataset and how to use its annotations for the localization of vehicles in the wild through audio models.

Index Terms— audio-visual, urban research, traffic, dataset.

1. INTRODUCTION

The automatic understanding of urban scenes is a growing area of research, with many potential applications of value to industry, academia, and the public sector. In particular, the automatic understanding of audio-visual urban traffic information is gaining increasing attention due to the availability of large quantities of online multimedia content, and has potential applications such as assistive devices for the hearing-impaired, the quantification of traffic for policy making, autonomous driving, among others. Audio-visual information is fundamental for the full understanding of real-world scenes, as visual and acoustic modals provide complementary information: images help identify sources and understand their motion, audio help understand the proximity of sources, the presence of relevant off-screen sounding objects, and help solve occlusions and improve estimations with poor lighting. Understanding an audio-visual urban scene includes estimating the class, spatial location, direction and speed of movement of beings and objects in real environments by the sounds they make and the way they look. Ideally, automatic solutions would be robust across a wide range of sound scenes and sensing conditions: noisy, sparse, with varying compositions of sources, with moving sources, with moving sensors.

While there is a large body of research in related computer vision (e.g. object detection and pedestrian counting [1, 2, 3]), and machine listening areas (e.g. urban sound event detection and classification, [4, 5]), there is little work on audio-visual classification and localization of sounding sources in realistic urban settings. Recently the machine listening community has turned its attention to localization, seeking to apply the same deep learning techniques that have proven successful in classification before [6, 7, 8, 9, 10], mostly using synthetic datasets.Research on the co-occurrence of audio and video has recently received increasing attention due to the development of self-supervised models that exploit audiovisual cues for their pretext-task [11, 12, 13, 14]. Most of this research is carried out using unlabeled videos from Youtube or Audioset [15], and models learn a representation of the data (either audio, visual or both) to later be applied to a downstream task [16]. Except for a few exceptions [17], these works have focused on audio-visual localisation mostly of sources such as musical instruments or in low-complexity settings, where the objects are relatively close to the camera and central to the scene.

One of the main challenges to audio-visual urban research is the lack of labeled data. While most of the existing resources involve either audio [18] or video [19] alone, the available audio-visual datasets of urban scenes have limited annotations, restricted to audio events only [20] or clip labels intended for scene classification [21]. Moreover, since manually annotating real-world data is very arduous and time consuming, the amount of labeled data tends to be small for machine learning standards. A way to alleviate the work of manual annotation is to create synthetic audio mixtures using isolated sound events [22] or synthetic visual scenes from video games [23], but they fail to capture the diversity and complexity of naturally occurring sound scenes. Another challenge is how to annotate moving sources in such complex settings: dealing with off-screen sounds, occlusions, or objects that can be seen but not heard.

The goal of this work is to take the first steps to address the challenges mentioned above. To that purpose, in Section 2, we introduce Urban Sound & Sight (Urbansas), a curated dataset of labeled and unlabeled audio-visual urban traffic data, with stereo audio, and annotations in audio and video. In Section 3 we discuss the challenges faced and the decisions made to annotate the data, and describe the annotations. In Section 4 we discuss ways of using the annotations of this dataset to approximate the position of vehicles in the wild using sound. Finally, in Section 5 we discuss the challenges of the data via baseline experiments for sound event detection and localization.

2. THE URBAN SOUND & SIGHT DATASET

We set four main goals for creating this dataset: 1) to compile a set of real-field audio-visual recordings; 2) the recordings should be stereo to allow exploring sound localization in the wild; 3) the compilation

should be varied in terms of scenes and recording conditions to be meaningful for training and evaluation of machine learning models; 4) the labeled collection should be accompanied by a bigger unlabeled collection with similar characteristics to allow exploring self-supervised learning in urban contexts in the future. In the following we explain how we have compiled Urbansas to fulfill these goals.

Data Sources. We have compiled and manually annotated Urbansas from two publicly available datasets, plus the addition of unreleased material. The public datasets are the TAU Urban Audio-Visual Scenes 2021 Development dataset [21] and the Montevideo Audio-Visual Dataset (MAVD) [20]. The TAU dataset consists of 10-second segments of audio and video from different scenes across European cities, traffic being one of the scenes. Only the subset of scenes labeled as traffic were included in Urbansas. MAVD is an audio-visual traffic dataset curated in different locations of Montevideo, Uruguay, with annotations of vehicles and vehicle components sounds (e.g. engine, brakes) for sound event detection. Besides the published datasets, we include a total of 9.5 hours of unpublished material recorded in Montevideo, with the same recording devices of MAVD but including new locations and scenes.

Data Capture. Recordings for TAU were acquired using a GoPro Hero 5 (30fps, 1280x720) and a Soundman OKM II Klassik/studio A3 electret binaural in-ear microphone with a Zoom F8 audio recorder (48kHz, 24 bits, stereo). Recordings for MAVD were collected using a GoPro Hero 3 (24fps, 1920x1080) and a SONY PCM-D50 recorder (48kHz, 24 bits, stereo).

Data Organization. In total, we gathered 15 hours of high quality material, which we organized as follows: 3 hours of data (1.5 hours TAU, 1.5 hours MAVD) manually annotated by our team both in audio and image, and 12 hours of unlabeled data (2.5 hours TAU, 9.5 hours of unpublished material). To reduce redundancy in the labeled data, we curated 1.5 hours from each dataset maximizing the variance of the data in terms of locations. An overview of Urbansas is presented in Table 1. Following the format of TAU, all audio and video are split into 10 second clips and are stored as separate MPEG4 and WAV files. Clips are uniformly formatted as 1280x720, 24fps video and 48kHz, 24 bit, stereo audio.We anonymized the data from MAVD and the newly curated data following [21].[1]

city	places	clips	mins	frames	labeled mins
Montevideo	8	4085	681	980400	92
Stockholm	3	91	15	21840	2
Barcelona	4	144	24	34560	24
Helsinki	4	144	24	34560	16
Lisbon	4	144	24	34560	19
Lyon	4	144	24	34560	6
Paris	4	144	24	34560	2
Prague	4	144	24	34560	2
Vienna	4	144	24	34560	6
London	5	144	24	34560	4
Milan	6	144	24	34560	6
Total	50	5472	912	1.3M	180

Table 1. Breakdown of Urbansas per city and location. Last column indicates the portion data in the labeled set.

[1]We used the anonymizer https://github.com/understand-ai/anonymizer.

3. ANNOTATING AUDIO-VISUAL URBAN SCENES

In order to understand an audio-visual urban scene, we want to estimate the class and location of each source as it moves over time. To that goal, we have annotated: 1) bounding boxes of objects with a class assignment and object id; 2) "strong" audio labels, with beginning and end timestamps and the correspondent class of the acoustic event; 3) relevant metadata about lighting and weather conditions (e.g. night vs. day). This dataset focuses on traffic since vehicles are a compelling case-study of sounding moving objects in urban settings. Consequently, our ontology focuses on the four most predominant vehicle types: *car*, *truck*, *bus*, and *motorbike*. In the following, we discuss the decisions we have made to annotate Urbansas.

Annotation Tools. We used CVAT [2] for the bounding box annotations, and VIA [24] for annotating the audio with the video as reference. For scenes where it was beneficial, we pre-computed bounding boxes using YOLO [2] and a customized location-based algorithm which used manual priors about the street orientation to track bounding boxes. Due to the large effort and time investment required for annotating, the video annotations were performed at 2fps to reduce redundant annotations, improve annotation quality, and allow for a larger volume of annotated clips (10 seconds at full fps is 240 frames to annotate which was found to be impractical).

Annotation types. There are two types of annotations in Urbansas: object annotations and scene annotations. Object annotations refer to vehicles in the scene that we are interested in and describe the class and position over time in both audio and video (e.g. bounding boxes and audio events). Scene annotations apply to entire clips at a scene and are informative of the context of those vehicles (e.g. takes place at night, too many vehicles to hear distinctly, etc.).

Notation. For a specific file in the dataset, let us define an audio annotation as a tuple $(t_{s,i}, t_{e,i}, l_i),\ i \in [1, N_A]$, where $t_{s,i}$ and $t_{e,i}$ are the start and end time of an audio event with label l_i, and N_A is the total number of audio annotations for the file. We also define a video annotation as a tuple $(t_j, l_j, tr_j, v_j, x_j, w_j, y_j, h_j),\ i \in [1, N_V]$, where t_j is the timestamp of an object with label l_j and visibility flag v_j; track id tr_j is used to identify a single object across frames in a file; the bounding box for the object is defined in terms of horizontal (x_j) and vertical (y_j) shift between the top-left corner of the frame and bounding box, with corresponding height (h_j) and width (w_j); N_V is the total number of video annotations for the file.

Video annotations of sounding vehicles. Vehicles in the video are annotated if they are believed to contribute to the acoustic scene. Primarily, this includes vehicles that either drive past or idle near the observer, while excluding vehicles with their engines off (i.e. parked). In complex scenes, there are often multiple roads at different distances. In these scenarios, acoustic masking is taken into account - e.g., if vehicles from closer road mask sounds from the further road, then only the closer vehicles are annotated. If the closer road is less busy, then the further road may be annotated as well. If a vehicle is temporarily occluded (hidden behind something, partially or fully) it is still annotated with an estimate of its true location, with an additional flag ($v_j = 0$) identifying it as occluded.

Integrating audio and video annotations. The audio annotations can be used in combination with the video annotations to identify vehicles that are both audible and visible. In some cases, an object could have no audio events (and vice versa) if the sound occurs before or after the vehicle enters/leaves the scene (this can happen for certain camera angles). In other cases, an audio event may have no corresponding object in the video, which may happen when a

[2]https://github.com/openvinotoolkit/cvat

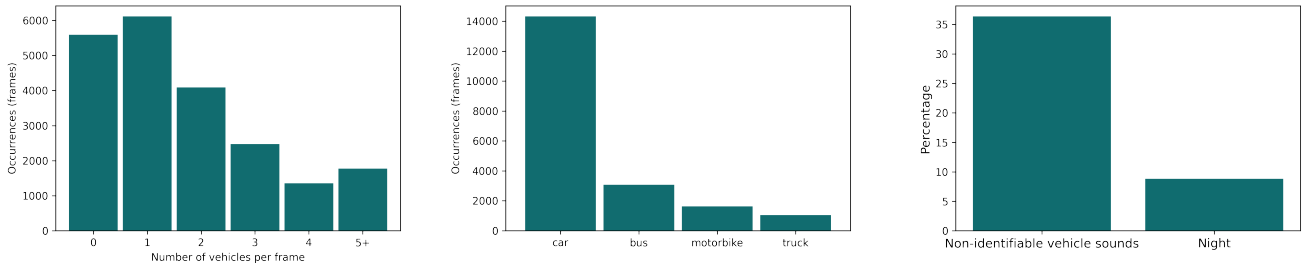

Fig. 1. Breakdown of events in Urbansas labeled set (at 2fps). *Left:* concurrency of vehicles base on image annotations; *Center:* Number of frames that each vehicle type appears in the scene (considering unique vehicles per scene); *Right:* Percentage of clips with clip-level annotations *non_identifiable_vehicle_sound* and *night*.

vehicle passes outside of the camera's view; these are labeled as off-screen sounds. Since we have the audio annotations to disambiguate when the object is both present in the image and producing sound at the same time, we annotate vehicles when they are "close enough" so is informative of error types in visual-only or audio-visual models.

Scene annotations. Whenever possible, we annotate beginning and end of vehicle events in the audio using the video to determine the vehicle class. However, some scenes have many vehicles passing at the same time and it is perceptually very hard to attribute sounds to a particular vehicle, they rather produce a "constant background sound" altogether. To address this, we include a binary flag at the clip level indicating the presence of *non_identifiable_vehicle_sound*. In cases where particular vehicles are identifiable on top of this constant sound, we annotate them with strong labels as well as indicate the presence of non identifiable vehicle sounds. Usually these scenes will present many bounding boxes at the same time. Additionally, we include flags indicative of the lighting: night vs. day.

4. LOCALIZING SOURCES IN THE WILD

Since Urbansas consists of data captured in real world conditions, it differs significantly from the synthetic datasets normally used in SELD tasks (e.g. in DCASE challenges). Besides the contrast in the realism of the data, the main two differences are: 1) the exact position of the sounding sources is not known in the ground truth, but instead we have an approximated idea of where the sources are and how are they moving given the video; and 2) there are overlapping sounds of the same source (e.g. multiple audible cars in the same scene at different positions). For this reason, the methods and metrics typically used for SELD would not be suitable to work with this data. Instead, we make approximations to work with partial information in the wild, as indicated in the following.

Indexing of video annotations for audio localization. We approximate the vehicles position using linearly spaced regions corresponding to the angles within the camera's field of view (FoV). For each video annotation, we approximate the position (θ_j) of the object based on the coordinates of the bounding box, and then we quantize θ_j to the closest region. We explore two ways of computing θ_j: 1) We consider the vehicles as point sources. For this we used the center point of the bounding box as the position indicator. Formally:

$$\theta_j(x_j) = \left(\frac{x_j + \frac{w_j}{2}}{W} - \frac{1}{2} \right) f_{ov}, \tag{1}$$

where W is the width of the frame and f_{ov} is the FoV of the camera. Working with θ_j approximated this way allow us to combine data with different FoVs and resolutions (W) in the future. 2) We relax the point-wise estimation and instead approximate the vehicle location to be the entire bounding box. For this, we use two positional values $\theta_{j,L}$ and $\theta_{j,R}$ which define the left and right limits of the source location respectively:

$$\theta_{j,L} = \theta_j \left(x_j - \frac{w_j}{2} \right), \text{ and } \theta_{j,R} = \theta_j \left(x_j + \frac{w_j}{2} \right). \tag{2}$$

Finally, we map θ_j to a specific region r_j as $r_j = \text{argmin}_i |\theta_j - r_i|$ for $i \in \{1, .., R_N\}$, where r_i denotes the region i and R_N is the total number of regions the FoV is divided, which we set to 5. For simplicity, we assume all videos are recorded with a pin-hole camera with a FoV of 120°, estimated from the cameras specification sheet.

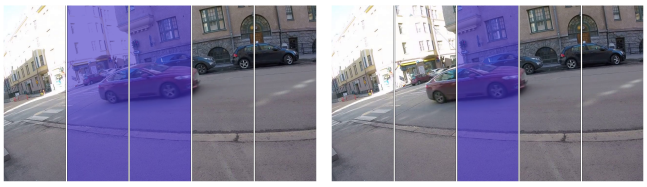

Fig. 2. *Left:* Regions activated using the full bounding box, *Right:* regions activated using the bounding box center. Each region conveys 24°. *Note: black vehicle is parked so it is not annotated.*

Audio annotations as filter to video objects. At training time, we only consider video events that are confirmed by audio annotations. This condition is met if the timestamps for a given video object overlap with the start and end time of an audio annotation with the same label. Note that this is not always the case since there are scenes where vehicles are visible but not audible and the other way around. Formally, given a video object $\mathcal{V}^k$ characterized by label l^k and a set of timestamps $\{t_p^k \; p \in [1, N_{\mathcal{V}_k}]\}$, we consider the video object as valid for training if it exists at least one audio annotation $(t_{s,i}, t_{e,i}, l_i)$, $i \in [1, N_A]$ such that $l_i = l^k$ and $\sum_{p=1}^{N_{\mathcal{V}^k}} \left[(t_p^k \geq t_{s,i}) \wedge (t_p^k \leq t_{e,i}) \right] >= 0$, i.e. audio and video overlap and their labels coincide.

Metrics. Usual SELD metrics [25, 26] are designed to work with a single source per class, and they compute either the angular or euclidean distance of the reference and estimated positions. This hypothesis does not apply in Urbansas, since multiple vehicles of the same type are often encountered in the scene. Instead, we propose to tackle the problem considering the overlapping regions between

the estimation and the ground truth, which would generalize for "angular" regions or image regions and would allow the comparison of audio-visual models and audio-only models in the future. In the computer vision community, this is evaluated using a metric called intersection over union (IoU), which intuitively computes the intersection between the ground truth bounding boxes and the estimated ones [14], a technique that can also be extended to 1-dimensional regions. In our case, the regions correspond to the horizontal location of the vehicle in the image, and its quantized θ_j derived from it. Following the ideas in [14] we propose to compute the IoU as:

$$IoU(\tau, c) = \frac{\sum_{i \in A_c(\tau)} g_{i,c}}{\sum_i g_{i,c} + \sum_{i \in A_c(\tau) - G_c} 1} \quad (3)$$

where i indicates the region in the image, c is the class index, τ is the threshold to determine if a prediction is positive or not so $A_c(\tau) = \{i | p_i > \tau\}_c$ and $G_c = \{i | g_i > 0\}_c$. IoU scores range in [0,1]. Because of the multi-label nature of the data we made the IoU score dependent on the class, so each class score is computed independently. The IoU is a promising metric for dealing with multi-source multi-direction scenes, and to bridge audio-only and audio-visual models making their estimations comparable, but it has no information about distance, and thus penalizes the models the same if the error is small or large. Extending the metric to deal with distance is out of the scope of this paper and will be studied in the future.

5. DATA CHALLENGES AND BASELINE

To learn about the challenges of the data and usefulness of the metric, we ran a set of experiments with simple baselines. We are not searching for an optimal model that maximizes accuracy but rather we are interested in understanding the characteristics of the dataset and metric themselves, and identifying venues for future research.

Baselines. Our baselines are based on the convolutional-recurrent network in [7]. This method predicts the probability of a class being present, and the horizontal and vertical direction for each class using multi-channel audio. It makes the assumption that there is at most one instance of each class in the scene at a time. As shown in Figure 1, a large portion of the frames contain multiple sources, and cars make up a significant majority of the observed vehicles, meaning that limiting the scenes to only those that contain a single instance of the class would significantly reduce the size of the applicable data, and would limit the utility of the model predictions on common, real-world scenarios involving multiple sources. Instead, we adapt the architecture of [7] to use stereo audio, and to be multi-class and multi-direction model, i.e. to predict overlapping sources of the same class and with different positions. To do so, our model predicts a tensor $T(i, c, j) = (t_i, c, r_k)$ for each time t_i, $i \in [1, N_f]$ with N_f the number of frames, vehicle class $c \in \{C_1, .., C_4\}$, and region $r_j \in [R_1, R_5]$. We use a sigmoid layer to allow for multiple activations at once. We train and evaluate the box-wise model using the regions covering the entire bounding box, and the point-wise using the regions activated at the center of the bounding box (see Figure 2). We also include two random baselines: a point-wise baseline that can predict up to two active regions at a time, and a box-wise baseline that estimates up to five regions. Each one is compared to the matching ground-truth (point- and box-wise).

Training and evaluation protocol. We split the labeled set into 5 folds stratified by location and we perform cross-fold (4-1) training and validation. We train using 4 second chunks as in [7]. We train two models following the ideas in Section 4: a *point-wise* model that predicts the position of vehicles as the region corresponding to the center of the bounding boxes, and a *box-wise* model that predicts the regions where the bounding box is present. We used a weighted binary cross-entropy loss for training (see implementation).[3]

Results. Results are depicted in Table 2. We compute the IoU score for non-empty frames (i.e. frames containing at least one bounding box that overlaps with the audio). The first observation is that both models perform better than random, the box-wise model being the best. This is to expect since the bounding box conveys more regions than the point-wise case and thus is an easier problem. We see a considerable drop in performance for the least frequent class (truck) whose sound resembles to cars and buses, unlike motorbikes.

model	*IoU* ($\tau = 0.05$)				
	bus	*car*	*motorbike*	*truck*	*all*
point-wise (pw)	0.332	0.344	0.231	0.143	0.260
box-wise (bw)	0.473	0.468	0.285	0.180	0.351
pw-random	0.045	0.045	0.048	0.037	0.044
bw-random	0.102	0.100	0.089	0.115	0.102

Table 2. IoU per-class of baseline models on non-empty frames.

We also compute the IoU for all frames, including inactive frames, to assess whether the baseline can determine the presence (and absence) of vehicles in a clip. For those empty frames, we compare the prediction mask of the model with an empty ground truth, obtaining a score of 1 if the model did not predict the class at any direction. We obtained better scores overall in this setting: cars ($IoU = 0.361$), buses ($IoU = 0.779$), motorbikes ($IoU = 0.808$) and trucks ($IoU = 0.917$) for the box-wise model. A counter-intuitive result is that the highest scores correspond to the least represented classes in the dataset. Taking a closer look at the results, we believe that this is due to the low frequency of such vehicles in the scenes and the fact that the baseline models have low confidence values in general, favoring empty predictions and scoring high in empty frames. This indicates that the joint detection and localization of vehicles is a highly imbalanced and hard learning problem. Regarding the usefulness of the IoU metric for localization of sources in the wild, we believe that the formulation of the problem as detection and localization makes it hard to judge with this metric how good the models are at localizing and detecting respectively, and we plan to explore them separately in the future.

6. CONCLUSIONS AND FUTURE WORK

We present Urbansas, an audio-visual dataset of traffic scenes, containing 12 hours of unlabeled data, suitable for unsupervised and self-supervised research in visual sound source detection and localization, and 3 hours of human-annotated data, containing bounding boxes, classes, and tracking information to be used for supervised research and validation of self-supervised models as a downstream task. To the best of our knowledge, Urbansas is the first audio-visual urban traffic dataset with human-annotated labels both in audio and video. We believe the dataset will open the path to new research on audio and audio-visual sound source localization, vehicle tracking, self-supervised audio-visual representation for real world applications, among others. We present first experiments on vehicle localization and detection, including a baseline and evaluation metric for the task. The data and code are open to the research community.

[3] https://github.com/magdalenafuentes/urbansas.

Work partially supported by NSF award 1955357 and Bosch RTC.

7. REFERENCES

[1] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, "SSD: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[2] Joseph Redmon and Ali Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[3] Vishwanath A Sindagi and Vishal M Patel, "A survey of recent advances in CNN-based single image crowd counting and density estimation," *Pattern Recognition Letters*, vol. 107, pp. 3–16, 2018.

[4] Justin Salamon and Juan Pablo Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal processing letters*, vol. 24, no. 3, pp. 279–283, 2017.

[5] Annamaria Mesaros, Toni Heittola, Tuomas Virtanen, and Mark D Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 2021.

[6] Pasi Pertilä, Emre Cakir, Aapo Hakala, Eemi Fagerlund, Tuomas Virtanen, Archontis Politis, and Antti Eronen, "Mobile microphone array speech detection and localization in diverse everyday environments," *arXiv preprint arXiv:2106.14787*, 2021.

[7] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2018.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[9] Carl Doersch, Abhinav Gupta, and Alexei A. Efros, "Unsupervised visual representation learning by context prediction," 2016.

[10] Israel D. Gebru, Silèye Ba, Georgios Evangelidis, and Radu Horaud, "Tracking the active speaker based on a joint audio-visual observation model," in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2015, pp. 702–708.

[11] Relja Arandjelovic and Andrew Zisserman, "Objects that sound," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[12] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba, "The sound of pixels," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Sept. 2018, pp. 570–586.

[13] Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel PW Ellis, and John R Hershey, "Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds," *arXiv preprint arXiv:2011.01143*, 2020.

[14] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman, "Localizing visual sounds the hard way," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16867–16876.

[15] Jort F Gemmeke, Daniel PW 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[16] Jason Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo 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.

[17] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba, "Self-supervised moving vehicle tracking with stereo sound," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7053–7062.

[18] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "TUT database for acoustic scene classification and sound event detection," in *In 24rd European Signal Processing Conference 2016 (EUSIPCO 2016), Budapest, Hungary*, 2016.

[19] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The Cityscapes Dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[20] Pablo Zinemanas, Pablo Cancela, and Martín Rocamora, "MAVD: A dataset for sound event detection in urban environments.," *Detection and Classification of Acoustic Scenes and Events, DCASE 2019, New York, NY, USA, 25–26 oct, page 263–267*, 2019.

[21] Shanshan Wang, Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "A curated dataset of urban scenes for audio-visual scene analysis," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 626–630.

[22] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello., "Scaper: A library for soundscape synthesis and augmentation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, october 2017.

[23] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun, "Playing for data: Ground truth from computer games," in *European Conference on Computer Vision (ECCV)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, Eds. 2016, vol. 9906 of *LNCS*, pp. 102–118, Springer International Publishing.

[24] Abhishek Dutta and Andrew Zisserman, "The VIA annotation software for images, audio and video," in *Proceedings of the 27th ACM International Conference on Multimedia*, New York, NY, USA, 2019, MM '19, ACM.

[25] Annamaria Mesaros, Sharath Adavanne, Archontis Politis, Toni Heittola, and Tuomas Virtanen, "Joint measurement of localization and detection of sound events," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 333–337.

[26] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, March 2018.