ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447655

ASPED: AN AUDIO DATASET FOR DETECTING PEDESTRIANS

Pavan Seshadri[b], Chaeyeon Han[a], Bon-Woo Koo[d], Noah Posner[c], Subhrajit Guhathakurta[a], Alexander Lerch[b]

[a]CSPAV, [b]Music Informatics Group, and [c]IPaT, Georgia Institute of Technology
[d]Toronto Metropolitan University

ABSTRACT

We introduce the new audio analysis task of pedestrian detection and present a new large-scale dataset for this task. While the preliminary results prove the viability of using audio approaches for pedestrian detection, they also show that this challenging task cannot be easily solved with standard approaches.

Index Terms— pedestrian detection, audio classification, dataset

1. INTRODUCTION

The intelligent analysis of urban soundscapes plays an increasingly important role in the design of smart cities. Microphones can complement or even replace other forms of sensors because (i) they are affordable, (ii) have low power requirements, (iii) can cover large angles up to 360 degrees, and (iv) are not negatively impacted by light conditions, weather patterns such as fog, or obstacles blocking the angle of view.

In this paper, we propose a new challenging task in urban sound analysis: the detection of pedestrians from audio-only signals. The detection of pedestrians helps in alleviating bottlenecks and in triggering advance warnings about potential dislocations. Understanding the temporal and spatial variation in demand for pedestrian infrastructure can also lead to better resource use, more equitable service delivery, and greater sustainability and resilience. Detecting pedestrians through audio poses some unique challenges. The audio signal pedestrians produce are often low volume and can be as diverse as steps and speech. These signals have to be detected in a poly-timbral and time-varying mixture of multiple urban sound sources with overlapping frequency content.

To allow investigation of the viability of this novel task as well as to enable and encourage future research on the task of pedestrian detection through audio signals, we present a new, large-scale dataset containing audio and video data recorded in multiple separate recording sessions at different locations at the Georgia Tech campus, Atlanta.[1] The number of pedestrians in proximity to the microphones is annotated through video analysis with three different proximity radii.

The main contributions of this paper are (i) the introduction of a new task in urban sound analysis: pedestrian detection, (ii) the publication of a new large-scale audio dataset for this task called ASPED (Audio Sensing for PEdestrian Detection), and (iii) the presentation of baseline results for benchmarking and for viability analysis.

This work was funded by NSF Award 2203408.

[1] urbanaudiosensing.github.io/ASPED, last access date Sep 6, 2023

2. RELATED WORK

Identifying the environmental context through Sound Event Detection (SED) has been an active area of research in the past decade [1, 2]. The challenge of SED in a typical outdoor environment is the detection of an event from multiple known and unknown sources of sound that are emitted simultaneously. Initial approaches to SED have used Mel-frequency cepstral coefficients (MFCC) or other time-frequency representations such as Fourier transform and the wavelet transform [3, 4]. Other approaches included non-negative matrix transformations (NMF) and spectrogram analysis with image processing techniques [5, 6, 7]. Recent advances in feedforward neural networks (FNN) and multilabel recurrent neural networks (RNN) have been particularly promising for SED [8, 9, 3].

The advances in SED have led to a small but emerging field focusing on the detection and classification of urban sounds [1, 2]. This research has been instrumental in the automatic detection of crime indicators such as screams and gunshots and in monitoring urban noise pollution [10, 11, 12, 13]. A large-scale research effort in this domain has been an NSF-funded project called SONYC for detecting noise and tagging urban sound sources [14]. This project has provided a large dataset of audio recordings tagged by citizen science volunteers who annotated the presence of 23 fine-grained categories of events. Another such dataset is AudioSet, which was developed by the Machine Perception Research Organization at Google [15]. AudioSet is a large-scale collection of human-labeled 10 s sound clips from over 2 million YouTube videos and contain 527 classes of annotated sounds. The same group at Google has also released the YouTube-100M data set labeled with one or more topic identifiers from a set of 30,871 labels [16]. These labels are assigned automatically based on the metadata and image content. A number of labeled data sets for SED have also been developed from contributions to freesound.org, including ESC-50 and FSD50K [17, 18]. In addition, VGGSound is another audio-visual dataset released in 2020 containing more than 310 audio classes [19]. However, previous research have not focused on sensing pedestrians using SED techniques. Thus, existing datasets are not annotated with pedestrian counts and cannot be easily relabeled. Furthermore, capturing meaningful audio data from pedestrians is particularly difficult because it is not clear what kind of sound can clearly identify pedestrian movement – is it footsteps, conversation, rustling of clothes, or a combination of these components? The challenge also is to detect such sound in a loud multilayered soundscape within a street environment with vehicular noise.

3. DATASET

We aimed to reach the following goals by planning and creating the dataset: (i) large scale: large amount of data to accommodate advanced data-hungry machine learning models, (ii) high quality: to

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on March 31,2025 at 20:32:06 UTC from IEEE Xplore. Restrictions apply.

Fig. 1: Research team installing audio recorders in the field.

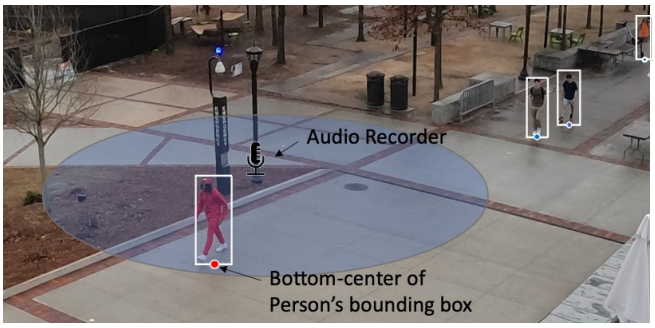

Fig. 2: Pedestrian detection video setup.

allow for investigation of the impact of audio quality levels (sample rate, word length), recordings should be provided in high audio quality, and (iii) diversity: different recording locations and times to account for a variety of scenarios.

3.1. Data acquisition

Two hardware setups were used for data acquisition. The audio collection setup consisted of multiple Tascam DR-05X audio recorders with power banks for extended duration recording, Saramonic SR-XM1 microphones, and 5L OverBoard Waterproof Dry Flat Bags for audio permeable weatherproofing. The video setup is a GoPro HERO9 Black cameras with power banks (housed in a Seahorse 56 OEM Micro Hard Case) for extended duration recording.

Multiple audio sensors and cameras were deployed for each data collection session. For each session, the recorders were placed in their weatherproof bags once started, then secured to their recording locations using zip ties. Recorders were secured at approx. chest height as it was determined that sub-meter variation in height did not affect audio quality. Figure 1 shows the installation of one audio recorder.

The cameras were set to time-lapse mode with a 1 s duration. Wi-Fi functionality was disabled to extend battery life. Multiple cameras were utilized to keep all recorders in view. The camera mounts were secured at $\approx$2.5 m using zip ties.

In order to time sync the cameras, the time as listed on www.time.gov was shown on a mobile device to each camera after starting recording. A fox 40 pearl whistle was then blown and the precise time was recorded. This whistle was used to sync the audio recorders. In deployment locations over larger areas, multiple whistle blows were conducted.

Recorders were deployed at two on-campus locations, the Cadell Courtyard, and the Tech Walkway. Both locations are near areas with restaurants and cafes but are off-limits to vehicular traffic. The battery life of the recording devices limited the length of each recording session to approx. 2 days per session.

In total, we captured 1-fps video recordings that sum up to 3,406,229 frames and the corresponding audio recordings of nearly 2,600 hours. All but one recorded days are weekdays.

3.2. Annotations

The number of pedestrians that actually passed the audio recorders was detected and annotated by applying the Masked-attention Mask Transformer (Mask2Former) [20], with a prediction threshold of 0.7 on video the recordings. This study used a Mask2Former implementation by OpenMMLab,[2] trained on Microsoft COCO [21].

	Pedestrian Counts				
Radius	**0**	**1**	**2**	**3**	**4+**
1	99.4757	0.4723	0.0453	0.0058	0.0008
3	97.2538	2.1906	0.4075	0.0962	0.0519
6	91.9113	5.2566	1.7578	0.6262	0.4481
9	86.6895	7.7196	3.0175	1.2999	1.2734

Table 1: Percentage proportion of each pedestrian count per the labels for each recorder radius

For each video frame, bounding boxes of the detected 'person' class were first extracted from the prediction from Mask2Former. Next, circular buffers of different radius $r \in [1\,\text{m}, 3\,\text{m}, 6\,\text{m}, 9\,\text{m}]$ were overlaid on the video frames around the poles to which audio recorders were attached. The buffers were angled to match the perspective of each video recording instance. Finally, the number of pedestrians with the bottom center of the bounding box intersecting with recorder buffers was counted and labeled in each frame. Figure 2 visualizes an example buffer and pedestrians with bounding boxes.

Each frame has four sets of annotation data for the four different recording radii (see Sect. 3.2). Among the annotated videos, frames without any detected pedestrians were the most common. Frames with one pedestrian were next most frequent, followed by those with two, then three, four pedestrians, and so on. Proportions of each count within the labels for each radius are displayed in Table 1. The labeled data contains more pedestrians detected during the daytime as shown in Fig. 3. Pedestrian activities were at peak around noon, especially during lunch time (11AM–14PM).

4. EXPERIMENTS

4.1. Experimental setup

We determine a baseline level of performance for this task with three different models, all targeting a binary classification (pedestrians present/not present) at different microphone radius settings and for different pedestrian count thresholds separating the present/not present classes. We aim to investigate the effects of these permutations to establish an understanding of how a basic classifier responds to the change of data parameters of this task.

[2]openmmlab.com, last access date Sep 5, 2023

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on March 31,2025 at 20:32:06 UTC from IEEE Xplore. Restrictions apply.

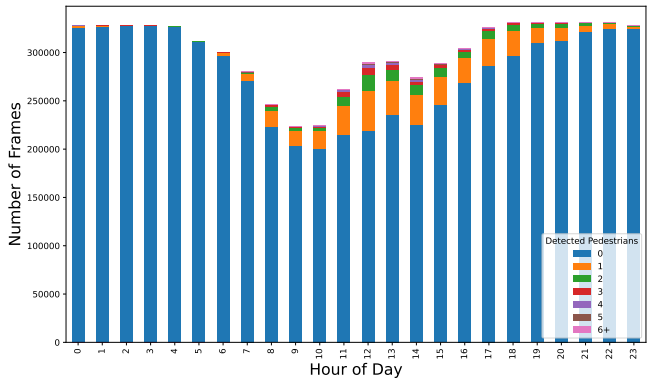

Fig. 3: Number of pedestrians radius $r = 6\,\mathrm{m}$ by hour of day.

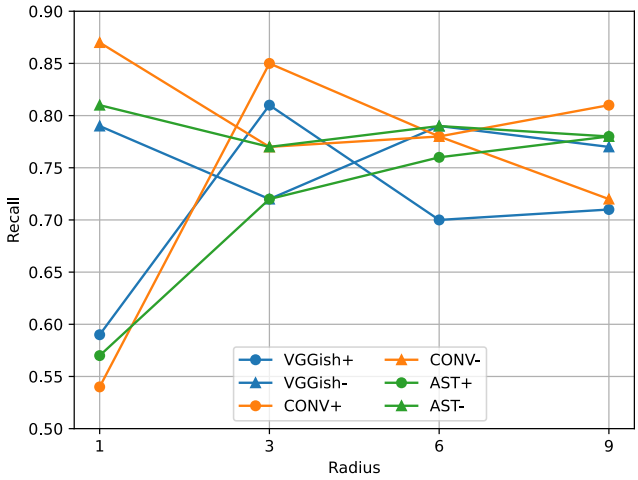

Fig. 4: Recall for each class over recording radius. Positive and negative classes are denoted by "+" and "-", respectively.

4.1.1. *Model architectures*

First, we investigate using the VGGish embeddings [16], pre-trained on AudioSet [15], as input to a transformer encoder to learn temporal relationships across each segment (referred to as *VGGISH*). Second, we use a convolutional encoder with a log-mel spectrogram input, followed by the aforementioned transformer encoder (referred to as *CONV*). Third, we explore using the Audio Spectrogram Transformer, which has been shown to deliver state-of-the-art performance for audio scene classification tasks [22] (referred to as *AST*). All models compute class output probabilities through an appended linear classification layer with a sigmoid activation function.

4.1.2. *Feature extraction*

All network inputs are extracted in time frames of approx. 1 s length. Both VGGISH and CONV follow the pre-processing procedure for the pre-trained VGGish network [16], resulting in a 128-dimensional VGGish embedding or a 96×64 dimensional (time $\times$ freq) log-mel spectrogram, respectively. The AST input is a spectrogram with dimensionality 100×128 (time $\times$ freq), following the original publication [22].

The input of the VGGish and CONV models are a sequence of 10 concurrent features, corresponding to each 1 s frame per 10 s audio segment. The input to the AST are single features per 1 s frame. Each classification is done per frame, for every second of audio.

4.1.3. *Training procedure*

As our data contains pedestrian counts per frame, we create classification labels where values of 0 are counted as negative-activity, and any value above 0 is counted as positive-activity.

The dataset was randomly split into train/test/validation subsets with 80/10/10 proportion, respectively. For testing and validation, any overlapping segments are removed so that labels are not re-used multiple times.

The loss function for all models is binary cross-entropy. As Fig. 3 shows, the label distribution is highly skewed towards no-activity; To promote the learning of pedestrian activity, we use the following augmentations for the underrepresented classes: (i) *weighted batch sampling* — in each mini batch, audio segments are sampled with replacement such that roughly half will contain at least one pedestrian activity event; (ii) *variable weighted loss* — each classes loss is weighted dynamically per batch relative to its density in the training samples, such that both positive and negative pedestrian-activity contribute roughly equally to the loss per batch. The weighting function used is shown below:

$$\mathcal{L} = \lambda \mathcal{L}_{\mathrm{BCE}+} + (1-\lambda)\mathcal{L}_{\mathrm{BCE}-} \tag{1}$$

$$\lambda = \begin{cases} \frac{1/num^+}{1/num^+ + 1/num^-}, & \text{if } num^+ \neq 0 \\ 0, & \text{if } num^+ = 0 \end{cases} \tag{2}$$

4.1.4. *Hyperparameters and implementation*

For CONV and VGGISH, we use 1 transformer encoder with 4 attention heads, with a hidden dimensionality of 128. CONV contains 6 convolutional blocks each containing a conv2D, batchnorm, and leakyReLU layer. Both networks are trained with a learning rate of 0.0005. For the AST, we use the base configuration per the authors implementation[3] pre-trained on ImageNet [23] and AudioSet [15] with hidden dimensionality of 768, and a learning rate of 5e-7. We train the CONV and VGGISH models for 20 epochs and the AST for 10 epochs, with the best performing model selected via performance on the validation set. Parameters are optimized using the ADAM optimizer [24]. We use a batch size of 256 for VGGISH and CONV, and 32 for AST.

4.1.5. *Experiments*

We evaluate the baseline performance measured by class-level and macro-average recall with the following experiments:

E1 — Comparison of baseline architectures: In order to capture task performance using general audio classification methods as well as to evaluate performance across architectures of varying complexity, the three models introduced above are compared. The complexity ranges from $\sim$ 100K trainable parameters (VGGISH) to $\sim$ 80M trainable parameters (AST).

E2 — Impact of recording radius on accuracy: With this experiment, the impact of the recording radius on the performance is investigated. Spatial consideration for determining pedestrian activity affects both the count and diversity of pedestrian noises: smaller radii

[3] github.com/YuanGongND/ast, last access date Sep 5, 2023

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on March 31,2025 at 20:32:06 UTC from IEEE Xplore. Restrictions apply.

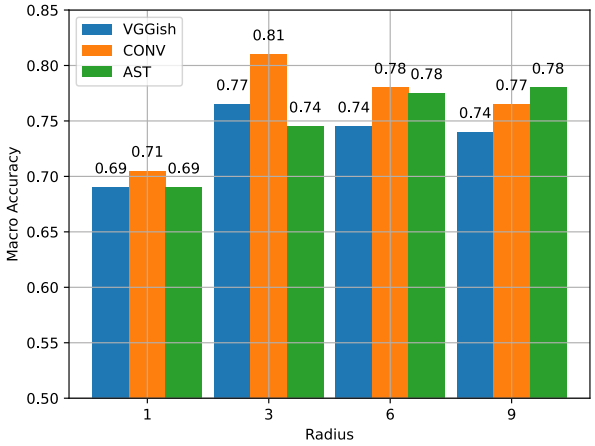

Fig. 5: Macro average accuracy using the VGGISH, CONV, and AST models.

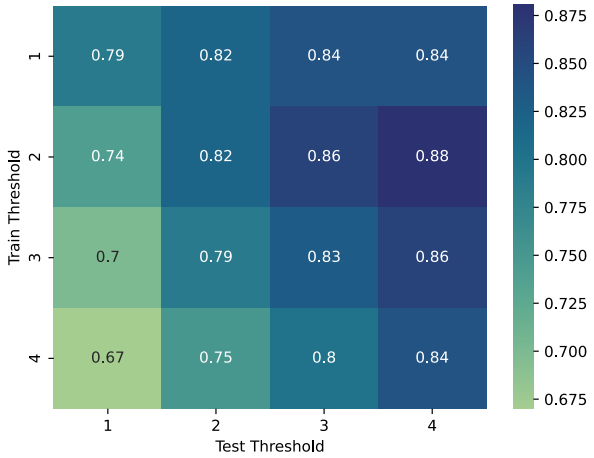

Fig. 6: Macro average accuracy over each train and test pedestrian count threshold for radius $r = 6\,\mathrm{m}$.

contain a lower number of pedestrians that should be easier to classify while larger radii contain a higher number of pedestrians with harder to classify samples. As such, larger radii should provide a greater diversity of pedestrian signals to our models with the downside that counted pedestrians are more difficult to detect.

E3 — Impact of pedestrian count during training and testing on performance: As the threshold for binary classification can be set at arbitrary pedestrian counts, it does not necessarily have to be identical for both training and testing. Therefore, we determine the impact of different training thresholds on different inference thresholds and thus investigate the model generalizability to different pedestrian activity. This experiment utilizes a CONV model at radius $r = 6\,\mathrm{m}$ while thresholding labels with values $p_\mathrm{T} \in [1,2,3,4]$ such that any value lower than p_T is set to 0. We then test each trained model on the 4 resulting test sets.

4.2. Results

E1: Figures 4 and 5 detail the results using the VGGISH, CONV, and AST models, respectively. We can make the following observations. First, the VGGISH model is in most cases outperformed by both the CONV and AST models, which is expected as the pre-trained VGGish embeddings are not fine-tuned for pedestrian detection. Second, in terms of macro accuracy, the performance of the VGGISH and AST model are fairly constant across radii 3 to $9\,\mathrm{m}$, while the CONV model achieves highest performance on radii 3 and $6\,\mathrm{m}$. The reasons for these radii working better could be a combination of less imbalanced class distribution and reasonable proximity to the microphone. Third, the AST generally has closest parity between performance on both classes. Lastly, the negative class recall seems to slightly outperform the recall for the positive (pedestrian) activity, although the dramatic class imbalance observed in the data is not reflected in the results showing the effectiveness of the sampling and loss weighting applied during training.

E2: When attempting to compare the performance across different radii in Fig. 4, it is important to note that the test sets are not identical; although all audio content is identical, the labels and, therefore, class-proportions differ. The performance per class tends to be most balanced using radii 3 and $6\,\mathrm{m}$. The performance for radius $1\,\mathrm{m}$ likely suffers due to pedestrian signals just outside the radius being labeled as no-activity, while radii 6 and $9\,\mathrm{m}$ see a slight decline from the opposite issue: low-volume pedestrian signals on the edge are labeled as pedestrians while potentially not detectable from audio.

E3: Figure 6 visualizes the macro accuracy for each permutation of combinations of train threshold and test threshold for pedestrian count. We can make the following observations. First, as the threshold for the *test* pedestrian count increases, a greater proportion of the samples are classified correctly. It is unsurprising that the classifier can perform better as an increased pedestrian count likely correlates to stronger signals and thus more easily detected frames. Second, performance generally decreases with increasing threshold for the *training* pedestrian count, indicating that the classifier benefits from harder to classify training samples. In general, the performance seems to be best when trained with a low pedestrian count threshold and evaluated with a high pedestrian count threshold (upper right triangle).

5. CONCLUSION

We have introduced the new large-scale dataset ASPED for the challenging task of detecting pedestrians from audio data. The dataset includes high quality audio recordings plus the video recordings used for labeling the data with pedestrians counts. The baseline results indicate the feasibility of using audio sensors for pedestrian tracking, although the performance needs to be improved before systems become practically usable.

Plans for future work include extending the dataset to additional environment such as locations with car traffic. Future directions of inquiry will include (i) robustness against noise, (ii) impact of the recording quality (sample rate etc.), (iii) impact of weather such as wind on rain on the performance, (iv) accuracy of regression approaches to predict exact pedestrian counts, and (v) the performance of more sophisticated classification approaches for pedestrian detection. Accurate detection of pedestrians can offer valuable information and guidance for transportation infrastructure planning and urban planning in general. Knowing why people choose to walk in certain urban areas and the destinations they prefer can help improve pedestrian infrastructure, reduce overcrowding, and provide information about untapped opportunities for economic development. Future plans also include using the ASPED dataset for pedestrian route prediction using network flow models.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on March 31,2025 at 20:32:06 UTC from IEEE Xplore. Restrictions apply.

6. REFERENCES

[1] E. Babaee, N. B. Anuar, A. W. Abdul Wahab, S. Shamshirband, and A. T. Chronopoulos, "An Overview of Audio Event Detection Methods from Feature Extraction to Classification," *Applied Artificial Intelligence*, vol. 31, no. 9-10, pp. 661–714, 2017.

[2] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound Event Detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 2021.

[3] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent Neural Networks for Polyphonic Sound Event Detection in Real Life Recordings," in *ICASSP*, 2016.

[4] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *EUSIPCO*, 2010.

[5] S. Innami and H. Kasai, "NMF-based environmental sound source separation using time-variant gain features," *Computers & Mathematics with Applications*, vol. 64, no. 5, pp. 1333–1342, 2012.

[6] A. Dessein, A. Cont, and G. Lemaitre, "Real-time detection of overlapping sound events with non-negative matrix factorization," *Matrix information geometry*, pp. 341–371, 2013.

[7] J. Dennis, H. D. Tran, and E. S. Chng, "Overlapping sound event recognition using local spectrogram features and the generalised hough transform," *Pattern Recognition Letters*, vol. 34, no. 9, pp. 1085–1093, 2013.

[8] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection," *IEEE/ACM TASLP*, vol. 25, no. 6, pp. 1291–1303, 2017.

[9] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *IJCNN*, 2015, pp. 1–7.

[10] L. Gerosa, G. Valenzise, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection in noisy environments," in *EUSIPCO*, 2007.

[11] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *IEEE AVSS*, 2007.

[12] S. S. Hammad, D. Iskandaryan, and S. Trilles, "An unsupervised TinyML approach applied to the detection of urban noise anomalies under the smart cities environment," *Internet of Things*, vol. 23, pp. 100848, 2023.

[13] J. Socoró, F. Alías, and R. Alsina-Pagès, "An Anomalous Noise Events Detector for Dynamic Road Traffic Noise Mapping in Real-Life Urban and Suburban Environments," *Sensors*, vol. 17, no. 10, pp. 2323, 2017.

[14] P.M. Cartwright, A. E. Mendez, J. Cramer, V. Lostanlen, G. Dove, H.-H. Wu, J. Salamon, O. Nov, and J. P. Bello, "SONYC Urban Sound Tagging (SONYC-UST): A Multilabel Dataset from an Urban Acoustic Sensor Network," in *DCASE*, 2019.

[15] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017.

[16] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN Architectures for Large-Scale Audio Classification," in *ICASSP*, 2017.

[17] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *ACMMM*, 2015.

[18] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An Open Dataset of Human-Labeled Sound Events," *IEEE/ACM TASLP*, vol. 30, pp. 829–852, 2022.

[19] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "VGGSound: A Large-scale Audio-Visual Dataset," 2020, arXiv:2004.14368.

[20] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *CVPR*, 2022.

[21] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[22] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Interspeech*, 2021.

[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on March 31,2025 at 20:32:06 UTC from IEEE Xplore. Restrictions apply.