Polyphonic training set synthesis improves self-supervised urban sound classification

Félix Gontier, Vincent Lostanlen, Nicolas Fortin, Mathieu Lagrange, Catherine Lavandier, Jean-François Petiot

▶ **To cite this version:**

Félix Gontier, Vincent Lostanlen, Nicolas Fortin, Mathieu Lagrange, Catherine Lavandier, et al.. Polyphonic training set synthesis improves self-supervised urban sound classification. Journal of the Acoustical Society of America, 2021. hal-03262863

HAL Id: hal-03262863
https://hal.science/hal-03262863

Submitted on 23 Jul 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

JASA ARTICLE

Polyphonic training set synthesis improves self-supervised urban sound classification[a)]

Félix Gontier,[1] Vincent Lostanlen,[1] Mathieu Lagrange,[1,b)] Nicolas Fortin,[2] Catherine Lavandier,[3] and Jean-François Petiot[4]

[1]*CNRS, LS2N, F-44322 Nantes, France*

[2]*Unité Mixte de Recherche en Acoustique Environnementale, Université Gustave Eiffel, Centre d'Etudes et d'Expertise sur les Risques, l'Environnement, la Mobilité et l'Aménagement, F-44344 Bouguenais, France*

[3]*CY Cergy Paris Université École Nationale Supérieure de l'électronique et de ses Applications (ENSEA), CNRS, ETIS, F-95000 Cergy, France*

[4]*École Centrale de Nantes, LS2N, F-44322 Nantes, France*

ABSTRACT:

Machine listening systems for environmental acoustic monitoring face a shortage of expert annotations to be used as training data. To circumvent this issue, the emerging paradigm of self-supervised learning proposes to pre-train audio classifiers on a task whose ground truth is trivially available. Alternatively, training set synthesis consists in annotating a small corpus of acoustic events of interest, which are then automatically mixed at random to form a larger corpus of polyphonic scenes. Prior studies have considered these two paradigms in isolation but rarely ever in conjunction. Furthermore, the impact of data curation in training set synthesis remains unclear. To fill this gap in research, this article proposes a two-stage approach. In the self-supervised stage, we formulate a pretext task (Audio2Vec skip-gram inpainting) on unlabeled spectrograms from an acoustic sensor network. Then, in the supervised stage, we formulate a downstream task of multilabel urban sound classification on synthetic scenes. We find that training set synthesis benefits overall performance more than self-supervised learning. Interestingly, the geographical origin of the acoustic events in training set synthesis appears to have a decisive impact.

© *2021 Acoustical Society of America*. https://doi.org/10.1121/10.0005277

(Received 2 February 2021; revised 20 May 2021; accepted 25 May 2021; published online 16 June 2021)

[Editor: Bozena Kostek] Pages: 4309–4326

I. INTRODUCTION

A. Monitoring sound quality with urban acoustic sensors

The urban population of the world is growing rapidly: from 751×10^6 in 1950, it reached 4.2×10^9 in 2018, and it is expected to increase further over the next decades (United Nations, 2018). In this context, noise pollution jeopardizes the well-being of many residents in dense industrialized areas. For example, in the United States, 70×10^6 people suffer from harmful levels of noise (Hammer *et al.*, 2014), with effects including sleep disruption, cardiovascular disease, and hearing loss (Basner *et al.*, 2014). Moreover, repeated exposure to noise has been proved to reduce the learning abilities of children during class (Bronzaft, 2002). Beyond the scope of public health, the issue of urban noise pollution has many other sociopolitical implications, such as effects on tourism and real estate values (Bristow and Thanos, 2015).

Modeling the appraisal of citizens of their sound environment is a difficult task. Different studies have proposed a taxonomy of urban sounds (Berglund and Nilsson, 2006; Brown *et al.*, 2011; Salamon *et al.*, 2014) and have concluded that the perception of the acoustic environment of urban areas is primarily influenced by three sources (Ricciardi *et al.*, 2015): *traffic* (presence of vehicles), *voices* (presence of humans), and *birds* (presence of nature). Furthermore, these studies show that it is possible to predict the perceived pleasantness of outside sound environment from the point of view of pedestrians nearby. To this end, it is necessary to take into account the overall loudness of this sound environment but also the presence of traffic, voices, and birds. In this context, the motivation of our paper is to detect and classify the presence of these sources on a frame-level basis, with a frame duration that is typically equal to 1 s. We will leave as future work the integration of urban sound classification into predictive perceptual models of the urban auditory environment.

The characteristics of sound environments may vary at small spatiotemporal scales, i.e., typically $1000\,\mathrm{m}^2 \times 10\,\mathrm{min}$ (Brocolini *et al.*, 2013). Therefore, a map of all major noise sources in a given city cannot be achieved by human inspection alone, be it expert or crowdsourced (New York City Department of Health and Mental Hygiene, 2014). Rather, the prospect of monitoring urban noise in real time calls for the deployment of an acoustic sensor network (Vidaña-Vila *et al.*, 2020). Indeed, prior research has demonstrated the potential of acoustic sensors for noise

[a)]This paper is part of a special issue on Machine Learning in Acoustics.

[b)]Electronic mail: mathieu.lagrange@ls2n.fr, ORCID: 0000-0002-1253-4427.

mitigation, particularly via cartography (Park *et al.*, 2014) and partnership with city agencies (Mydlarz *et al.*, 2017).

Meanwhile, the past decade has witnessed a surge of deep learning models for the computational analysis of sound scenes and events, with remarkable results in the automatic classification of environmental sounds (Virtanen *et al.*, 2018). These statistical models may detect the presence of sound sources near each sensor at any time of the day or night (Cartwright *et al.*, 2019c). In addition, urban sound classifiers may serve as pre-processing systems for soundscape ecology (Pijanowski *et al.*, 2011).

A recent publication has shown that estimating the presence of three coarse categories (traffic, voice, and birds) by means of a deep convolutional network suffices to predict the pleasantness of polyphonic urban soundscapes, as judged by human listeners (Gontier *et al.*, 2019). Henceforth, the effective monitoring of audible patterns in a city requires high-throughput analysis software paired with low-cost sensing hardware (Picaut *et al.*, 2020).

B. Limitations of supervised machine listening models

The ultimate goal of urban sound classification is to reach a high degree of agreement between the outputs of the machine and the judgments of one or multiple human listeners, over a diverse range of recording conditions. The most straightforward way to achieve this goal is to train the machine listening model in a supervised way, i.e., to replicate the classification of humans over a collection of annotated samples (Andén *et al.*, 2019; Lagrange *et al.*, 2015; Piczak, 2015). Despite its simplicity, supervised learning in acoustic sensor networks suffers from a number of practical drawbacks.

First, annotating sounds is a tedious and time-consuming task (Cartwright *et al.*, 2019b). Auditory perception is inherently tied to the constraints of real time: the physical duration of an annotated audio collection scales in proportion with the number of human-hours spent annotating. On the one hand, the "weak labeling" task, where listeners only annotate acoustic sources in terms of presence or absence, can be accomplished almost on par with real time. On the other hand, the "strong labeling" task also involves the precise onset and offset of each acoustic event in the recording: as such, it often requires multiple playbacks and is thus much slower than real time (Cartwright *et al.*, 2017). While recent advances in multiple-instance learning have proposed "strong" machine listening systems that are trained on "weak" labels only (McFee *et al.*, 2018), the topic of efficient audio annotation remains central in urban sound classification (Méndez Méndez *et al.*, 2019).

Second, supervised machine listening is exposed to sampling bias as well as label noise (Fonseca *et al.*, 2019). In the age of user-generated content, media sharing platforms harvest massive amounts of crowdsourced audio data in general and urban sounds in particular (Font *et al.*, 2013). Specifically, scraping YouTube (for-profit) and Freesound (nonprofit) has led to the curation of the AudioSet (Gemmeke *et al.*, 2017) and FSD50k (Fonseca *et al.*, 2021) datasets, respectively. That being said, the probability distribution of acoustic events in crowdsourced data does not reflect the real-world use case of sound quality monitoring. This is a form of convenience sampling, which incurs sampling bias in the machine learning process. Besides, delegating the annotation process to crowdsourcers tends to lower inter-rater agreement, a phenomenon known as label noise (Zhu *et al.*, 2020).

Third, and perhaps most fundamentally, supervised models for urban sound classification cannot be re-used verbatim from one city to another. This is because different cities will typically enforce different "noise codes," depending on geographical and cultural factors (Brown *et al.*, 2011). For example, the Sounds of New York City (SONYC) project has deployed a network of 50 acoustic sensors to monitor noise pollution in New York City (Bello *et al.*, 2019). Meanwhile, the CENSE project operates a network of 100 sensors to monitor sound quality, which are attached onto street lights at 3 m high, in the city of Lorient, France (Ardouin *et al.*, 2018). According to the New York City Department of Environmental Protection, the list of most frequent causes of noise complaints in New York includes *pile driver* and *large rotating saw* (Cartwright *et al.*, 2019c). These construction tools are virtually unheard of in the center of a small city such as Lorient (population 57 149). Therefore, although SONYC and CENSE have similar aims, mutualizing their machine listening systems would be far from trivial. More generally, the robust deployment of machine listening systems across mismatched acoustic environments remains a challenging task (Lostanlen *et al.*, 2019).

For the reasons mentioned above, the paradigm of supervised learning does not suffice to train machine listening systems for acoustic sensor networks. Sections I C and I D review two alternatives to this paradigm: training set synthesis and self-supervised learning.

C. Training set synthesis: Real-world tasks on fake data

Training set synthesis consists in synthesizing a training set for the task of multilabel classification (Lafay *et al.*, 2016). The key idea is to curate a relatively small collection of acoustic events of interest, which are then combined iteratively to form complex polyphonic scenes. Because the onset and offset times of all events are under control, there is no need for manual annotation: instead, a "virtual annotator" assigns strong labels to the synthetic scene at hand based on its monophonic constituents.

The most widely used software libraries for training set synthesis in machine listening are simScene in MATLAB (Lafay *et al.*, 2016) and Scaper in Python (Salamon *et al.*, 2017a). These libraries have led to the public release of synthetic audio datasets, such as DCASE OS (Mesaros *et al.*, 2018; Stowell *et al.*, 2015) for office sounds, URBAN-SED (Salamon *et al.*, 2017b) for urban sounds, BirdVox-scaper-10k (Mendoza *et al.*, 2019) for avian flight calls, and

DESED (Turpault and Serizel, 2020) for domestic sounds. Furthermore, state-of-the-art systems in singer identification (Lee and Nam, 2019) and audio segmentation in radio broadcasts currently rely on fully synthetic training sets. Beyond the scope of machine listening, the topic of learning in simulated environments has a long history in computer vision as well as robotics (Gaidon *et al.*, 2018). Therefore, training set synthesis is a promising, yet largely unexplored, technique for urban sound classification with limited labeled data.

D. Self-supervised learning: Fake tasks on real-world data

Self-supervised learning is a more recent approach than training set synthesis. It proposes to circumvent the cost of human annotation by formulating a machine learning task whose ground truth is trivially available to the machine itself. This paradigm, known as self-supervised learning, optimizes the trainable parameters of the machine listening system according to computer-generated labels rather than crowdsourced ones. In the context of urban acoustic sensor networks, one example of a self-supervised task consists in predicting the time of day that is associated with a given audio recording (Cartwright *et al.*, 2019a). Here, the key idea is that anthropogenic noise follows a certain circadian pattern, thus making local time approximately identifiable from audio measurements.

Of course, this "pretext task" has no practical interest *per se*: clocks already provide the local time with sufficient precision. However, the postulate of self-supervised learning is that solving the pretext task will indirectly train the machine to extract informative features for general-purpose machine listening. These features operate at an intermediate level of abstraction, between the raw data and the associated class labels (Kolesnikov *et al.*, 2019). Once the self-supervised learning stage on the pretext task has converged, the mid-level features serve as an input representation to a "downstream task," in our case, urban sound classification. Unlike the pretext task, the downstream task does require manual labeling and is optimized via conventional supervised learning.

Interestingly, training set synthesis and self-supervised learning play dual roles: while the former simulates fake data to accomplish real-world classification tasks, the latter formulates fake tasks as a pretext for analyzing real-world unlabeled data. We refer to Pascual *et al.* (2019) for a recent review of the state of the art in self-supervised machine listening.

E. Contributions of the present paper

Our paper aims at alleviating the cost of human labor in the process of training a machine listening system. The motivation for this study resides in the deployment of an acoustic sensor network for monitoring. Taking the CENSE project as an example use case, we present a new approach for training a multilabel audio classifier. Specifically, we combine two emerging techniques: self-supervised learning, which requires large-scale data acquisition but no annotation, and training set synthesis, which requires small-scale data acquisition and small-scale annotation.

Prior studies have considered training set synthesis and self-supervised learning separately but rarely ever in conjunction. To the best of our knowledge, our publication is the first to combine both these techniques in the context of machine listening. We note that a recent publication relies on audio-visual correspondence to separate sounds and synthesize isolated sources from an audio mixture (Zhao *et al.*, 2018). That publication regards audio synthesis as its end goal, not as a technique for generating computer-annotated data. Outside of the field of machine listening, Tung *et al.* (2017) have trained a deep neural network to perform motion capture (mocap) from a single-camera video input via a combination of strong supervision on synthetic video data and self-supervised rendering of three-dimensional (3D) keypoints.

Figure 1 illustrates our approach. The key idea is to decompose the machine learning process into two stages. In the first stage, we formulate a "pretext task" to analyze unlabeled spectrogram data from the sensor network: our paper relies on Audio2Vec skip-gram (SG) inpainting (Tagliasacchi *et al.*, 2020) for that purpose. In the second stage, we curate and annotate a limited amount of audio recordings for sources of interest (40 min in our case) and fine-tune the self-supervised model to accomplish the supervised task of multilabel urban sound classification.

After having trained a convolutional recurrent neural network (CRNN) with the two-stage approach described above, we evaluate it on a hold-out dataset of audio recordings from the city of Lorient. We name this dataset Lorient-1k, since it has an approximate duration of 1k seconds (see Appendix C). On this dataset, our system reaches an average accuracy of 73.6% across the three classes of interest: *traffic*, *voice*, and *birds*. Removing self-supervised learning from our pipeline reduces accuracy slightly (72.9%), whereas reducing polyphonic training set synthesis reduces it dramatically (49.6%). In comparison, the official pretrained classifier of TensorFlow Hub (YAMNet) achieves an average accuracy of 71.4%.

Beyond the raw performance comparison, we note that YAMNet was trained on a large-scale corpus (AudioSet) in a supervised way, by relying on extensive crowdsourced annotation (Gemmeke *et al.*, 2017). Meanwhile, our approach accomplishes the task of urban sound classification despite having been trained only on a pretext task (Audio2Vec) and fine-tuned on synthetic data. Crucially, we acquired the training datasets (CENSEgram-5M and simCENSE-18k) and evaluation dataset (Lorient-1k) in the same city but with different hardware: respectively, stationary sensors [Micro-Electro-Mechanical System (MEMS)-based] and Zoom (Hauppauge, NY) H4n handheld devices. Therefore, our approach generalizes to sensor technologies that are unseen in the training set.

Our most surprising observation is that, in the preparation of training set synthesis, curating audio samples in the

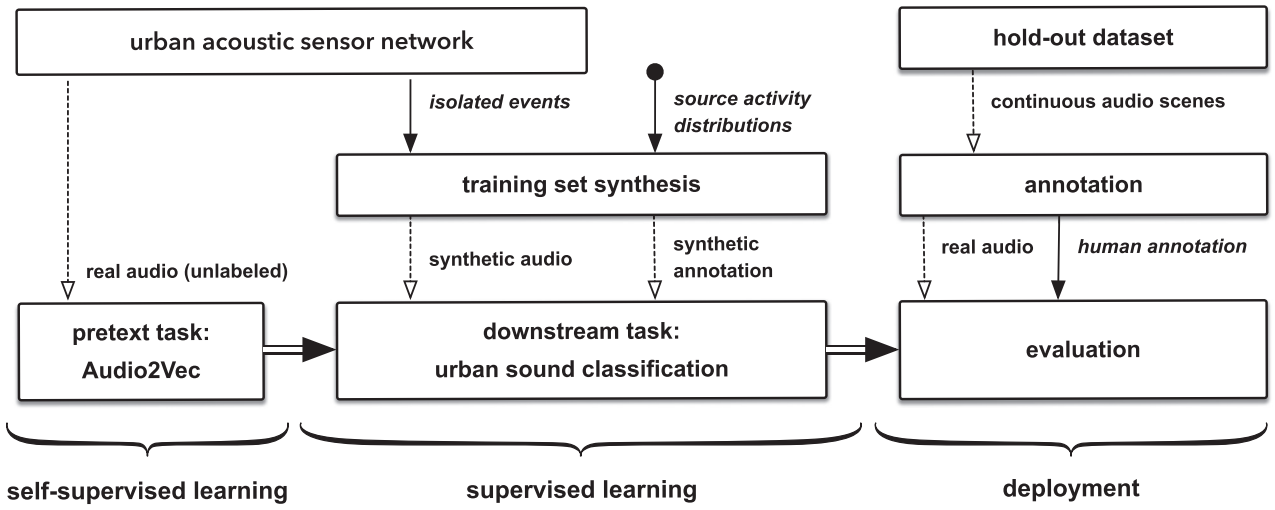

FIG. 1. Overview of the proposed system. Solid and dashed arrows, respectively, denote data collection procedures by humans and machines. Large double arrows denote transfers of model weights. See Sec. I E for details.

same city as the evaluation dataset appears to be necessary. On the one hand, training on synthetic scenes from Lorient samples (simCENSE-18k) yields state-of-the-art performance; on the other hand, replacing these local samples by external ones (from FreeSound and Librispeech in our case) brings the frame-wise classification accuracy down to 59.6% on average. This finding suggests that small-scale annotation in the acoustic environment of interest, combined with training set synthesis, has the potential to outperform general-purpose audio classifiers trained on large-scale corpora.

Section II describes the self-supervised learning stage, in particular, the Audio2Vec pretext task. Section III describes the supervised learning stage, which is based on training set synthesis with simScene. Section IV presents our evaluation dataset, metrics, and results. Last, Sec. V presents a detailed benchmark of our approach, during which we evaluate the effect of three methodological components: self-supervised learning, training set synthesis, and local curation of training data.

II. SELF-SUPERVISED LEARNING WITH AUDIO2VEC

This section applies the self-supervised learning paradigm to a network of urban acoustic sensors, named CENSE. Specifically, we train a deep convolutional network to solve a "pretext task" named Audio2Vec. This pretext task was initially proposed by Chung and Glass (2017) for similarity retrieval between spoken words and extended by Tagliasacchi *et al.* (2020) to general-purpose acoustic similarity retrieval. The originality of our protocol is that it operates on pre-computed spectrograms, without requiring persistent access to waveform audio.

A. The CENSE acoustic sensor network

The central goal of the CENSE project is to monitor urban acoustic environments. As a case study, CENSE currently operates a network of over 100 acoustic sensors in the French city of Lorient. This network has a greater density of sensors per unit area than comparable projects, such as SONYC in New York City (Mydlarz *et al.*, 2019), DYNAMAP in Rome (Bellucci *et al.*, 2017), SONORUS in Antwerp (Botteldooren *et al.*, 2018), or StadtLärm in Jena (Abeßer *et al.*, 2018). For more general information on the CENSE project, see CENSE (2019).

Another originality of the CENSE network lies in the extraction of spectrograms on the sensing device itself. This design choice is a form of *edge computing*, a paradigm in which acoustic sensors perform audio feature extraction or content analysis before transmitting data (Sheng *et al.*, 2019). Edge computing is opposed to cloud computing, in which sensors transmit audio data verbatim to a central server.

B. On-device extraction of third-octave spectrograms

Each sensor in the CENSE network extracts spectrograms via an STM32L4 microcontroller or a Raspberry Pi single-board computer, depending on node connectivity. Following the efficient algorithm of Antoni (2010), we decompose the audio input over 29 third-octave bands whose center frequencies range from 20 Hz to 12.5 kHz. Then we apply the pointwise squared modulus and integrate each band over non-overlapping windows of duration equal to 125 ms; i.e., 8 frames per second. Last, we apply pointwise logarithmic compression, thereby mapping the raw energy values in the spectrogram onto a decibel scale.

The advantage of storing third-octave spectrograms compared to audio waveforms is the drastic reduction in the volume of data to be transferred and stored. Indeed, the throughput of spectrogram data, as encoded in compressed JSON files, is on the order of 4.4 kilobytes per second (kbps) on average. In comparison, audio in "CD quality" (16-bit PCM at 44.1 kHz) has a bit rate of 705.6 kbps, while compressed audio in MP3 format has a typical bit rate of 128 kbps.

https://doi.org/10.1121/10.0005277

C. Scalability and privacy considerations

The drastic reduction of bit rate caused by spectrogram analysis on the edge has two implications. First, it alleviates the technical constraints of transmission and storage that are associated with acoustic monitoring. Instead of carrying audio over a high-throughput channel, such as Ethernet or VoLTE (Poikselkä *et al.*, 2012), it becomes possible to adopt emerging protocols for low-rate wireless personal area networking (LR-WPAN), such as 6LoWPAN or LoRaWAN (Turchet *et al.*, 2020). Although the present paper relies purely on data acquired via Ethernet communication, CENSE has recently extended its network to encompass solar-powered wireless sensors (Ardouin *et al.*, 2018). Beyond the example of CENSE, the edge computing paradigm has the potential to improve the scalability of machine listening systems by offloading server-side applications (Cerutti *et al.*, 2020).

Second, the fact that, except in few well-controlled cases (see Sec. III A), the CENSE network does not store waveforms permanently brings a guarantee of privacy. A prior publication has demonstrated that low-resolution spectrograms do not contain the necessary information to recover intelligible speech, even with state-of-the-art spectrogram inversion techniques (Gontier *et al.*, 2017). This is an instance of "privacy by design and by default" (Romanou, 2018), in compliance with the European General Data Protection Regulation (GDPR). An alternative approach, proposed by Cohen-Hadria *et al.* (2019), would be to acquire waveform audio, detect the presence of voice, and obfuscate it with digital audio effects. Although this approach is promising, we note that it remains error-prone and is still at an experimental stage. We refer to Appendix B for more details on privacy preservation in the CENSE project.

D. Audio2Vec pretext task

The gist of the Audio2Vec task is to learn sequential associations between neighboring snippets in real-world sounds (Chung and Glass, 2017; Tagliasacchi *et al.*, 2020). This task is inspired by Word2Vec, a well-established technique in natural language processing (Mikolov *et al.*, 2013). The analogy between Audio2Vec and Word2Vec consists in seeing acoustic scenes as sentences and short-term audio snippets as their constituent words. Once represented in the time–frequency domain, these snippets form spectrogram "clips" of fixed duration.

Just like Word2Vec, Audio2Vec comes in two flavors: continuous bag of words (CBoW) or skip-gram (SG). In the CBoW formulation, the self-supervised model takes one spectrogram clip as input and attempts to predict the content of a predefined number of adjacent clips, i.e., the past ones and future ones. Conversely, in the SG formulation, it takes a predefined number of disjoint clips as input and attempts to predict the content of the central clip. Both formulations of Audio2Vec resemble context encoders in computer vision (Pathak *et al.*, 2016), in the sense that they perform self-supervised feature learning by inpainting unobserved portions of a two-dimensional (2D) input. We adopt the SG formulation in this article. We set the context length to two past clips and two future clips, with one skipped clip at the center.

In accordance with the original implementation of Audio2Vec, we set the clip duration equal to 1 s, i.e., eight non-overlapping frames of 125 ms. However, note that the original implementation of Audio2Vec operates on mel-frequency spectrograms with a higher frequency resolution (64 mel-frequency bins in the range from 60 to 7800 Hz) as well as a higher time resolution (window size of 25 ms and hop size of 10 ms).

Figure 2 illustrates our implementation of the Audio2Vec SG task. Note that this task is a "pretext task" for self-supervised learning: it may be formulated on real-world acoustic scenes without any human intervention. In line with Word2Vec models, we solve the task by means of a deep generative encoder–decoder architecture. We share the synaptic weights of the encoder across all four context clips. We set the output dimension of the encoder to 128, hence a concatenated embedding of 512 for the entire context. This concatenated embedding serves as input to the decoder, which predicts a third-octave spectrogram clip of $29 \times 8 = 232$ decibel-scaled values.

E. Deep convolutional encoder–decoder architecture

Figure 3 illustrates the architecture of our encoder for the Audio2Vec SG task. It is a deep convolutional neural network (CNN) with six convolutional layers and one dense (fully connected) layer. Convolutional layers grow in width as depth increases, with, respectively, 64, 64, 128, 128, 256, and 256 filters. Each filter covers a receptive field of $3 \times 3 = 9$ adjacent values in the time–frequency activation of the previous layer. After the learned operations of convolution and additive bias, each layer applies batch normalization (Ioffe and Szegedy, 2015) and rectified linear unit (ReLU) activation. Furthermore, we apply maximum pooling to every other convolutional layer, i.e., three pooling layers in total. We set the pooling size to (2×2) and apply pooling without overlap, hence a subsampling by a factor of 2 in time and 2 in frequency. The last layer connects the "flattened" response of the third convolutional block onto 128 output units, i.e., the chosen dimension of the embedding. In total, the encoder contains 1.2×10^6 parameters.

Figure 4 illustrates the architecture of the decoder. This decoder takes a 512-dimensional vector as input and predicts a third-octave spectrogram clip of shape 29×8, corresponding to 1 s of audio. Following the SG formulation of Audio2Vec, we train the encoder and decoder jointly: the 512-dimensional input to the decoder results from the concatenation of 128-dimensional embeddings produced by the encoder over four context clips.

The sequential composition of layers in the Audio2Vec encoder and decoder are symmetric to each other. The decoder contains one dense layer followed by six convolutional layers of decreasing widths: 256, 128, 128, 64, 64, and 1 filter, respectively. We re-use the same receptive field

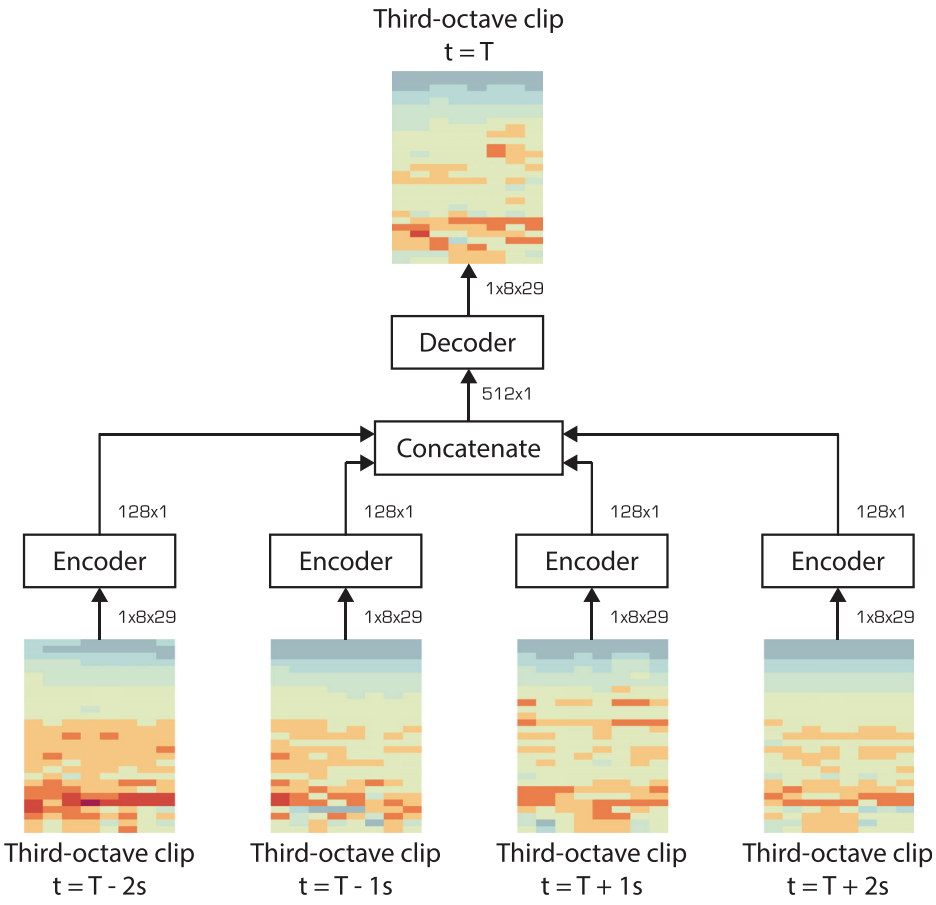

FIG. 2. (Color online) Encoder–decoder architecture used to solve the Audio2Vec pretext task by predicting a third-octave spectrogram clip from neighboring clips. See Sec. II D for details.

size (3×3), batch normalization, and activation function (ReLU) in the encoder and decoder. However, we replace maximum pooling in the encoder by nearest-neighbor upsampling in the decoder. The decoder has roughly the same overall number of parameters as the encoder, i.e., around 1.2×10^6.

F. Self-supervised training on the CENSEgram-5M dataset

Between December 1 and 5, 2019, we collected third-octave spectrogram data from 16 urban acoustic sensors belonging to the CENSE network. The resulting dataset,

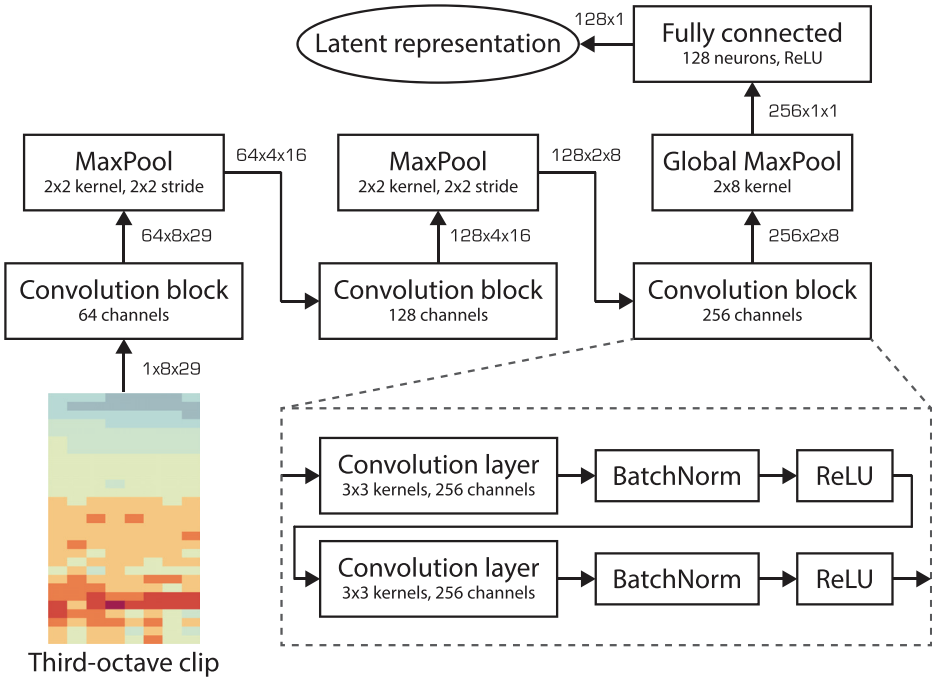

FIG. 3. (Color online) Encoder architecture used to extract information from 1 s third-octave spectrogram clip. The architecture is common to the pretext and downstream tasks in a self-supervised learning setting. See Sec. II E for details.

https://doi.org/10.1121/10.0005277

named CENSEgram-5M, contains a total of 4.6×10^6 s of spectrogram data, i.e., 1280 h. We divide CENSEgram-5M into training, validation, and test subsets, with proportions of 70%, 20%, and 10%, respectively. In doing so, we ensure that these subsets are temporally disjoint across all sensors.

We initialize the encoder and decoder with random independent and identically distributed weights and train them on the CENSEgram-5M dataset in an end-to-end fashion. We define the reconstruction error of a given spectrogram clip in terms of Euclidean distance on a decibel scale. With the encoder–decoder architecture, we aim to minimize the mean square error (MSE) across all spectrogram clips in the training set (3.2×10^6 s). To this end, we apply the Adam algorithm (Kingma and Ba, 2014), an improved variant of stochastic gradient descent with the PyTorch framework (version 1.1.0). We set the learning rate to 10^{-3} and leave all other hyperparameters of Adam to their default values. We train for 20 epochs (65×10^6 samples in total) with a minibatch size equal to 64. The training lasted about 20 h on a graphics processing unit (GPU), i.e., NVIDIA RTX2080Ti.

At the random initialization of the deep neural network, the MSE of the Audio2Vec task on the validation set is equal to 1973.35. This MSE decreases exponentially during training until reaching a plateau. At epoch 20, the MSE of Audio2Vec is equal to 9.29, i.e., 2 orders of magnitude below the MSE at epoch zero. This decrease indicates that the self-supervised learning stage updates the deep neural network toward solving the pretext task, as expected.

III. TRAINING SET SYNTHESIS WITH SIMSCENE

This section proposes to synthesize a training set for urban sound classification by means of a software library named simScene, which is publicly available (Lagrange, 2018). Specifically, we collect real-world monophonic recordings of sources of interest in Lorient (traffic, voice, and birds) to build the CENSE-2k dataset. We then assemble them to form a dataset of polyphonic scenes named simCENSE-18k.

A. Semi-automatic curation of monophonic acoustic events

We aim to curate a dataset of audio samples for three sources of interest: traffic, voice, and birds. In this regard, a widespread approach is to record audio data in bulk by means of a sensor network and then to annotate these sources manually in terms of presence or absence. Despite its computational simplicity, this approach incurs a high amount of human labor, especially for infrequent sources. Furthermore, in the case of the CENSE project, the principle of "privacy by default" forbids the bulk collection of audio data in the waveform domain. Rather, the CENSE network may only store third-octave spectrograms in its normal operation regime, while the collection of audio waveforms must be kept to a minimum.

Thus, for reasons of scalability and privacy, we restrict the collection of audio data to the three sources of interest by means of a template matching algorithm implemented "on the edge." Specifically, we begin by curating a small corpus of public-domain audio samples from the Freesound archive. We compute the third-octave spectrogram corresponding to each of these samples and average them over time to produce a spectral template. In addition, we define a "flat" third-octave template, corresponding to the power spectral density of pink noise, to extract "neutral" background noise. Then we implement a template matching algorithm, based on the cosine similarity in the third-octave spectrogram domain, on four sensors from the CENSE network.

Between May 1 and July 1, 2020, these sensors recorded 182 waveform audio samples, each of them corresponding to a high cosine similarity with one of the templates. We listen to each of these audio samples to verify that they contained an example of the source of interest. We trim their duration to exclude silent regions. On some *voice*

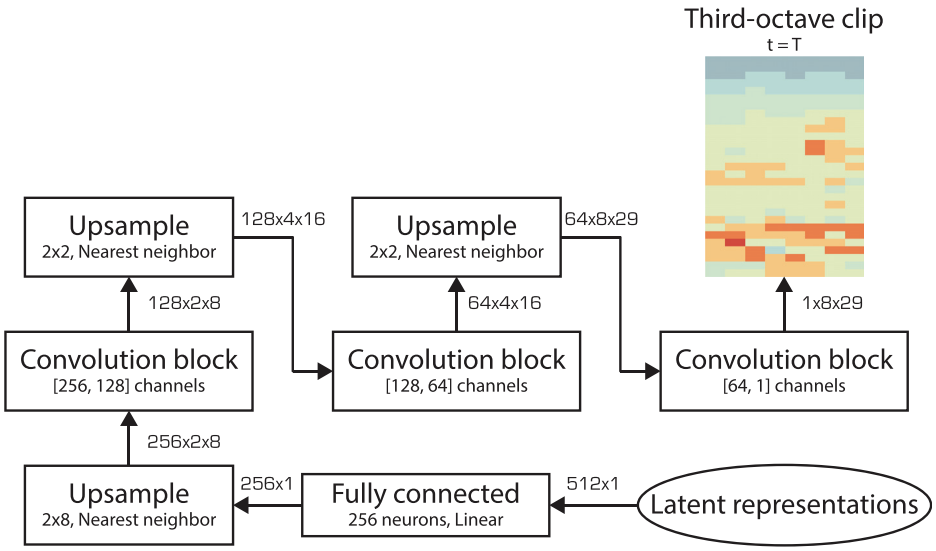

FIG. 4. (Color online) Decoder architecture proposed to solve the Audio2Vec pretext task. See Sec. II E for details.

TABLE I. Specifications of the CENSE-2k database.

Source type	Source class	Extracts	Duration (min)
Background	Neutral noise	16	3:41
Event	Traffic	128	31:20
	Voice	10	2:22
	Birds	28	3:19

and *bird* samples, we apply background noise reduction by means of the Adobe Audition software to enhance the presence of the source.

Table I summarizes the content of the resulting dataset, named CENSE-2k. These 182 samples have a maximum duration of 10 s and a total duration of 41 min, i.e., 2400 s. Note that this total duration corresponds to about 0.01% of the operating time of the template matching algorithm, i.e., 2 months in four sensors. We refer to Appendix B for more details on privacy preservation in the CENSE project.

B. Probabilistic simulation of polyphonic scenes

Simulation tools allow the production of large datasets of polyphonic sound scenes. Specifically, the simScene library (Lafay *et al.*, 2016) simulates a sound scene from two components, as illustrated in Fig. 5. The first component is a small-scale dataset of monophonic samples, in our case CENSE-2k. The second component, named "scenario," refers to the activity of sound sources through time. This scenario includes one background source, whose properties remain stationary throughout the scene, and foreground sources, which are characterized by a class, an onset time, and an "event-to-background ratio" in dB.

The simScene library generates original scenarios by sampling from distributions that set source activity parameters. Event sources are then associated with a probability of appearance, as well as Gaussian distributions of inter-onsets and event-to-background ratios. Source probabilities of appearance and activity distributions are obtained by manually annotating a corpus of 74 sound scenes recorded in Paris (Aumond *et al.*, 2017). This annotation is introduced in Gloaguen *et al.* (2019). Distributions are conditional to five types of sound environments: *quiet street*, *noisy street*, *very noisy street*, *park*, and *square* environment type with predominant voice content (Gontier *et al.*, 2019). This conditioning should allow a more complete coverage of typical urban situations by simulated corpora.

The simScene library instantiates every scenario by selecting monophonic samples uniformly at random from within the CENSE-2k database, according to the class of foreground acoustic events. Each of these monophonic samples is scaled in amplitude according to event-to-background ratios and shifted in time according to onset timestamps. Last, simScene combines all acoustic events with the background audio track to produce a polyphonic mixdown.

A development set designated simCENSE-18k is simulated with the CENSE-2k database presented in Sec. III A. The dataset contains 400 scenes of 45 s (total duration 5 Hz). It is balanced in terms of types of sound environments, i.e., 80 scenes are simulated for each type. After simulation, each sound scene is scaled to a sound level, drawn randomly from Gaussian distributions conditionally to the type of environment. For training purposes, the simCENSE-18k dataset is split into training and validation subsets containing 70% and 30% of simulated scenes, respectively. This split is done for each type of sound environment to conserve balanced characteristics.

C. Virtual annotator

Due to the additive combination process, ground truth contributions of types of sound sources in simulated scenes are known. This information enables trivially labeling acoustical source presence or absence by application of an energy threshold on individual source contributions. However, the present study is oriented toward urban sound classification in a perceptual context. In some cases, auditory masking may occur in critical bands of the spectrum, where a source is not heard within the mixed scene despite being objectively present. Thus, we define the presence of a source in terms of its audibility in context, that is, as an element of a polyphonic scene. This definition approximately

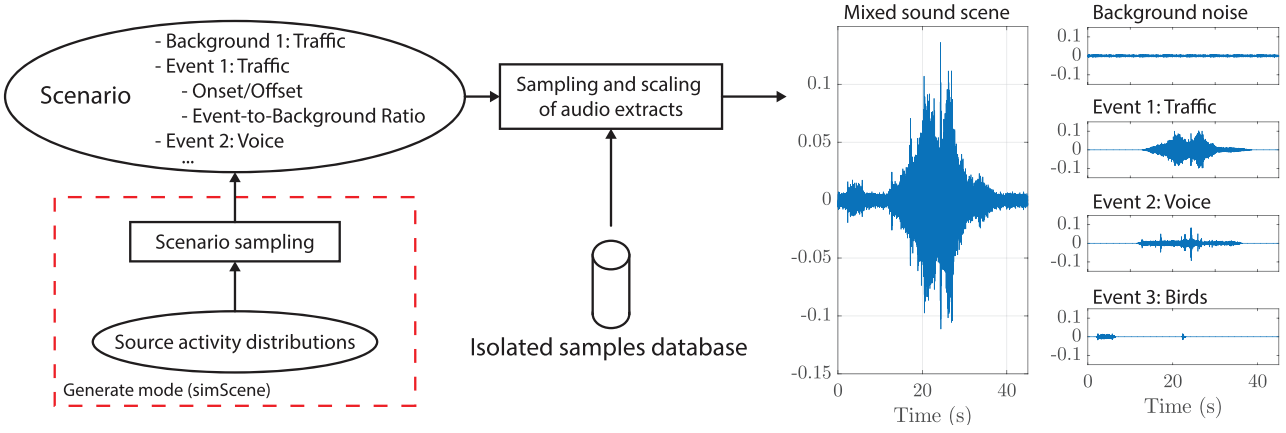

FIG. 5. (Color online) Overview of the scene simulation process from a scenario and a database of monophonic source samples.

https://doi.org/10.1121/10.0005277

corresponds to the behavioral response of an active human listener—e.g., a pedestrian passing by—who would be continuously attending to all concurrent sound sources.

We simulate auditory masking in polyphonic acoustic scenes via the method of Gontier *et al.* (2019), which is based on relative energy thresholding in the time–frequency domain. Let $\mathcal{S}$ denote the set of sources of interest: in our case, $\mathcal{S}$ contains *traffic*, *voices*, and *birds*. For any given source s in S, we denote by $\mathbf{X}_s(t,f)$ the third-octave decibel-scaled spectrogram associated with the audio track of s in the virtual polyphonic scene. Furthermore, we denote by $\mathbf{X}_{\bar{s}}(t,f)$ the third-octave decibel-scaled spectrogram associated with all other tracks; that is, sources $s' \notin \mathcal{S}$ as well as the background noise track.

To model auditory masking, a binary time–frequency mask, proposed in Gontier *et al.* (2019), is computed from the third-octave ratio of energy $\Delta X_s(t,f)$ of source s with respect to all other sources combined $\bar{s}$,

$$\Delta X_s(t,f) = X_s(t,f) - X_{\bar{s}}(t,f), \tag{1}$$

where X_s is the third-octave spectrogram in dB SPL. The first processing step determines whether the source of interest is present on each time frame t and frequency band f by application of a threshold parameter α to the emergence spectrogram,

$$Y_{s,\alpha}(t,f) = \mathbb{1}_{\Delta X_s(t,f)>\alpha}, \tag{2}$$

where $\mathbb{1}$ denotes the indicator function. Thus, the operation returns 1 if the emergence is greater than α and 0 otherwise. A single presence label for a given time frame is obtained by averaging the source emergence on selected bands and applying a second threshold β,

$$Y_s(t) = \mathbb{1}\left[\frac{\sum_{f=1}^{N_f} \Delta X_s(t,f)\,\mathbb{1}_{\Delta X_s(t,f)>\alpha}}{\sum_{f=1}^{N_f} \mathbb{1}_{\Delta X_s(t,f)>\alpha}} > \beta\right]. \tag{3}$$

Note that the average over the N_f third-octave bands is taken for logarithmic sound levels; thus, it cannot be interpreted as a sound level.

The values of hyperparameters α and β have been optimized by (Gontier *et al.*, 2019) with respect to subjective assessments collected during a listening test. In this paper, we re-use the optimal values that arose from that listening test, i.e., $\alpha = -14$ dB and $\beta = -7$ dB.

D. Urban sound classification model

The downstream task is the multilabel classification of three sources of interest—*traffic*, *voice*, and *birds*—at the time scale of 1 s. Figure 6 describes the system we propose to address the downstream task. It is a CRNN taking the third-octave spectrogram of a 1-s audio clip as input. The acoustic frontend corresponds to the encoder of the Audio2Vec pretext task (see Sec. II). We set the hop length between adjacent clips to 125 ms, i.e., one spectrogram frame.

This encoder produces a 128-dimensional representation of the current audio clip, which feeds a single-layer

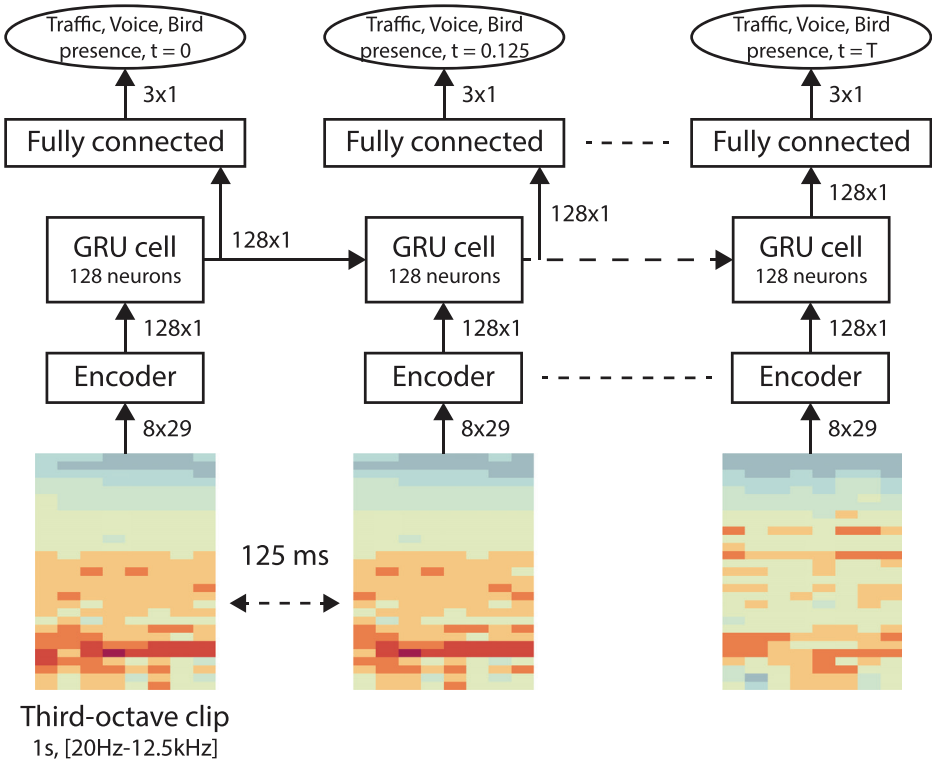

FIG. 6. (Color online) Decision architecture for the downstream task of source presence prediction.

gated recurrent unit (GRU), whose recurrent state comprises 128 neurons (Cho *et al.*, 2014). Every second, this recurrent state is passed through a LeakyReLU nonlinearity whose slope equals 0.1 and to a fully connected layer.

The fully connected layer predicts a three-dimensional vector $\boldsymbol{y}$, containing the predicted probability of presence of all three sources of interest. The entire neural network architecture is trained in an end-to-end fashion so as to minimize the binary cross-entropy (BCE) loss function,

$$\mathrm{BCE}(y,\hat{y}) = -\sum_{s} y_s \log(\hat{y}_s) + (1-y_s)\log(1-\hat{y}_s), \quad (4)$$

where s is the source, $y_s \in \{0,1\}$ is the target presence label for source s, and $\hat{y}_s \in [0,1]$ is the predicted presence.

As in Gontier *et al.* (2019), we apply the Adam algorithm with default hyperparameters to train the model above on the simCENSE-18k dataset. We set the batch size to 32 and keep the learning rate constant at 10^{-4}.

IV. EVALUATION METHODOLOGY AND RESULTS

A. From localized software to localized machine learning

In the terminology of software development, "localization" refers to the adaptation of a software product to a group of users from a particular geographic region (Esselink, 2000). Such adaptation includes, for example, the translation of texts appearing on screen to a different language, the conversion of physical units, the formatting of times and dates, and other national or cultural conventions. Our paper proposes to borrow from this terminology and extend it to the realm of machine learning, specifically urban sound classification.

Thus, we envision the combination of self-supervised learning and training set synthesis as a procedure for data-efficient "localization of models." In other words, our ambition is not to advance the state of the art in general-purpose audio classification (e.g., AudioSet), but rather to focus on the acoustic events that are most relevant to a specific location. We approach this problem as the auditory equivalent to "recognition in terra incognita" in computer vision (Beery *et al.*, 2018).

Going back to the example of the CENSE project, the location of interest is the city of Lorient (France). Moreover, the relevant acoustic events correspond to the known predominant factors to the perceived pleasantness of urban soundscapes in Lorient (Gontier *et al.*, 2019): *traffic*, *voice*, and *birds*.

B. Recording and annotation of the Lorient-1k dataset

To evaluate a "localized machine learning model" (see above), the evaluation set must belong to the same location as the training set. However, collecting the evaluation set with the same hardware equipment as the training set would give an unfair advantage to our methodology in comparison with pre-trained systems. Indeed, in the absence of any acoustic matching frontend (Su *et al.*, 2020), deep learning systems tend to overfit the frequency response of the microphone in the training set (Das *et al.*, 2014). However, we aim to specialize the model to a particular location while generalizing to potentially unseen recording equipment. Therefore, while our training set purely consists of sensor network data (CENSE), we decided to collect an evaluation set with handheld devices.

On July 30, 2020, we visited ten different locations in the downtown area of Lorient, within the area of coverage of the CENSE network. With Zoom H4n handheld devices, we recorded 30 acoustic scenes of 45 s each. The microphone was attached at the top of a backpack, thus at approximately 2 m high, which leads to a discrepancy of recording height between the training and the evaluation data of about 1 m.

In parallel, we measured the A-weighted sound pressure level of the scenes by means of a class-1 sound level meter. In this way, we were able to calibrate the amplitude of the waveform incoming from the Zoom H4n device according to the measured A-weighted decibel level (dBA). The resulting dataset, named Lorient-1k, has a total duration of 22.5 min, i.e., 1350 s.

The 30 recorded sound scenes are then annotated by a panel of four researchers with expertise on acoustics or auditory perception. Following the recommendations of Cartwright *et al.* (2017), we allow all participants to pause, repeat, and view spectrograms during the annotation process. The participant first annotates the scenes in terms of perceived sound presence and then annotates them in terms of activity onsets and offsets for each of the three sources, respectively, *traffic*, *voice*, and *birds*. No distinction is made between subclasses of sounds, e.g., small birds and seagulls. Participants are instructed not to change the software audio level during the procedure to preserve relative sound levels between sound scenes. To match the hop and clip sizes of the predictors, annotations are sampled with 125 ms, and sound events separated by less than 1 s are merged.

Starting from those annotations, an expert in the perception of the urban acoustic environment (C.L. of the authors), provides a unique ground truth to compare the predictors with, using the following procedure. First, she annotates the 30 scenes in terms of time of presence for each source by taking into account the time of presence of the four annotators. For the traffic class, some discrepancies between annotators are observed due to the lack of formal definition of the *traffic* class, from the background noise to the pass-by of vehicles.

For her "main" annotation, C.L. has tried to understand the strategies chosen by the annotators and kept a middle-ground approach for strategies in line with literature about urban sound quality. She eliminates the annotations that were insufficiently compliant with this definition. In the case of *traffic*, she only considers the sound as active when the sound due to the traffic class is not perceived as fully stationary within a period of a few seconds.

Second, she annotates activity onsets and offsets for each of the three sources and goes back to the first annotation step if necessary so that the difference between the sum

https://doi.org/10.1121/10.0005277

of the time spans of each source and the perceived time of presence is below 10%. This value is assumed as satisfactory, as it can be considered to be the human precision in terms of perceived time of presence (Aumond *et al.*, 2017).

C. Current state of the art: YAMNet pre-trained classifier

Our approach requires the collection of local audio material and the training of dedicated machine learning models. To motivate the need for such a procedure, we evaluate the comparative performance of a state-of-the-art system that does not require any of those interventions.

YAMNet is a deep neural network that predicts 521 audio event classes and is pre-trained on the AudioSet-YouTube (Gemmeke *et al.*, 2017) corpus of more than 2×10^6 sound clips.[1] It relies on the Mobilenet_v1 (Howard *et al.*, 2017) depthwise-separable convolution architecture. We employ YAMNet as an off-the-shelf system without any form of transfer learning. Rather, we use domain-specific knowledge to aggregate labels semantically: for example, *squawk*, *caw*, and *hoot* all fall under the umbrella term *birds* for our purpose. Appendix A describes exhaustively how we map AudioSet labels onto our three sources of interest, namely, *traffic*, *voice*, and *birds*.

On the Lorient-1k dataset, YAMNet achieves a classification accuracy of **71.4**% on average: 63.9% for *traffic*, 80.5% for *voice*, and 69.8% for *birds*. The lower performance of *traffic* and *birds* in comparison with *voice* can partly be explained by the class imbalance of training data in AudioSet: whereas *speech* appears in 50% of AudioSet examples, *car* and *bird* appear in 2% and 1% of examples, respectively. Another plausible explanation is the choice of time–frequency representation that is passed as input to YAMNet. Indeed, this representation is a mel-frequency spectrogram with 64 filters ranging between 125 Hz and 7.5 kHz, that is, roughly the bandwidth of human voice, yet frequencies below 125 kHz have a crucial role in denoting the presence of traffic. Conversely, some bird families, such as warblers and sparrows, often vocalize above 7.5 kHz (Lostanlen *et al.*, 2018). In comparison, our third-octave spectrogram (see Sec. II B) covers the audible range from 20 Hz to 12.5 kHz.

V. BENCHMARK

In this section, we analyze the main factors to overall performance in our proposed model. Figure 7 summarizes our results.

We present five variations of our proposed neural network architecture, that is, a deep CRNN taking third-octave spectrograms as input representation (see Secs. II B, II E, and II D). All five models have the same number of trainable parameters and the same computational complexity at prediction time, yet the models differ in terms of how they were trained: with or without self-supervised learning, with or without polyphonic training set synthesis, and with local or external annotated data.

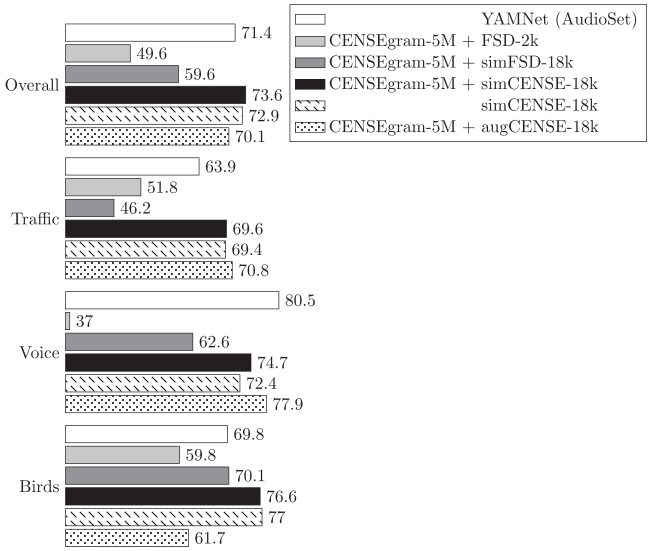

FIG. 7. Overall, traffic, voice, and bird accuracies in percentages achieved by the several flavors compared to the YAMNet detector.

A. Baseline: Local self-supervised learning followed by external supervised learning

We begin with a simple-minded pipeline in which we skip local data curation (Sec. III A) training set synthesis (Sec. III) entirely. Rather than curating a local dataset of sound events for the city of Lorient (CENSE-2k), we collect these sound events via external sources. Specifically, we download audio samples of traffic and bird vocalizations from the Freesound archive and speech samples from the Librispeech (Panayotov *et al.*, 2015) dataset. In this way, we obtain a freely licensed dataset containing monophonic examples of all three sources of interest: *traffic*, *voice*, and *birds*.

The resulting dataset, which we name FSD-2k (FreeSound Dataset, 2k seconds), has the same size and event taxonomy as CENSE-2k. However, FSD-2k differs from CENSE-2k in terms of intra-class variability: for example, FSD-2k contains speech in other languages besides French and bird vocalizations from species whose habitats exclude the region of Lorient. We split recordings from FSD-2k into segments of 3 s or less. We mix 3 s of low-level pink noise with each segment to guarantee that all segments last this duration exactly. Because all segments are monophonic, it is trivial to obtain labels of source activity for these samples. Table II summarizes the contents of this dataset on a per-source basis. We refer to Appendix C 6 for more details on FSD-2k.

As a first stage, we train the encoder for the baseline model on CENSEgram-5M (see Sec. II F) in a self-supervised way by means of the Audio2Vec pretext task. We call this first stage "local self-supervised learning" because the unlabeled spectrogram data of CENSEgram-5M are local to the city of Lorient. Then, as a second stage, we transfer the weights of this pre-trained encoder into a CRNN architecture to solve the downstream task of multilabel sound classification. We train this CRNN on FSD-2k in an end-to-end fashion, i.e., "fine-tuning" the encoder weights instead of

TABLE II. Contents of the isolated samples database used to generate simFSD-18k.

Source type	Source class	Extracts	Duration (min)
Background	Neutral noise	16	3:41
Event	Traffic	82	24:50
	Voice	160	9:10
	Birds	22	4:18

"freezing" them. We call this second stage "external supervised learning" because the labeled audio data of FSD-2k come from online audio archives, most of them external to the city of Lorient. The motivation behind this baseline is that it does not require any annotation of local data: CENSEgram-5M is local but unlabeled, whereas FSD-2k is already labeled by the original contributors of FreeSound and LibriSpeech.

On the hold-out test set (Lorient-1k), the baseline classifier reaches a classification accuracy of 49.6% on average, i.e., about 20% points below YAMNet (71.4%, Sec. IV C). This result suggests that our pretext task, Audio2Vec, does not suffice to encode the properties of urban auditory environments. In the baseline procedure, there is a mismatch between the self-supervised learning stage, which is local and polyphonic, and the supervised learning stage, which is external and monophonic. We hypothesize that, after having learned the peculiarities of the local urban environment (CENSEgram-5M) during the pretext task, the encoder "forgets" these peculiarities when being fine-tuned on an external dataset (FSD-2k) to solve the downstream task.

B. Role of polyphonic training set synthesis

To improve the baseline performance, we apply polyphonic training set synthesis on the external dataset of urban sound recordings, FSD-2k. Specifically, we run the simScene algorithm as described in Sec. III). In doing so, we take foreground samples (*traffic*, *voice*, and *birds*) from an external source (FSD-2k) and background samples from a local source (CENSE-2k). We obtain 400 polyphonic scenes of duration equal to 45 s each, yielding a dataset that we call simFSD-18k (simulated FreeSound Dataset, 18k seconds). We refer to Appendix C 7 for more details on simFSD-18k.

Not only does simScene increase the amount of training data by a factor of 6 or so; it also makes the training data more reflective of the downstream task: that is, multilabel urban sound classification. Indeed, many spectrogam clips in the hold-out test set (Lorient-1k) contain overlapping sound events from different classes: see Sec. IV B.

After self-supervised training on CENSEgram-5M and supervised training on simFSD-18k, the model reaches a classification accuracy of **59.6**% on average. Although this result remains below YAMNet performance (71.4%, Sec. IV C), it fares ten percentage points above the baseline (49.6%, Sec. V A). This result suggests that polyphonic training set synthesis has a crucial role to play in improving the generalization of self-supervised learning for sound event classification. To our knowledge, our paper is the first in reporting an experimental observation of this kind.

C. Role of local data collection and best performing system

We now turn to the question of training the model with annotated local data, as opposed to external data, for the downstream task of multilabel sound event classification. To address this question, we pass the CENSE-2k dataset of local sound events from Lorient (Sec. III A) as input to simScene. Thus, we obtain 400 polyphonic scenes of duration equal to 45 s each, yielding a dataset that we call simCENSE-18k. The simCENSE-18k and simFSD-18k datasets have the same total duration and share the same hyperparameters in terms of per-source probability distributions (Sec. III B). However, simCENSE-18k and simFSD-18k differ in terms of their acoustic constituents. In simCENSE-18k, both the background noise and the foreground events come from CENSE-2k and are thus local to the city of Lorient. Meanwhile, in simFSD-18k, the background noise is from a local source (CENSE-2k), but all foreground events come from external sources (FSD-2k, i.e., FreeSound and LibriSpeech).

After self-supervised training on CENSEgram-5M and supervised training on simCENSE-18k, the model reaches a classification accuracy of **73.6**% on average. This performance is above YAMNet, which we took as current state of the art (71.4%, see Sec. IV C). Crucially, our model surpasses YAMNet even though it was never exposed to any real-world annotated data: we have formulated the pretext task on real-world unlabeled data and the downstream task on synthetic labeled data. Furthermore, to allow a fair comparison with YAMNet, our test set (Lorient-1k, see Sec. IV B) does not come from the sensor network that produced the local training data but from mobile handheld devices.

On the flip side, we note that our system surpasses YAMNet on average but does not do so on every sound source. Specifically, the per-source classification accuracy of our system is: 69.6% for *traffic*, 74.7% for *voice*, and 76.6% for *birds*. Thus, our system outperforms YAMNet in the classification of *traffic* and *birds* but remains below YAMNet on the classification of *voice*. Section VI will discuss the possible causes of such discrepancy.

D. Ablation of self-supervised learning

We have seen thus far that polyphonic training set synthesis improves self-supervised learning, and even more so when the isolated events that serve as input to simScene come from the same recording location (Lorient in our case) as the test set. It remains to be seen whether self-supervised learning plays an essential role in our proposed pipeline or whether polyphonic training set synthesis suffices on its own.

To answer this question, we experiment with removing the self-supervised learning stage from our pipeline. Instead of using the large-scale unlabeled spectrogram data in CENSEgram-5M for the pretext task, we initialize the deep neural network with random Gaussian weights (Giryes *et al.*, 2016) and train the CRNN directly on simCENSE-18k in a supervised fashion.

https://doi.org/10.1121/10.0005277

After ablation of self-supervised learning, the classification accuracy slightly decreases from 73.6% to 72.9% on average. On a per-class basis, the accuracy breaks down as 69.4% for *traffic*, 72.4% for *voice*, and 77.0% for *birds*. Interestingly, the impact of self-supervised learning is more noticeable on the *voice* class than on *traffic* and *birds*. Although this anecdotal finding deserves further inquiry, we note that *voice* is the least frequent class in CENSE-2k, with only ten different extracts (see Table I). This is because speech is rarely heard in isolation in an urban acoustic sensor network such as CENSE. Thus, *voice* is also the least varying source in simCENSE-18k, with ten or so different speaker identities. Meanwhile, our unlabeled dataset for self-supervised learning (CENSEgram-5M) contains over 1000 h of audio data and thus potentially thousands of different speaker identities. Together, these observations suggest that self-supervised learning potentially mitigates class imbalance and strengthens the detection of acoustic events that infrequently appear on their own.

E. Ablation of polyphonic training set synthesis

Last, we experiment with the removal of polyphonic training set synthesis while maintaining self-supervised learning and local data curation. As shown in Secs. V A and V B, simScene brings a noticeable performance gain: from 49.6% to 59.6% classification accuracy on average. However, the cause of this performance gain might not be the shift from monophonic (CENSE-2k) to polyphonic (simCENSE-18k) training but, more simply, from a sixfold increase in the amount of training data.

To refute this hypothesis, we perform artificial data augmentation on the monophonic examples of CENSE-2k, thus yielding a new synthetic dataset, which we name augCENSE-18k. Following Salamon and Bello (2017), our data augmentation procedure includes pitch shifting in a range of −6 to 6 semitones and time stretching with a factor from 0.7 to 1.3 or a combination of both at random. In doing so, we augment monophonic samples from the *voice* and *birds* sources, therefore matching the duration of the most frequent class: *traffic*. In this way, the total duration of augCENSE-18k is approximately 5 h, i.e., the same as simCENSE-18k. Furthermore, by construction, augCENSE-18k is free from any class imbalance, unlike CENSE-2k.

After self-supervised learning on CENSEgram-5M and supervised learning on augCENSE-18k, the model achieves an average classification accuracy of **70.1**% on the Lorient-1k test set. This underperforms our best performing model (73.6% on average), which is trained on simCENSE-18k during the supervised stage. It also fares below the YAMNet off-the-shelf classifier (71.4% on average) which is trained on AudioSet. That being said, the replacement of polyphonic training set synthesis (simCENSE-18k) by artificial data augmentation (augCENSE-18k) produces a model that remains competitive with the state of the art. We observe that the combination of artificial data augmentation and polyphonic training set synthesis, as proposed by the Scaper library, for example (Salamon *et al.*, 2017a), does not improve the state of the art any further in our experimental benchmark, yet we believe that this sort of combination deserves further investigation.

VI. DISCUSSION

This section proposes some qualitative comments on the outcome of our experimental benchmark.

A. Hardware mismatch between training and test set

Even with microphones used for all the sensors of the same type and model, difference in manufacturing and in aging may lead to subtle changes in their frequency properties (Picaut *et al.*, 2020). To design detectors that are more resilient to change in frequency response, we thus chose to use different microphones with different frequency responses to build the training sets (CENSEgram-5M and simCENSE-18k) and the evaluation set (Lorient-1k).

We assume that the impact of this discrepancy is weak, but sensitivity and evaluation of robustness of classifiers to modifications of the frequency response should be addressed in further research nonetheless. At this stage, using different microphones for training and evaluation is a good way to ensure that this matter does not compromise our conclusion on the performance of the proposed system.

B. Choice of pretext task

The gain achieved in this study with pretext-based learning is rather marginal. This is potentially due to two factors. First, the design of the Audio2Vec pretext task is unrelated to a choice of downstream task. Indeed, it relies purely on the mutual information between successive clips. Other pretext tasks may be considered in the future: for example, the TriCycle task (Cartwright *et al.*, 2019a), which relies on domain-specific knowledge about cycling patterns in urban acoustic sensor networks. Second, we expect that increasing our training set, both in terms of time span and in terms of spatial coverage, will improve the generalization of self-supervised learning.

C. Diversity of detected sources

From an application perspective, there is a need to expand the diversity of sources to produce a more versatile predictive model for high-level auditory perception. This could be achieved by refining the taxonomy of urban sounds: for example, breaking down *traffic* into *motorcycle*, *car*, *truck*, and so forth. Likewise, the *birds* class could be divided into various orders, families, and ultimately species (Cramer *et al.*, 2020). One advantage of training set synthesis is that such an expansion in taxonomy would come at a moderate cost in human workload: indeed, the curation of a dataset of isolated samples for every source of interest suffices to train a deep learning system on polyphonic scenes.

D. Application to the prediction of high-level perceptual attributes

The original motivation of our study, as stated in the Introduction, is to predict the time of presence of three sources (*traffic*, *voice*, and *birds*) in urban acoustic scenes, as represented by third-octave spectrograms. Indeed, linear combinations between these three times of presence and overall loudness provide a good approximation of the perceived pleasantness of the scene as well as other high-level perceptual attributes (Aumond *et al.*, 2017).

For this reason, we evaluate all proposed models, not only in terms of accuracy, but also in terms of relative predicted time of presence. Specifically, for each of the three classes in our taxonomy, we compute the ratio between the number of positive frames and the total number of frames in the scene. This ratio ranges between zero and one and can be readily compared with human judgments. To this end, we use the square root of the mean square error (RMSE) as an auxiliary metric to the task of multilabel urban sound classification. Similarly to the micro-averaging of accuracy across all three classes of interest, we compute the RMSE of each source across all scenes and then average the per-source RMSE to obtain a unified metric per model. Note that accuracy is a "higher-is-better" metric, whereas RMSE is a "lower-is-better" metric.

We find an average RMSE of 60.3% for the baseline: local self-supervised learning on CENSEgram-5M followed by external supervised learning on FSD-2k (see Sec. V A). In comparison, the YAMNet model achieves an average RMSE of 35.4% (see Sec. IV C). Last, our full-fledged model (see Sec. V C) achieves a RMSE of 32.5%.

Together, these results suggest that accuracy improvements correlate with reductions in RMSE. That being said, we should take the RMSE results with caution because they are aggregated across scenes, whereas accuracy results are aggregated across frames. In Lorient-1k, there are 1350 frames per source yet only 30 scenes. This explains why we conducted our benchmark with accuracy as the primary metric of interest and measure the predicted time of presence only as a secondary metric.

Last, we should note that the new state-of-the-art performance, i.e., 73.6% accuracy and 32.5% RMSE, remains perfectible. Ideally, a reliable model for urban sound monitoring should estimate the time of presence of each source with a deviation of 10% or less with respect to human annotation. Future work is needed to close the gap in performance between deep learning models for urban sound classification and the response of expert listeners.

VII. CONCLUSION

A prior study (Gontier *et al.*, 2019) has demonstrated that CNNs are effective for estimating the time of presence of sources.

In this paper, we show that the design of the training methodology for learning the deep architecture in order to perform well for a given spatial location is critical.

In this paper, we employed YAMNet as a baseline system. This is a pretrained classifier learned on AudioSet, a very large dataset of more than 20×10^6 s of audio. Another option would be to extract a task-agnostic embedding, such as OpenL3 (Cramer *et al.*, 2019) or the penultimate layer or YAMNet. In this case, a classification layer would have to be learned on top of the embedding, which requires annotated local data.

We demonstrate that simulating polyphonic sound scenes is an efficient method of training set synthesis when the volume of labeled data is limited. Crowdsourcing audio annotations produces noisy labels (Fonseca *et al.*, 2019) and demands intensive unskilled labor; in contrast, training set synthesis relies on computer-generated labels and demands a small amount of highly skilled labor.

The simScene software library automates the process of large-scale training set synthesis and produces a "virtual annotation" as a by-product. This library leads to the generation of arbitrary polyphonic urban sound datasets of arbitrarily large duration, such as simCENSE-18k. Our results suggest that recording sound sources for training set synthesis at the same place where the sensor is operating is important for achieving good performance.

Some limitations of the study have been discussed in Sec. VI. The impact of the difference of microphone frequency response between sensors should be studied. The ambiguity between the background noise and the *traffic* source should be tackled both from a perception and computational architecture design point of view. The training set synthesis approach proposed in this paper seems to be complementary with self-supervised learning. While we employed a simple-minded pretext task (Audio2Vec) with a short receptive field of 1 s, we believe that the design of new pretext tasks with longer receptive fields has the potential to improve performance even further.

ACKNOWLEDGMENTS

This research was funded by the French National Agency for Research (Agence Nationale de la Recherche) Grant No. ANR-16-CE22-0012.

The data supporting this study are openly available from Zenodo, an open-access repository, at: CENSEgram-5M at https://doi.org/10.5281/zenodo.4687030, CENSE-2k at https://doi.org/10.5281/zenodo.4694522, simCENSE-18k at https://doi.org/10.5281/zenodo.4694524, Lorient-1k at https://doi.org/10.5281/zenodo.4687057, augCENSE-18k at https://doi.org/10.5281/zenodo.4733681, FSD-2k at https://doi.org/10.5281/zenodo.4730390, simFSD-18k at https://doi.org/10.5281/zenodo.4733698.

APPENDIX A: THE YAMNET BASELINE

For each clip, the three most likely sound classes among the 521 classes of the AudioSet ontology are selected. If one of those three classes belongs to the set of events of a given source E_s, the source s is set as active in this clip. The use of the YAMNet baseline thus requires the definition of source

https://doi.org/10.1121/10.0005277

TABLE III. Selected sound event set for the *traffic* set E_t. Only the classes with underlined IDs are considered for the small sound event set.

ID	Name
300	Motor vehicle (road)
307	Tire squeal
308	Car passing by
309	Race car, auto racing
310	Truck
315	Bus
320	Motorcycle
321	Traffic noise, roadway noise

TABLE IV. Selected sound event classes for the *voice* set E_v. Only the classes with underlined IDs are considered for the small sound event set.

ID	Name
0	Speech
1	Child speech, kid speaking
2	Conversation
3	Narration, monologue
4	Babbling
5	Speech synthesizer
6	Shout
7	Bellow
8	Whoop
9	Yell
10	Children shouting
11	Screaming
12	Whispering
13	Laughter
14	Baby laughter
15	Giggle
16	Snicker
17	Belly laugh
18	Chuckle, chortle
19	Crying, sobbing
20	Baby cry, infant cry
21	Whimper

TABLE V. Selected sound event classes for the *bird* set E_b. Only the classes with underlined IDs are considered for the small sound event set.

ID	Name
106	Bird
107	Bird vocalization, bird call, bird song
108	Chirp, tweet
109	Squawk
110	Pigeon, dove
111	Coo
112	Crow
113	Caw
114	Owl
115	Hoot
116	Bird flight, flapping wings

event sets E_s for the three classes of interest: *traffic*, *voice*, and *bird*.

Curating the AudioSet ontology for building the three source sets may bias the performance of the baseline. We thus consider two alternative curations: one with source sets of low cardinality and the other with higher cardinality. We list both variants in Table III for *traffic*, Table IV for *voices*, and Table V for *birds*. We observe that both variants are within one-tenth of a percentage point in terms of our evaluation metric, $\max\bar{\kappa}$. With this observation in mind, we consider only the low-cardinality variant of YAMNet in the remainder of this study.

APPENDIX B: ETHICS STATEMENT

1. Data acquisition

Local recordings of 10-s clips made from the network are stored on an offline server with fully end-to-end encrypted access, restricted to a few researchers within the project. Each audio clip is deleted if it contains intelligible speech, does not illustrate a source of interest, or is not monophonic. Monophonic samples, sound scenes produced from these samples, and the evaluation dataset are not distributed or made available publicly in the waveform domain.

APPENDIX C: SUPPLEMENTARY MATERIAL

Our study introduces seven new datasets. All of them have been made available along with the processing code (Lagrange, 2021). This section summarizes their characteristics.

1. CENSEgram-5M: A large-scale dataset of spectrograms from an urban acoustic sensor network in Lorient (France)

CENSEgram-5M contains third-octave spectrograms from the CENSE network of acoustic sensors. These spectrograms correspond to 5 days of continuous measurements obtained in December 2019 $\times$ 16 sensors. The total duration of the dataset is on the order of 5×10^6 s, i.e., 1280 h. Our paper uses CENSEgram-5M for self-supervised learning. See Sec. II for details.

2. CENSE-2k: An annotated dataset of sound events from an urban acoustic sensor network in Lorient (France)

CENSE-2k contains 182 monophonic audio clips from the CENSE network of acoustic sensors. One expert annotated these audio clips in terms of four classes: background noise, traffic, voice, and birds. Our paper uses CENSE-2k for training set synthesis. The total duration of the dataset is on the order of 2400 s, i.e., 40 min. Our paper uses CENSE-2k for training set synthesis. See Sec. III A for details.

3. simCENSE-18k: A dataset of synthetic acoustic scenes from Lorient (France)

simCENSE-18k contains 400 acoustic scenes of duration equal to 45 s. We synthesized these polyphonic scenes via the simScene software, based on monophonic audio clips from the CENSE-2k dataset. The total duration of the dataset is equal to 18 000 s, i.e., 5 h. Our paper uses simCENSE-18k to train a multilabel classifier of urban sounds on synthetic data. See Sec. III for details.

4. Lorient-1k: An annotated dataset of acoustic scenes from handheld devices in Lorient (France)

Lorient-1k contains 30 acoustic scenes of duration equal to 45 s. These scenes were recorded with Zoom H4n handheld devices at 10 different locations of Lorient (France). Four experts annotated the onset and offset times of three sources of interest: traffic, voice, and birds. The total duration of the dataset is on the order of 1350 s, i.e., 22.5 min. Our paper uses Lorient-1k as a hold-out evaluation set for benchmarking urban sound classifiers. See Sec. IV for details.

5. augCENSE-18k: An artificially augmented version of CENSE-2k

augCENSE-18k is a derivative of CENSE-2k, obtained by time stretching and pitch shifting audio clips of the *voice* and *birds* classes at random. The total duration of the dataset is equal to 18k seconds, i.e., the same as simCENSE-18k, with balanced material over classes. Our paper uses augCENSE-18k to demonstrate that artificial data augmentation (yielding augCENSE-18k) underperforms training set synthesis with simScene (yielding simCENSE-18k). See Sec. V for details.

6. FSD-2k: A dataset of events collected from Freesound and Librispeech

FSD-2k contains about 200 monophonic audio clips collected from online resources, which are unrelated to the city of Lorient: Freesound for birds and traffic and Librispeech for voice.

7. simFSD-18k: A dataset of synthetic acoustic scenes made with Freesound and Librispeech samples

simFSD-18k contains 400 acoustic scenes of duration equal to 45 s. We synthesized these polyphonic scenes via the simScene software, based on FSD-2k. The total duration of the dataset is equal to 18000 s, i.e., the same as simCENSE-18k. Our paper uses simFSD-18k to demonstrate that training set synthesis with external audio samples (yielding simFSD-18k) underperforms training set synthesis with local audio samples (yielding simCENSE-18k). See Sec. V for details.

[1]More information on YAMNet available at https://www.tensorflow.org/hub/tutorials/yamnet (Last viewed 06/08/2021).

Abeßer, J., Gotze, M., Kuhnlenz, S., Grafe, R., Kuhn, C., ClauB, T., and Lukashevich, H. (**2018**). "A distributed sensor network for monitoring noise level and noise sources in urban environments," in *Proceedings of the IEEE International Conference on Future Internet of Things and Cloud (FiCloud)*, August 6–8, Barcelona, Spain, pp. 318–324.

Andén, J., Lostanlen, V., and Mallat, S. (**2019**). "Joint time–frequency scattering," IEEE Trans. Signal Process. **67**(14), 3704–3718.

Antoni, J. (**2010**). "Orthogonal-like fractional-octave-band filters," J. Acoust. Soc. Am. **127**, 884–895.

Ardouin, J., Charpentier, L., Lagrange, M., Gontier, F., Fortin, N., Ecotière, D., Picaut, J., and Mietlicky, C. (**2018**). "An innovative low-cost sensor for urban sound monitoring," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, August 26–29, Chicago, IL, Vol. 258, pp. 2226–2237.

Aumond, P., Can, A., De Coensel, B., Botteldooren, D., Ribeiro, C., and Lavandier, C. (**2017**). "Modeling soundscape pleasantness using perceptual assessments and acoustic measurements along paths in urban context," Acta Acust. united Acust. **103**(3), 430–443.

Basner, M., Babisch, W., Davis, A., Brink, M., Clark, C., Janssen, S., and Stansfeld, S. (**2014**). "Auditory and non-auditory effects of noise on health," Lancet **383**(9925), 1325–1332.

Beery, S., Van Horn, G., and Perona, P. (**2018**). "Recognition in terra incognita," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 8–20, Munich, Germany, pp. 456–473.

Bello, J. P., Silva, C., Nov, O., DuBois, R. L., Arora, A., Salamon, J., Mydlarz, C., and Doraiswamy, H. (**2019**). "SONYC: A system for monitoring, analyzing, and mitigating urban noise pollution," Commun. ACM **62**(2), 68–77.

Bellucci, P., Peruzzi, L., and Zambon, G. (**2017**). "LIFE DYNAMAP project: The case study of Rome," Appl. Acoust. **117**, 193–206.

Berglund, B., and Nilsson, M. E. (**2006**). "On a tool for measuring soundscape quality in urban residential areas," Acta Acust. united Acust. **92**(6), 938–944.

Botteldooren, D., Dekoninck, L., Meeussen, C., and Van Renterghem, T. (**2018**). "Early stage sound planning in urban re-development: The Antwerp case study," in *Proceedings of the International Congress and Exposition on Noise Control Engineering (Inter-Noise)*, August 26–29, Chicago, IL.

Bristow, A., and Thanos, S. (**2015**). "What do hedonic studies of the costs of road traffic noise nuisance tell us?," J. Acoust. Soc. Am. **138**(3), 1750–1750.

Brocolini, L., Lavandier, C., Quoy, M., and Ribeiro, C. (**2013**). "Measurements of acoustic environments for urban soundscapes: Choice of homogeneous periods, optimization of durations, and selection of indicators," J. Acoust. Soc. Am. **134**(1), 813–821.

Bronzaft, A. L. (**2002**). "Noise pollution: A hazard to physical and mental well-being," in *Handbook of Environmental Psychology* (Wiley, New York), Chap. 32, pp. 499–510.

Brown, A., Kang, J., and Gjestland, T. (**2011**). "Towards standardization in soundscape preference assessment," Appl. Acoust. **72**(6), 387–392.

Cartwright, M., Cramer, J., Salamon, J., and Bello, J. P. (**2019a**). "TriCycle: Audio representation learning from sensor network data using self-supervision," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 20–23, New Paltz, NY, pp. 278–282.

Cartwright, M., Dove, G., Méndez Méndez, A. E., Bello, J. P., and Nov, O. (**2019b**). "Crowdsourcing multi-label audio annotation tasks with citizen scientists," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, May 4–9, Glasgow, Scotland, pp. 1–11.

Cartwright, M., Mendez, A. E. M., Cramer, J., Lostanlen, V., Dove, G., Wu, H.-H., Salamon, J., Nov, O., and Bello, J. (**2019c**). "SONYC Urban Sound Tagging (SONYC-UST): A multilabel dataset from an urban acoustic sensor network," in *Proceedings of the International Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, October 25–26, New York.

Cartwright, M., Seals, A., Salamon, J., Williams, A., Mikloska, S., MacConnell, D., Law, E., Bello, J. P., and Nov, O. (**2017**). "Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations," Proc. ACM Hum. Comput. Interact. **1**(CSCW), 1–21.

CENSE (**2019**). "Caractérisation des environnements sonores urbains," https://cense.ifsttar.fr/ (Last viewed 06/08/2021).

Cerutti, G., Prasad, R., Brutti, A., and Farella, E. (**2020**). "Compact recurrent neural networks for acoustic event detection on low-energy low-complexity platforms," IEEE J. Sel. Top. Signal Process. **14**, 654.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (**2014**). "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, October 25–29, Doha, Qatar, pp. 1724–1734.

Chung, Y.-A., and Glass, J. (**2017**). "Learning word embeddings from speech," in *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, December 4–9, Long Beach, CA.

Cohen-Hadria, A., Cartwright, M., McFee, B., and Bello, J. P. (**2019**). "Voice anonymization in urban sound recordings," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, October 13–16, Pittsburgh, PA, pp. 1–6.

Cramer, J., Lostanlen, V., Farnsworth, A., Salamon, J., and Bello, J. P. (**2020**). "Chirping up the right tree: Incorporating biological taxonomies into deep bioacoustic classifiers," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 901–905.

Cramer, J., Wu, H.-H., Salamon, J., and Bello, J. P. (**2019**). "Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 4–8, Barcelona, Spain, pp. 3852–3856.

Das, A., Borisov, N., and Caesar, M. (**2014**). "Do you hear what I hear? Fingerprinting smart devices through embedded acoustic components," in *Proceedings of the SIGSAC Conference on Computer and Communications Security (CCS)*, November 3–7, Scottsdale, AZ, pp. 441–452.

Esselink, B. (**2000**). *A Practical Guide to Localization* (John Benjamins Publishing, Amsterdam).

Fonseca, E., Favory, X., Pons, J., Font, F., and Serra, X. (**2021**). "FSD50k: An open dataset of human-labeled sound events," (published online 2020); arXiv:2010.00475, https://10.5281/zenodo.4060432.

Fonseca, E., Plakal, M., Ellis, D. P., Font, F., Favory, X., and Serra, X. (**2019**). "Learning sound event classifiers from web audio with noisy labels," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 12–17, Brighton, UK, pp. 21–25.

Font, F., Roma, G., and Serra, X. (**2013**). "Freesound technical demo," in *Proceedings of the ACM International Conference on Multimedia*, September 23, New York, pp. 411–412.

Gaidon, A., Lopez, A., and Perronnin, F. (**2018**). "The reasonable effectiveness of synthetic visual data," Int. J. Comput. Vision **126**(9), 899–901.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (**2017**). "Audio Set: An ontology and human-labeled dataset for audio events," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March 5–9, New Orleans, LA, pp. 776–780.

Giryes, R., Sapiro, G., and Bronstein, A. M. (**2016**). "Deep neural networks with random Gaussian weights: A universal classification strategy?," IEEE Trans. Signal Process. **64**(13), 3444–3457.

Gloaguen, J.-R., Can, A., Lagrange, M., and Petiot, J.-F. (**2019**). "Road traffic sound level estimation from realistic urban sound mixtures by non-negative matrix factorization," Appl. Acoust. **143**, 229–238.

Gontier, F., Lagrange, M., Aumond, P., Can, A., and Lavandier, C. (**2017**). "An efficient audio coding scheme for quantitative and qualitative large scale acoustic monitoring using the sensor grid approach," Sensors **17**(12), 2758.

Gontier, F., Lavandier, C., Aumond, P., Lagrange, M., and Petiot, J.-F. (**2019**). "Estimation of the perceived time of presence of sources in urban acoustic environments using deep learning techniques," Acta Acust. united Acust. **105**(6), 1053–1066.

Hammer, M. S., Swinburn, T. K., and Neitzel, R. L. (**2014**). "Environmental noise pollution in the United States: Developing an effective public health response," Environ. Health Perspect. **122**(2), 115–119.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (**2017**). "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv:1704.04861.

Ioffe, S., and Szegedy, C. (**2015**). "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, July 6–11, Lille, France, Vol. 37, pp. 448–456.

Kingma, D. P., and Ba, J. (**2014**). "Adam: A method for stochastic optimization," arXiv:1412.6980.

Kolesnikov, A., Zhai, X., and Beyer, L. (**2019**). "Revisiting self-supervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 15–20, Long Beach, CA, pp. 1920–1929.

Lafay, G., Lagrange, M., Rossignol, M., Benetos, E., and Roebel, A. (**2016**). "A morphological model for simulating acoustic scenes and its application to sound event detection," IEEE/ACM Trans. Audio Speech Language Process. **24**(10), 1854–1864.

Lagrange, M. (**2018**). "simScene," https://bitbucket.org/mlagrange/simscene (Last viewed 06/08/2021).

Lagrange, M. (**2021**). "gontier2021training," https://github.com/mathieulagrange/gontier2021training (Last viewed 06/08/2021).

Lagrange, M., Lafay, G., Défréville, B., and Aucouturier, J.-J. (**2015**). "The bag-of-frames approach: A not-so-sufficient model for urban soundscapes," J. Acoust. Soc. Am. **138**(5), EL487–EL492.

Lee, K., and Nam, J. (**2019**). "Learning a joint embedding space of monophonic and mixed music signals for singing voice," in *Proceedings of the International Society for Music Information Retrieval (ISMIR) Conference*, November 4–8, Delft, Netherlands.

Lostanlen, V., Salamon, J., Farnsworth, A., Kelling, S., and Bello, J. P. (**2018**). "Birdvox-full-night: A dataset and benchmark for avian flight call detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 15–20, Calgary, Canada, pp. 266–270.

Lostanlen, V., Salamon, J., Farnsworth, A., Kelling, S., and Bello, J. P. (**2019**). "Robust sound event detection in bioacoustic sensor networks," PLoS One **14**, e0214168.

McFee, B., Salamon, J., and Bello, P. (**2018**). "Adaptive pooling operators for weakly labeled sound event detection," IEEE/ACM Trans. Audio Speech Language Process. **26**(11), 2180–2193.

Méndez Méndez, A. E., Cartwright, M., and Bello, J. P. (**2019**). "Machine-crowd-expert model for increasing user engagement and annotation quality," in Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, May 4–9, Glasgow, Scotland, pp. 1–6.

Mendoza, E., Lostanlen, V., Salamon, J., Farnsworth, A., Kelling, S., and Bello, J. P. (**2019**). "BirdVox-scaper-10k: A synthetic dataset for multilabel species classification of flight calls from 10-second audio recordings (version 1.0) [data set]," Zenodo, https://doi.org/10.5281/zenodo.2560773 (Last viewed 06/08/2021).

Mesaros, A., Heittola, T., Benetos, E., Foster, P., Lagrange, M., Virtanen, T., and Plumbley, M. D. (**2018**). "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," IEEE/ACM Trans. Audio Speech Language Process. **26**(2), 379–393.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (**2013**). "Efficient estimation of word representations in vector space," in *Proceedings of the International Conference on Learning Representations (ICLR)*, May 2–4, Scottsdale, AZ.

Mydlarz, C., Shamoon, C., and Bello, J. P. (**2017**). "Noise monitoring and enforcement in New York City using a remote acoustic sensor network," in *Proceedings of INTER-NOISE and NOISE-CON Congress*, August 27–30, Hong Kong, Vol. 255, pp. 5509–5520.

Mydlarz, C., Sharma, M., Lockerman, Y., Steers, B., Silva, C., and Bello, J. P. (**2019**). "The life of a New York City noise sensor network," Sensor **19**(6), 1415.

New York City Department of Health and Mental Hygiene (**2014**). "Ambient noise disruption in New York City," Epi Data Brief 45 (New York City Department of Health and Mental Hygiene, New York).

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (**2015**). "Librispeech: An ASR corpus based on public-domain audio books," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 19–24, South Brisbane, Australia, pp. 5206–5210.

Park, T. H., Turner, J., Musick, M., Lee, J. H., Jacoby, C., Mydlarz, C., and Salamon, J. (**2014**). "Sensing urban soundscapes," in *Proceedings of the EDBT/ICDT Workshop*, March 28, 2014, Athens, Greece, pp. 375–382.

Pascual, S., Ravanelli, M., Serrà, J., Bonafonte, A., and Bengio, Y. (**2019**). "Learning problem-agnostic speech representations from multiple self-supervised tasks," in *Proceedings of the International Speech*

Communication Association Conference (INTERSPEECH), September 15–19, Graz, Austria, pp. 161–165.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. (**2016**). "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 27–30, Las Vegas, NV, pp. 2536–2544.

Picaut, J., Can, A., Fortin, N., Ardouin, J., and Lagrange, M. (**2020**). "Low-cost sensors for urban noise monitoring networks—A literature review," Sensor **20**(8), 2256.

Piczak, K. J. (**2015**). "Environmental sound classification with convolutional neural networks," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, October 19–20, Dalian, China, pp. 1–6.

Pijanowski, B. C., Farina, A., Gage, S. H., Dumyahn, S. L., and Krause, B. L. (**2011**). "What is soundscape ecology? An introduction and overview of an emerging new science," Landscape Ecol. **26**(9), 1213–1232.

Poikselkä, M., Holma, H., Hongisto, J., Kallio, J., and Toskala, A. (**2012**). *Voice over LTE: VoLTE* (Wiley, New York).

Ricciardi, P., Delaitre, P., Lavandier, C., Torchia, F., and Aumond, P. (**2015**). "Sound quality indicators for urban places in Paris cross-validated by Milan data," J. Acoust Soc. Am. **138**(4), 2337–2348.

Romanou, A. (**2018**). "The necessity of the implementation of privacy by design in sectors where data protection concerns arise," Comput. Law Security Rev. **34**(1), 99–110.

Salamon, J., and Bello, J. P. (**2017**). "Deep convolutional neural networks and data augmentation for environmental sound classification," IEEE Signal Process. Lett. **24**(3), 279–283.

Salamon, J., Jacoby, C., and Bello, J. P. (**2014**). "A dataset and taxonomy for urban sound research," in *Proceedings of the ACM International Conference on Multimedia*, November 3–7, New York, pp. 1041–1044.

Salamon, J., MacConnell, D., Cartwright, M., Li, P., and Bello, J. P. (**2017a**). "Scaper: A library for soundscape synthesis and augmentation," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 15–18, New Paltz, NY, pp. 344–348.

Salamon, J., MacConnell, D., Cartwright, M., Li, P., and Bello, J. P. (**2017b**). "URBAN-SED (version 2.0.0) [data set]," Zenodo, https://doi.org/10.5281/zenodo.1324404 (Last viewed 06/08/2021).

Sheng, Z., Pfersich, S., Eldridge, A., Zhou, J., Tian, D., and Leung, V. C. (**2019**). "Wireless acoustic sensor networks and edge computing for rapid acoustic monitoring," IEEE/CAA J. Automatica Sin. **6**(1), 64–74.

Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., and Plumbley, M. D. (**2015**). "Detection and classification of acoustic scenes and events," IEEE Trans. Multimedia **17**(10), 1733.

Su, J., Jin, Z., and Finkelstein, A. (**2020**). "Acoustic matching by embedding impulse responses," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 4–8, Barcelona, Spain, pp. 426–430.

Tagliasacchi, M., Gfeller, B., de Chaumont Quitry, F., and Roblek, D. (**2020**). "Pre-training audio representations with self-supervision," IEEE Signal Process. Lett. **27**, 600–604.

Tung, H.-Y. F., Tung, H.-W., Yumer, E., and Fragkiadaki, K. (**2017**). "Self-supervised learning of motion capture," in *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, December 4–9, Long Beach, CA.

Turchet, L., Fazekas, G., Lagrange, M., Ghadikolaei, H. S., and Fischione, C. (**2020**). "The Internet of Audio Things: State-of-the-art, vision, and challenges," IEEE Internet Things J. **7**, 10233.

Turpault, N., and Serizel, R. (**2020**). "Desed_synthetic (version v2.2)," Zenodo, http://doi.org/10.5281/zenodo.4307908 (Last viewed 06/08/2021).

United Nations (**2018**). "World Urbanization Prospects: The 2018 Revision, Methodology," Working Paper ESA/P/WP.252, Department of Economic and Social Affairs, Population Division (United Nations, New York).

Vidaña-Vila, E., Navarro, J., Borda-Fortuny, C., Stowell, D., and Alsina-Pagès, R. M. (**2020**). "Low-cost distributed acoustic sensor network for real-time urban sound monitoring," Electron **9**(12), 2119.

Virtanen, T., Plumbley, M. D., and Ellis, D. (**2018**). *Computational Analysis of Sound Scenes and Events* (Springer, New York).

Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., and Torralba, A. (**2018**). "The sound of pixels," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 8–14, Munich, Germany, pp. 570–586.

Zhu, B., Xu, K., Kong, Q., Wang, H., and Peng, Y. (**2020**). "Audio tagging by cross filtering noisy labels," IEEE/ACM Trans. Audio Speech Language Process. **28**, 2073–2083.