Leveraging Acoustic Images for Effective Self-Supervised Audio Representation Learning

Supplementary material

Valentina Sanguineti[1,2], Pietro Morerio[1], Niccolò Pozzetti[4], Danilo Greco[1,2], Marco Cristani[4], and Vittorio Murino[1,3,4]

[1] Pattern Analysis & Computer Vision, Istituto Italiano di Tecnologia
[2] University of Genova, Italy [3]Huawei Technologies Ltd., Ireland Research Center
{valentina.sanguineti,pietro.morerio,danilo.greco,vittorio.murino}@iit.it
[4] University of Verona, Italy
niccolo.pozzetti@studenti.univr.it marco.cristani@univr.it

The supplementary material is organized as follows. In Section 1, we provide more details, some statistics and examples about the ACIVW dataset. Section 2 presents an analysis of the classification performance using different modalities. In Section 3, we report implementation details, in particular regarding the setting of the several hyperparameters used. Finally, in Section 4, we show some cross-modal retrieval examples.

1 ACIVW: ACoustic Images and Videos in the Wild Dataset

As described in the main paper, we used a $0.45m \times 0.45m$ planar array of 128 MEMS microphones located according to an optimized aperiodic layout with a webcam (so that it doesn't have any appreciable geometric distortion) in the device center for dataset collection shown in Figure 1.

Fig. 1. DualCam acoustic-optic camera.

The device is capable of acquiring audio data in the useful bandwidth 500 Hz – 6.4 kHz and audio-video sequences at a frame rate of 12 frames per second (fps).

In particular, the data provided by the sensor consists in RGB video frames of 480×640 pixels and raw audio signals from 128 microphones acquired with a sampling frequency of 12.8 kHz. Thus the device can record all frequencies from 0 up to 6.4 kHz (the Nyquist frequency limit), however it is less directive below 500 Hz. Fourier harmonics make our device still sensitive to sound outside this range. The single-mic audio signal is upsampled only for the audio model, to allow a fair comparison with [5]. [5] upsampled the audio signal for HearNet [3] to allow a comparison with SoundNet [2], which employs 22050 samples/sec.

$36 \times 48 \times 512$ multispectral acoustic images are obtained from the raw audio signals of all the microphones combining them through the beamforming algorithm [7], which summarizes the audio intensity for every direction and discretized frequency bin. We adopted MFCC compression, so that not only experiments were less computational demanding and memory hungry, but also resulted in a better accuracy. MFCC have been proven to be good in audio compression while maintaining the characteristic sound properties, and 12 coefficients are often considered in literature. We also employ 12 MFCC for single-mic audio, which showed to be sufficient to classify audio signals. The acquisition of the latter modality is aligned not only in time with optical images, but also in space: each acoustic pixel corresponds to 13.3×13.3 RGB pixels. So the acoustic image has a lower resolution than the RGB image and we can see that when we overlap acoustic image energy to video frames since acoustic pixels are interpolated in the videos for visualization purposes. This is due to the size limitations of the planar array, which limits the directivity of the beampattern.

Acoustic and RGB images are geometrically aligned by a calibration procedure that selects the correct virtual field of view of the array of microphones in the beamforming algorithm. More detailed info can be found in ref. [4] (Sect. 2).

We acquired a big dataset outdoors in the wild containing 5 hours of three modalities aligned in time and space: RGB frames, audio and acoustic images. Planar arrays are very sensible to echos that are usually present in indoor environments, so we collected the whole dataset outdoors to record good quality sounds exploiting the planar array features.

Number of classes	10
Number of videos per class	60/26.80/10
Total number of videos	268
Length of videos in seconds	256/68.95 /2
Length of class in seconds	1898/1847.8/1794
Total length of videos in seconds	18478

Table 1. ACIVW dataset statistics. Where there are three entries in a field, numbers refer to the maximum/average/minimum.

Statistics about dataset are in Table 1. We show examples of the three modalities for each class in Figures 2, 3: on the left RGB image, in the center energy of the corresponding acoustic image overlaid on RGB frame and on the right

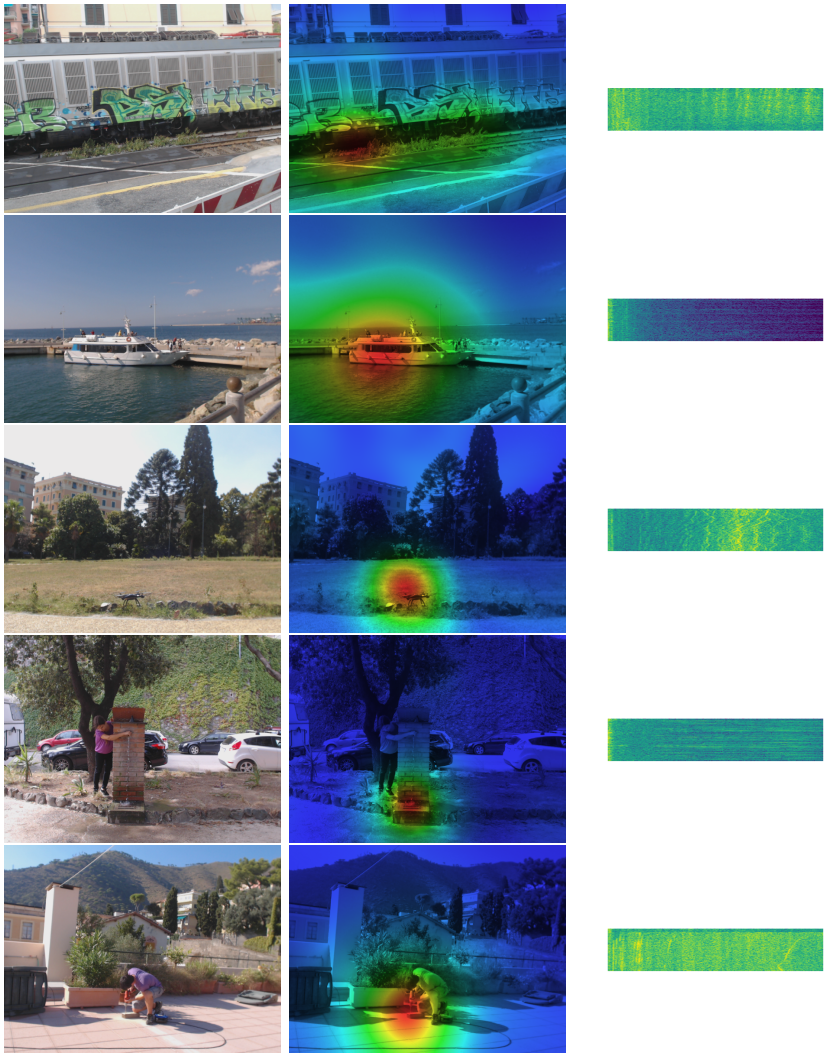

Fig. 2. Examples of ACIVW Dataset from the classes: from top to bottom *train*, *boat*, *drone*, *fountain*, *drill*. Left: RGB frame, center: acoustic energy map overlaid on the acoustic frame, right: single microphone spectrogram.

the spectrogram obtained from one single microphone. Spectrogram examples for each class give a qualitative idea of the class frequency content.

We accompany this pdf with some videos with acoustic image energy overlaid on video frames. Each video comes in two versions:

1. With the audio from a single omnidirectional microphone, which is fixed for all videos

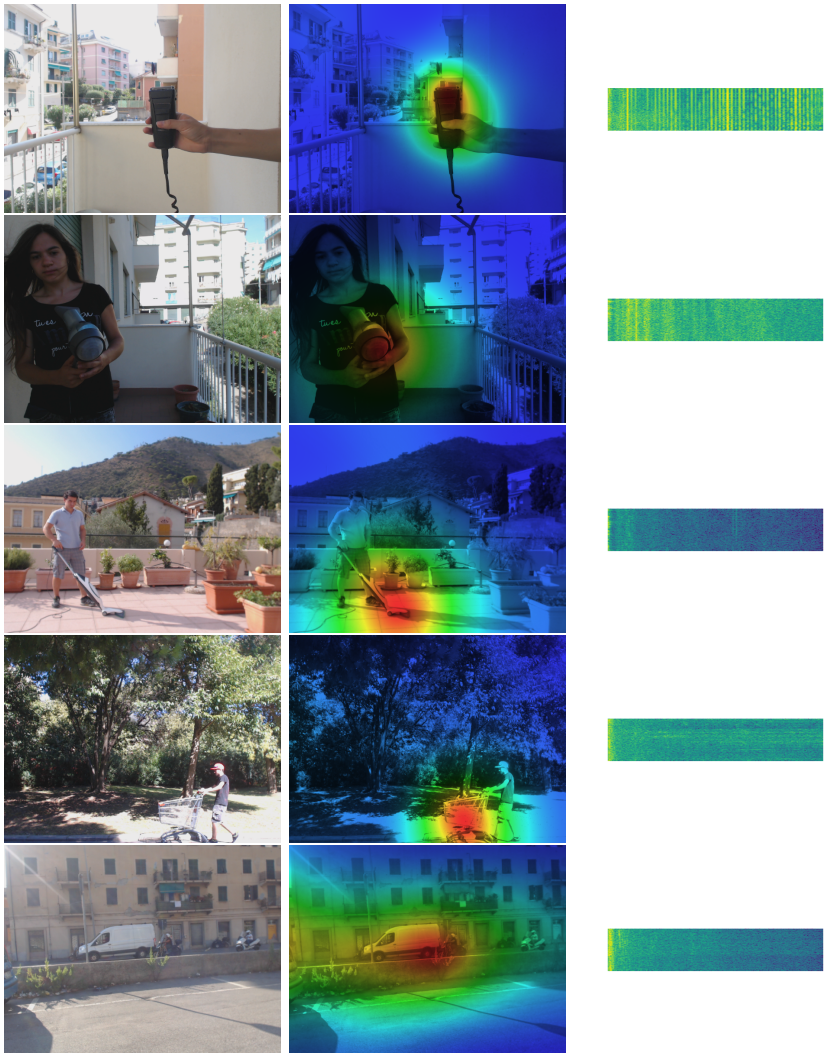

Fig. 3. Examples of classes of the ACIVW Dataset. From the top: *razor*, *hair dryer*, *vacuum cleaner*, *shopping cart*, *traffic*. Left: RGB frame, center: acoustic energy map overlaid on the acoustic frame, right: single microphone spectrogram.

2. With the audio coming from the the direction of the sound source. This is obtained by isolating the virtual directional microphone corresponding to the acoustic pixel where the source is located.

The latter is obtained through Inverse Fast Fourier Transform (IFFT) of the FFT of the acoustic pixel obtained with the beamforming algorithm. You can notice the difference between omnidirectional and directional sound.

Both the dataset and the code will be released at this link https://github.com/IIT-PAVIS/acoustic-images-self-supervision.

2 ACIVW Performance Analysis

We show the confusion matrices on one run for the three different supervised models, namely DualCamNet, HearNet and ResNet18, in Figures 4, 5, 6, respectively. Both DualCamNet and HearNet classification is good, as we can see from diagonal confusion matrices. DualCamNet only confuses drill with razor and vice versa. Concerning ResNet18, we notice that it makes a lot of confusion when classifying the classes drone, fountain, drill and mostly hair dryer. This is because we collected our dataset in real scenarios where not always the items are easy to detect because of occlusions, deceiving details in the scene and the tiny object size (for example a drone). In some cases, instead, the object is visually difficult to classify as its appearance changes a lot from one video to another one, for instance the fountain, instead hair dryer and drill sometimes are alone in the scene, sometimes used by a person.

As demonstrated in Table 2 (top box) in the main paper, in fact, we can actually classify very well both spectrograms and acoustic images, while video classification is more challenging.

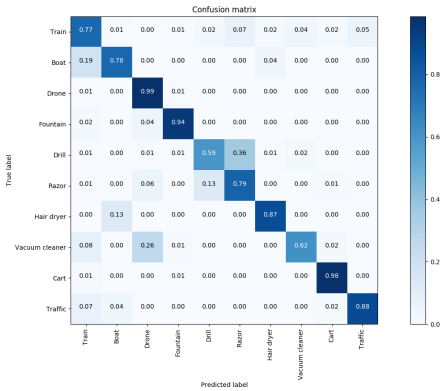

Fig. 4. Supervised DualCamNet confusion matrix.

3 Implementation Details

3.1 Data Preparation

We implemented all of our networks and our data processing pipeline using TensorFlow. In particular we stored our dataset in multiple compressed TFRecord

Fig. 5. Supervised HearNet confusion matrix.

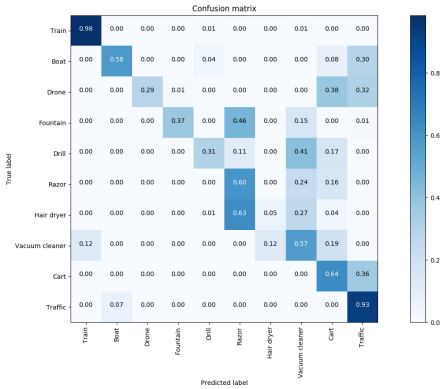

Fig. 6. Supervised ResNet18 confusion matrix.

files, each of which contains 1 second of synchronized data from the three modalities, video images, raw audio waveforms, and acoustic images. We use the `tf.data` API to retrieve this data and compose at runtime 2.0 s sequences grouping contiguous TFRecord files into full audio-video sequences.

3.2 Hyper-parameters

We cross-validated learning rate, number of epochs and margin. We chose the biggest batch size that could fit GPU available memory. Then, we employed the following hyper-parameters throughout all the experiments: learning rate 10^{-5}, margin $m = 0.2$, 20 epochs, batch size 64 for self-supervised case; learning rate 10^{-4}, 100 epochs, batch size 32 for the supervised case and for [5], where we also chose $\alpha = 0.5$ and $T = 1$, as indicated therein. We compare our results with [1], trained with learning rate 10^{-3}, 100 epochs, batch size 16 and we also trained

the audio and video sub-networks separately for 100 epochs with batch size 32 and learning rate 10^{-5} and 10^{-6}, respectively. Results are averaged over 5 runs.

3.3 k-NN

Regarding the k-NN classification accuracy results, we cross-validated the considered number of nearest neighbors k considering odd numbers between 7 and 15.

3.4 Triplet Loss Margin

We cross-validated margin m choosing it among $\{0.2, 0.5, 1.0, 1.5\}$. $m = 0.2$ was our best performing option in case of no distillation and it is usually chosen as default value [6] for the triplet loss. We kept it fixed to $m = 0.2$ for the second setup as well, i.e. when distilling from DualCamNet.

3.5 Knowledge distillation

To perform distillation, we consider a pre-trained acoustic image network. This model was trained in advance in a self-supervised manner together with video network using correspondence pretext task. We restored the teacher model corresponding to the epoch where the acoustic image network had best classification results on the validation set. This was done separately for each of the five runs of distillation.

4 Cross-modal Retrieval

We show some retrieval examples in Figures 7 and 8. The first sample on the left marked by a note is the audio embedding, the other images on the right from the second column on, are the corresponding retrieved RGB images with increasing distance from $k = 1$ to $k = 5$. In green, there are samples belonging to the same class, in blue belonging to same video, in red to different classes. Given an audio embedding (of a single microphone audio *or* acoustic image), we retrieve the corresponding video frame by matching the closest audio-visual embedding. We cannot do the opposite because the audio-visual embedding is a function of the audio and cannot be computed without its information.

Results for each class are shown in 2 rows: we plot the first 5 retrieved images for one audio embedding by considering, in the first row, acoustic image and, in the second row, single microphone audio.

We are also able to retrieve RGB frames from different clips of the same class, not only samples which belong to the same video.

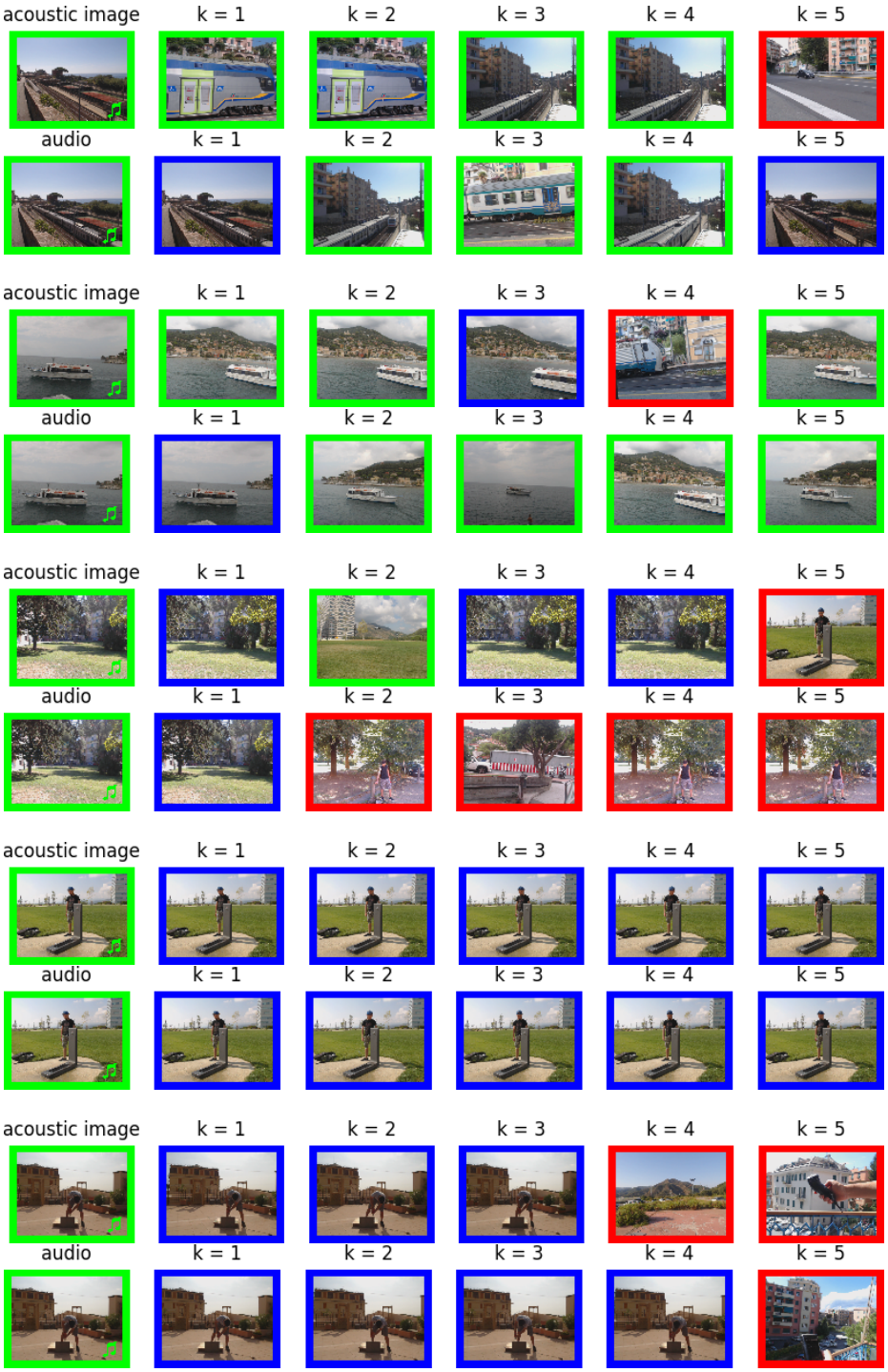

Fig. 7. Examples of ACIVW Dataset retrieved samples from the following classes. From top to bottom, two rows per class: *train*, *boat*, *drone*, *fountain*, *drill*.

Fig. 8. Examples of ACIVW Dataset retrieved samples from the following classes. From top to bottom, two rows per class: *razor*, *hair dryer*, *vacuum cleaner*, *shopping cart*, *traffic*.

References

1. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
2. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. pp. 892–900. NIPS'16, Curran Associates Inc., USA (2016), http://dl.acm.org/citation.cfm?id=3157096.3157196
3. Aytar, Y., Vondrick, C., Torralba, A.: See, hear, and read: Deep aligned representations. CoRR **abs/1706.00932** (2017), http://arxiv.org/abs/1706.00932
4. Crocco, M., Martelli, S., Trucco, A., Zunino, A., Murino, V.: Audio tracking in noisy environments by acoustic map and spectral signature. IEEE Transactions on Cybernetics **48**, 1619–1632 (May 2018)
5. Pérez, A.F., Sanguineti, V., Morerio, P., Murino, V.: Audio-visual model distillation using acoustic images. In: Winter Conference on Applications of Computer Vision (WACV) (2020)
6. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 815–823 (June 2015)
7. Van Trees, H.: Detection, Estimation, and Modulation Theory, Optimum Array Processing. Wiley (2002)