Leveraging Acoustic Images for Effective Self-Supervised Audio Representation Learning

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

Abstract. In this paper, we propose the use of a new modality characterized by a richer information content, namely acoustic images, for the sake of audio-visual scene understanding. Each pixel in such images is characterized by a spectral signature, associated to a specific direction in space and obtained by processing the audio signals coming from an array of microphones. By coupling such array with a video camera, we obtain spatio-temporal alignment of acoustic images and video frames. This constitutes a powerful source of self-supervision, which can be exploited in the learning pipeline we are proposing, without resorting to expensive data annotations. However, since 2D planar arrays are cumbersome and not as widespread as ordinary microphones, we propose that the richer information content of acoustic images can be distilled, through a self-supervised learning scheme, into more powerful audio and visual feature representations. The learnt feature representations can then be employed for downstream tasks such as classification and cross-modal retrieval, without the need of a microphone array. To prove that, we introduce a novel multimodal dataset consisting in RGB videos, raw audio signals and acoustic images, aligned in space and synchronized in time. Experimental results demonstrate the validity of our hypothesis and the effectiveness of the proposed pipeline, also when tested for tasks and datasets different from those used for training.

Keywords: Audio-visual representations, acoustic images, audio- and video-based classification, cross-modal retrieval, self-supervised learning

1 Introduction

Humans perceive and interpret the world by combining different sensory modalities. However, designing computational systems able to emulate or surpass human capabilities in this respect, although of utmost importance from both scientific and applicative standpoints, is still a far-reaching goal.

Among all modalities, vision and audio are surely the most commonly used and important ones that both humans and machines can use to sense the world.

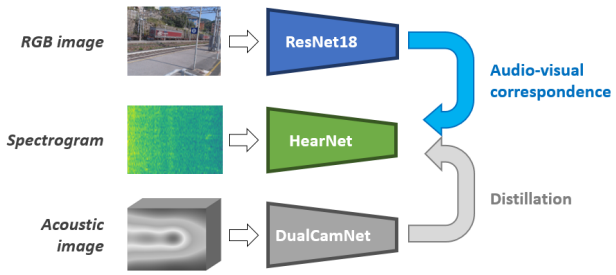

Fig. 1. We consider three modalities aligned in time and space: RGB, (monaural) audio signal (here in the form of spectrogram), and acoustic images. We exploit such correspondence to jointly learn audio-visual representations. We improve audio models with knowledge transfer from the acoustic image model.

This is also caused by the fact that they are often quite correlated, temporally synchronized, and support each other for interpretation tasks. In fact, sound helps to pay attention and visually focus on situations of interest, and may complement noisy or low-quality visual information, ultimately aiming to improve the interpretation of a scene. In such cases, humans take advantage of the spatial localization of the produced sound (obtained thanks to the binaural configuration of our auditory system), to shift visual attention to the event that generated the sound.

Unfortunately, artificial systems mimicking human performance are not so common, especially because video data typically comes with a monaural (single microphone) acoustic signal only. Hence, spatial localization is lost, and reliably recovering it is a difficult and only partially solved problem [9,23]. Thus, in order to have the possibility to emulate human performance by also exploiting spatially localized audio data, one needs to resort to an array of microphones positioned in special geometrical (e.g., planar) configuration, and able to provide an enriched audio description of a scene – an acoustic image – being formed by properly combining the signals acquired by all microphones. In an acoustic image, each pixel is characterized by the spectral signature corresponding to the audio signal coming from the corresponding direction, so, overall, allowing effectively to visualize the acoustic landscape of the sensed scene (see Fig. 2).

In particular, we take advantage of an audio-visual sensor composed by a microphone array coupled with a video camera, jointly calibrated, in order to get a sequence of acoustic images and associated video frames, aligned in space and time [36,6]. Examples of sample video frames overlaid with the energy map of the sound obtained from the corresponding acoustic images are shown in Fig. 2. The peculiar nature of this data, i.e., the spatial alignment and time synchronization of the data produced by such sensor, opens the door to the adoption of *self-supervised learning* approaches for model training. The motivation for this choice lies in the fact that such methods do not require data annotations. This

specifically suits to our case, since acoustic images would results quite expensive to fully annotate (i.e., assigning pixel-level or bounding box annotations to the same objects in both video frames and acoustic images while listening the signals coming from different directions). Instead, self-supervised methods just exploit the implicit supervision inherent in the signals themselves. For example, we can train audio-visual networks by simply looking and listening to a large number of unlabelled videos, and exploiting their natural alignment as a supervisory signal. More in detail, in deep self-supervised learning schemes, a network is trained to solve a so-called *pretext task*, and the quality of learned features is then assessed on a variety of *downstream tasks*, which are usually supervised (e.g., classification), showing beneficial effects [18].

More specifically, in this paper we investigate whether we can obtain more powerful features for downstream tasks by training audio-visual models with a self-supervised framework exploiting audio-visual correspondence. We also employ acoustic image modality as privileged information [34] used at training time in a knowledge distillation [21] framework (see Fig. 1) to enhance such audio-visual self-supervised features. The distillation framework was already exploited in the literature for classification tasks in several scenarios [15,21,11,10,28], but always in *supervised* settings. Instead, here we are proposing a novel *self-supervised distillation* framework, which does not require any time-consuming annotations, and allows to train audio and video models together. To the best of our knowledge, privileged information was never exploited before in a self-supervised learning pipeline. After training, individual models can be used as feature extractors for the sake of audio and video classification and cross-modal retrieval as downstream tasks.

To show the potentiality of acoustic images to improve feature learning, we collected a new multimodal audio-visual dataset, composed by RGB video frames, acoustic images and monaural audio signals[3]. This dataset contains 10 classes of real sound acquired outdoors in the wild, is bigger than AVIA dataset [28] and more suitable for self-supervised learning. With this novel dataset we carry out an accurate ablation study; subsequently, in two different benchmark datasets publicly available [28,22], we show that, when augmented with privileged information distilled from acoustic images, the obtained feature representations are more powerful than in the case of just training audio and visual models with the audio-visual correspondence task. In the end, acoustic images proved to have notable characteristics to be effectively transferred to other domains and tasks, when distilled by our training mechanism.

In summary, the main contributions of this work can be summarized as follows:

- We propose a multimodal deep learning framework to learn audio-visual models considering a novel modality, acoustic images, which is heavily underexplored in computer vision. This framework embeds a novel self-supervised distillation mechanism to transfer the information extracted from an acoustic

[3] https://github.com/IIT-PAVIS/acoustic-images-self-supervision

image model to an audio model for learning more powerful feature representations.
- We collect and release a new multimodal dataset of aligned audio (single microphone), RGB images and acoustic images, bigger than [28].
- Using this dataset for model training, we show the effectiveness of our framework for downstream tasks such as 1) audio and video classification, and 2) cross-modal retrieval. In particular, we prove that the features obtained by the distillation of acoustic images perform better than those obtained without using such privileged information, not only on our dataset, but also on other public benchmarks [28,22].

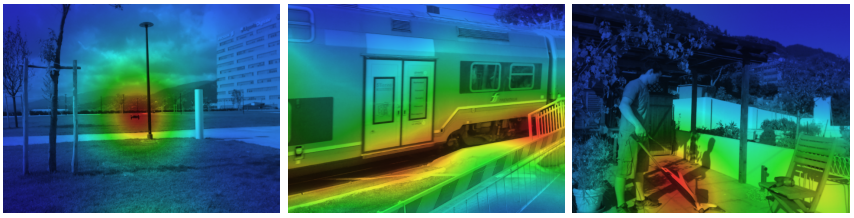

Fig. 2. Three examples from the collected dataset. We visualize the acoustic image by summing the energy of all frequencies for each acoustic pixel. The resulting map is overlaid on the corresponding RGB frame. From left to right: drone, train, and vacuum cleaner classes.

The rest of the paper is organized as follows. We review the state of the art and highlight the main differences with respect to our work in Section 2. Section 3 introduces our new dataset and briefly presents the sensor used. Section 4 explains the proposed self-supervised training method and, in Section 5, we evaluate our learning strategy and report the performance of the experiments in the downstream tasks. Finally, in Section 6, conclusions are drawn.

2 Related works

Our work lies at the intersection of two broad topics, namely self-supervised learning and knowledge distillation. In this section we give an overview of relevant works in both fields, mainly in the context of audio-visual learning, and discuss how our method relates to them. We also review literature dealing with acoustic images.

Audio-visual self-supervised learning. Multimodal learning takes advantage of data from different modalities [24] aiming at obtaining better semantic representations than those learned by segregated modalities.

There has been increased interest in using perception-inspired audio-visual (fusion) deep learning models because the correspondence between the visual and the audio streams is ubiquitous and free in unlabeled consumer videos.

Vision and sound are often informative about the same concept of the world. As a consequence of their correlations, concurrent visual and sound data provide a rich supervisory self-training signal that can be used to jointly learn useful audio and video representations. Early approaches trained single networks on one modality only, using the other one to derive a sort of supervisory signal [3,13,26,27]. For example, [3,13] train an audio network using pre-trained visual architectures as teachers. Instead, [26,27] directly predicts sound from video, thus using ambient sound as a supervisory signal for video.

Other works [19,1,17,2,12,31,29,8,25] jointly train visual and audio streams, aiming at learning multimodal representations useful for many applications, such as classification, cross-modal retrieval, sound source localization, speech separation, and on/off-screen audio-visual source separation. As in [1], we also use audio-visual correspondence verification task: networks are trained to determine whether a video frame and a short audio chunk overlap in time. Learned representations are then tested in a classification task. Within a similar framework, [19] uses hard samples, i.e., slightly out-of-sync audio and visual segments sampled from the same video in a self-supervised curriculum-based learning scheme. [2] enforces the alignment of features extracted by audio and visual networks by computing the correspondence score as a function of the Euclidean distance between the normalized visual and audio embeddings, hence making them amenable to retrieval.

The common factor in all these works is the natural *temporal* synchronization between (single) auditory signal and visual images, which is used to train the several models in a self-supervised manner. Some works instead explore the *spatial alignment* between stereo auditory signal and visual images [35].

In our case, the intrinsic *temporal synchronization*, but also the *spatial alignment* of visual and acoustic images are exploited as a supervisory signal. Our method takes inspiration from [2] and [31]: we force audio-visual agreement between feature maps to find aligned shared representations, however, both the task and the mechanism we propose for training are different, since they involve knowledge distillation and an extremely different modality.

Knowledge distillation. Our work is related to knowledge distillation, which can be coarsely and generally defined as the class of approaches trying to indeed condensate knowledge gained in a learning task and feed another learning task or another model [15]. Such framework was later unified with the privileged information framework [34] into the so-called generalized distillation theory [21], and recently exploited in the context of multi-modal learning with missing modalities at test time [16,11,10]

The seminal work [3] capitalizes on the natural synchronization between vision and sound to train a sound classification model using a teacher-student setup, transferring from video teachers (ImageNet and Places pre-trained networks) into sound. However, such teachers are trained with supervision, while we do not use any supervision at all. In fact, while traditional generalized distillation framework are applied in a *supervised* setup, since exploiting cross-entropy loss and teacher's soft predictions [21], we are here in the self-supervised scenario,

where labels are missing. We can thus only leverage embeddings as additional information from the teacher. Furthermore, the teacher network itself is also trained with self-supervision.

Acoustic images. Acoustic images are obtained using an array of microphones, typically distributed in a planar configuration, by properly combining the audio signals acquired by every microphone using an algorithm called beamforming [33]. To the best of our knowledge, acoustic image processing with deep learning methods was only preliminary explored in [28], which proposed an architecture able to classify acoustic images in a multimodal action dataset in a supervised way. Furthermore, it also showed how to distill acoustic image information to audio models, still in a *supervised* way. The substantial difference of our work with respect to [28] is that we use here a self-supervised learning approach, for which, as also above highlighted, the canonical supervised distillation [21] cannot be applied. Other applications of acoustic images regarded only the tracking of sound sources [6,36]. In the end, no other works are present in the literature aimed at using such unique source of information in a *self-supervised learning* setting.

3 ACIVW: ACoustic Images and Videos in the Wild

We acquired a multimodal dataset containing 5 hours of videos outdoors in the wild, using an acoustic-optical camera. The sensor captures both raw audio signals from 128 microphones acquired with a sampling frequency of 12.8 kHz and RGB video frames of 480×640 pixels, using a planar array of microphones located according to an optimized aperiodic layout [7] and a webcam placed at the device center. Audio data is acquired in the useful bandwidth 500 Hz – 6.4 kHz and audio-video sequences are acquired at a frame rate of 12 frames per second (fps). $36 \times 48 \times 512$ multispectral acoustic images are obtained from the raw audio signals of all the microphones combining them through the beamforming algorithm [33], which summarizes the audio intensity for every direction and discretized frequency bin. 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 visual pixels. Among the raw audio waveforms, we choose one microphone for training monaural audio networks.

We selected 10 classes of interest: drone, shopping cart, traffic, train, boat, fountain, drill, razor, hair dryer, vacuum cleaner. Figure 2 shows three sample RGB frames overlaid with the energy of the corresponding acoustic image. More examples, videos and details are provided in the supplementary material. We acquired data for half an hour for each class, in different locations and viewpoints. This implies more than 21,000 RGB and acoustic images for each class. The data is split in training, validation and test in the proportion 70%, 15% and 15%.

We use the training split of this dataset for the pretext task of learning correspondences. We then test for the downstream tasks of cross-modal retrieval and classification on the test set. For classification, we also test on two publicly available datasets proposed in [28] and [22].

4 The method

As mentioned above, we consider three data modalities, namely audio, acoustic images and RGB images, and we adopt a different stream network for each modality as they are extremely heterogeneous, as shown in Fig. 3.

Our aim is to train two models at a time using audio-visual correspondence pretext task: first, we train the acoustic images' stream jointly with the RGB stream, and, second, the audio stream with the RGB stream. After that, we exploit the trained acoustic image stream to distill additional knowledge to a new audio stream, trained again using the same pretext task as illustrated in Figure 4. We then compare the performances of audio and video models trained with and without the aid of the self-supervised pre-trained acoustic image stream.

4.1 Input data

For the three modalities we consider temporal windows of 2.0 s, which represent a good compromise between information content and computational load.

Monaural Audio. The audio amplitude spectrogram is obtained from an audio waveform of 2 seconds, upsampled to 22 kHz by computing the Short-Time Fourier Transform (STFT), considering a window length of 20 ms with half-window overlap. This produces 200 windows with 257 frequency bands. The resulting spectrogram is interpreted as a $200 \times 1 \times 257$ dimensional signal, so that the frequency bands can be interpreted as the number of channels in convolutions, as detailed in Figure 3.

Acoustic images. Acoustic images are generated with the frequency implementation of the filter-and-sum beamforming algorithm [36]. They are volumes of size $36 \times 48 \times 512$ (36×48 as image size, 512 frequency channels). These channels correspond to the frequency bins discretizing frequency content for each pixels. A more comprehensive description of acoustic images generation can be found in [33]. However, handling acoustic images with 512 channels is computationally expensive and most of the useful information is typically contained in the low frequencies. Consequently, we compressed the acoustic images along the frequency axis using Mel-Frequency Cepstral Coefficients (MFCC), which consider human audio perception characteristics [32]. We thus compute 12 MFCC, compressing from 512 to 12 channels, preserving most of the information but consistently reducing the computational complexity and the required memory. Acoustic images frame rate is $12\,\mathrm{s}^{-1}$, so we consider 24 acoustic images in input.

RGB video. RGB frames are $224 \times 298 \times 3$ volumes obtained by scaling the original $360 \times 480 \times 3$ video frames, keeping the original proportion. The images are then normalized by subtracting ImageNet mean [20]. Even if we have both acoustic images and RGB images frame rates are $12\,\mathrm{s}^{-1}$, we consider just one RGB frame per second to reduce computational burden, so we have 2 RGB frames in input.

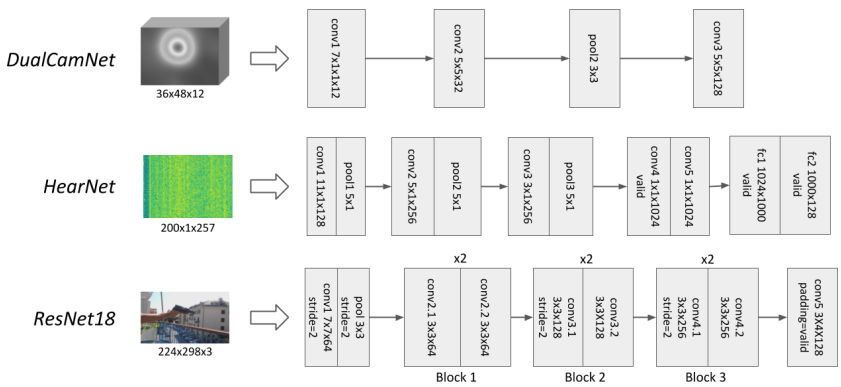

Fig. 3. The adopted models for the 3 data modalities. In convolutional layers stride=1 and padding=SAME unless otherwise specified.

4.2 Single data stream models

The chosen architectures for self-supervised learning are depicted in Figure 3: ResNet18 for RGB frames [14], DualCamNet for acoustic images [28], and HearNet [4] for the single audio signal. All networks were slightly modified for our purposes as illustrated in Figure 3. We consider ResNet18 since it can be trained from scratch on our dataset at relatively low computational cost and without the risk of overfitting. In fact, we do not want to rely on ImageNet pre-trained models to avoid employing labels at all. HearNet draws inspiration from [4] and [28], but it has been modified to consider our sampling time interval of 2.0 s (instead of 5.0 s). Such network takes spectrogram as input and has a limited size, again making it suitable to be trained from scratch.

We cut ResNet18 and DualCamNet in order to obtain feature volumes and then compute a similarity score map between them, via point-wise multiplication. In particular, we modified the original ResNet18 removing the 4-th block and the final average pooling, adding instead a 2D convolution at the end of the network. The feature volumes keep the same spatial relationships of the original image and the acoustic image. In fact, these maps are proportional to the 224×298 RGB image and to the 36×48 acoustic image.

From HearNet, we cannot get spatial feature maps as explained above in Subsection 4.1, since the signal is one-dimensional, but only a 128D array after 2 fully connected layers. In order to obtain the same similarity map as above, we propose to tile audio feature vectors along 2D spatial dimensions, to match the dimensions of the video feature map. This allows then to multiply the two maps in a point-wise fashion, as in the case of the acoustic image.

In order to obtain baselines, ResNet18 and DualCamNet can be trained in a supervised way (Table 2 - top) by adding a simple average pooling layer followed by a fully connected layer. HearNet also requires to add one fully connected layer for supervised training, in order to match the number of classes.

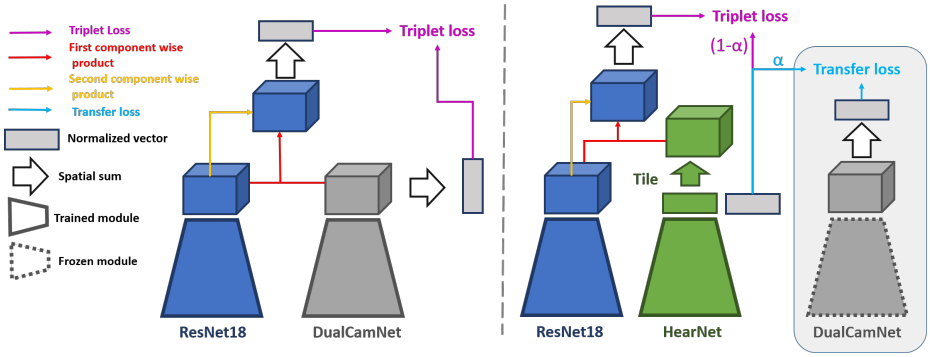

Fig. 4. Proposed distillation method. Left: self-supervised learning of the teacher DualCamNet. Right: The pre-trained teacher network contributes to the self-supervised learning of the monaural and video networks. Note that setting $\alpha = 0$ the audio network is trained without distillation.

4.3 Pretext task

We propose the self-supervised training procedures depicted in Fig. 4: we employ 2 trainable streams, from which 2 feature map volumes are extracted. We obtain a similarity score map by multiplying element-wise each $12 \times 16 \times 128$ acoustic image and video feature maps. For monaural audio, the feature cuboid is obtained by replicating the 128D vector by HearNet in each spatial location. This is different from both [31,2], as they use a dot product in each spatial location to obtain a scalar map. Keeping the original depth in the similarity score map allows to retain more information about the input.

The output of the architecture is one audio-visual feature vector and one audio vector, either obtained from DualCamNet or from HearNet. The former is a 128D vector obtained by a second element-wise product between the similarity score map and the video feature map itself, followed by a sum along the two spatial dimensions (spatial sum in Figure 4). This corresponds to a weighted sum, where the weights come from the similarity map. Instead, the 128D audio feature vector is obtained by a sum along two spatial dimensions of the acoustic feature map in case of DualCamNet, while no sum is needed in case of Hearnet, since it outputs a 128D already.

The two feature vectors are normalized and feed a triplet loss [30]:

$$\mathcal{L}_{XY}^{triplet}(x_i^a, y_i^p, y_i^n) = \sum_{i=1}^{N} [||f_X(x_i^a) - f_Y(y_i^p)||_2^2 - ||f_X(x_i^a) - f_Y(y_i^n)||_2^2 + m]_+ \quad (1)$$

where $[f(x)]_+$ represents $\max(f(x), 0)$, m is a margin between positive and negative pairs, and $f(x_i)$ are the normalized feature vectors. The triplet loss aims to separate the positive pairs from the negative ones by a distance margin. This is obtained by minimizing the distance between an anchor a and a positive, p, both of which have the same identity, and maximizing the distance between

the anchor a and a negative n of a different identity. In our case, we want an audio embedding $f_X(x_i^a)$ to have a small squared distance from video embeddings $f_Y(y_i^p)$ from the same clip, and a large one from video embeddings $f_Y(y_i^n)$ obtained from a different clip.

It is crucial to select carefully triplets to make the networks learn. In particular, we exploit a curriculum learning scheme [5]: in the first epochs, we use all the triplets that contribute to the training (i.e. with a loss greater than zero), and later on only the hardest triplets: for each anchor, we select the hardest negative as the one with smallest distance, and the hardest positive as the one with largest distance from the anchor.

4.4 Knowledge distillation

Distillation is carried out by exploiting a self-supervised pre-trained DualCamNet, as depicted in the right part of Fig. 4. To this end, we exploit an additional triplet loss between the single-audio and the acoustic-image embeddings vectors, which we name Transfer Loss $\mathcal{L}^{transfer} = \mathcal{L}_{HD}^{triplet}$, where H stands for HearNet, D for DualCamNet and $\mathcal{L}^{triplet}$ is the triplet loss. Such loss tries to transfer effective embeddings learned with DualCamNet to the monaural audio model.

The total loss is thus the weighted sum of the triplet loss between HearNet and ResNet18 embeddings and the transfer loss, which is calculated between (previously trained) DualCamNet and HearNet embeddings vectors:

$$\mathcal{L}^{tot} = \alpha\mathcal{L}^{transfer} + (1-\alpha)\mathcal{L}_{HR}^{triplet} \tag{2}$$

where $0 \leq \alpha \leq 1$, H is HearNet, R is ResNet18.

The imitation parameter α measures how much HearNet features will resemble DualCamNet features. Note that in the limit case where $\alpha = 0$ we fall in the standard self-supervised case with no knowledge transfer. We consider different values of the imitation parameter α to assess how much we have to weight the two losses.

5 Experiments

Classification and cross-modal retrieval are the downstream tasks used to evaluate the quality and generalization capability of the features learned with the proposed approach.

We compare our self-supervised method with supervised distilled audio model [28] and with L^3Net [1]. Features considered for our work and [28] are 128D.

The correspondence accuracy of L^3Net on our dataset is 0.8386 $\pm$ 0.0035. We consider both self-supervised audio and video sub networks obtained from L^3Net to extract features as well as audio and video supervised models trained separately adding after the final 512D feature vectors a fully connected layer with size equal to number of classes. Both supervised and self-supervised models of L^3Net [1] have features 512D. Training details are in the supplementary material.

5.1 Cross-modal retrieval

The target of cross-modal retrieval consists in choosing one audio sample and searching for the corresponding video frames of the same class. The audio sample comes either in the form of an acoustic image or of a spectrogram. We will specify whenever needed. Given an audio sample, corresponding audio-visual embeddings are ranked based on their distance in the common feature space. Rank K retrieval performance measures if at least one sample of the correct class falls in the top K neighbours.

Fixing a query audio we can compute audio-visual embeddings for any given video sample, while we cannot fix the audio-visual embedding, because its value will be different for different audios. Thus, we perform cross-modal retrieval only from audio to images and not vice versa.

Results are presented in Table 1 and refer to the test set of the ACIVW dataset. They clearly show that audio-visual representations learned with acoustic images (DualCamNet) are consistently better than those learned with monaural audio alone. Besides, results are good in absolute terms, considering that random chance on Rank 1 is 10% and that features learned with the pretext task proposed by [1] are less effective.

Table 1. CMC scores on ACIVW Dataset for $k = 1, 2, 5, 10, 30$.

Model	Rank 1	Rank 2	Rank 5	Rank 10	Rank 30
DualCamNet	33.41±3.65	37.01±3.17	42.97±2.25	48.21±1.80	62.44±1.33
HearNet	28.95±2.15	34.40±3.27	42.43±4.88	48.08±5.77	61.43±4.94
L^3 Audio Subnetwork [1]	9.74±0.33	11.91±4.09	24.23±9.02	26.78±8.60	30.14±10.00

5.2 Classification

For this task we use the trained models as feature extractors and classify the extracted features with a K-Nearest Neighbor (KNN) classifier. We consider both audio and audio-visual features computed as explained in Subsection 4.3. We benchmark against the proposed ACIVW dataset as reference and then test the generalization of features on two additional datasets: Audio-Visually Indicated Action Dataset (AVIA) [28] and Detection and Classification of Acoustic Scenes (DCASE - version 2018) [22]. AVIA is a multimodal dataset which provides synchronized data belonging to 3 modalities: acoustic images, audio and RGB frames. DCASE 2018 is a renowned audio dataset, containing recordings from six large European cities, in different locations for each scene class.

ACIVW Dataset. In Table 2 we report, in the top part, the fully supervised classification baseline accuracies for the single stream architectures described in Subsection 4.2. The bottom part lists instead the KNN classification accuracies for the models trained with the proposed self-supervised framework.

Table 2. Accuracy results for models on ACIVW dataset. Results are averaged over 5 runs. (H): HearNet model, (D): DualCamNet model.

Features	**Training**	**Test accuracy**
L^3 Audio Subnetwork [1]	supervised	0.6424 ± 0.2857
HearNet	supervised	0.8779 ± 0.0145
HearNet w/ transfer [28]	supervised	0.8578 ± 0.0198
L^3 Vision Subnetwork [1]	supervised	0.4647 ± 0.0225
ResNet18	supervised	0.5123 ± 0.0521
DualCamNet	supervised	0.8378 ± 0.0187
L^3 Audio Subnetwork [1]	self-supervised	0.3605 ± 0.0265
HearNet	self-supervised w/o transfer	0.7573 ± 0.0278
HearNet	self-supervised w/ transfer $\alpha = 0.1$	0.7697 ± 0.0147
HearNet	self-supervised w/ transfer $\alpha = 0.3$	0.7896 ± 0.0092
HearNet	self-supervised w/ transfer $\alpha = 0.5$	**0.7946** ± 0.0137
HearNet	self-supervised w/ transfer $\alpha = 0.7$	0.7810 ± 0.0206
HearNet	self-supervised w/ transfer $\alpha = 0.9$	0.7867 ± 0.0093
L^3 Video Subnetwork [1]	self-supervised	0.5444 ± 0.0839
Audio-visual (H)	self-supervised w/o transfer	0.6670 ± 0.0446
Audio-visual (H)	self-supervised w/ transfer $\alpha = 0.1$	0.7061 ± 0.0496
Audio-visual (H)	self-supervised w/ transfer $\alpha = 0.3$	0.7144 ± 0.0223
Audio-visual (H)	self-supervised w/ transfer $\alpha = 0.5$	0.7125 ± 0.0200
Audio-visual (H)	self-supervised w/ transfer $\alpha = 0.7$	0.7191 ± 0.0285
Audio-visual (H)	self-supervised w/ transfer $\alpha = 0.9$	**0.7322** ± 0.0070
Audio-visual (D)	self-supervised	0.5837 ± 0.0468
DualCamNet	self-supervised	0.7457 ± 0.0292

For supervised models we choose the model with best validation accuracy and provide its test performance. For self-supervised models we fix a number of iterations (20 epochs). Averages and standard deviations are computed over 5 independent runs. Results show that the videos in our dataset are quite challenging to classify. Audio models perform instead much better than video ones.

When training in a self supervised manner, audio models naturally experience a drop in performance. Such drop is partially recovered when training with the additional supervision of DualCamNet features. Hearnet w/ transfer for $\alpha = 0.5$ is indeed boosted by $\sim 4\%$.

Audio-visual features, although obtained with self-supervision, are better than visual features obtained with supervision using ResNet18. This is due to the fact that audio information can help to better discriminate the class. Also in this case the transfer is beneficial, increasing performance by $\sim 6\%$ for $\alpha = 0.9$. This is true also for self-supervised video subnetwork [1], which performs better than supervised one. This shows that when one modality is difficult to classify, self-supervision is able to improve accuracy.

Different values of the imitation parameter $\alpha \in \{0,1; 0,3; 0,5; 0,7; 0,9\}$ are investigated. We notice that both audio and audio-visual accuracies are always improved by the transferring, for all values of α.

Table 3. Accuracy results for models trained on ACIVW dataset and tested on AVIA. Results are averaged over 5 runs. (H): HearNet model, (D): DualCamNet model.

Features	Training	Test accuracy
L^3 Audio Subnetwork [1]	supervised	0.3713 ± 0.0233
HearNet	supervised	0.3108 ± 0.0114
HearNet w/ transfer [28]	supervised	0.3556 ± 0.0181
L^3 Vision Subnetwork [1]	supervised	0.0287 ± 0.0013
ResNet18	supervised	0.0263 ± 0.0073
DualCamNet	supervised	0.4783 ± 0.0224
L^3 Audio Subnetwork [1]	self-supervised	0.0571 ± 0.0175
HearNet	self-supervised w/o transfer	0.4103 ± 0.0248
HearNet	self-supervised w/ transfer $\alpha = 0.1$	0.4393 ± 0.0097
HearNet	self-supervised w/ transfer $\alpha = 0.3$	0.4749 ± 0.0305
HearNet	self-supervised w/ transfer $\alpha = 0.5$	**0.4817** ± 0.0165
HearNet	self-supervised w/ transfer $\alpha = 0.7$	**0.4851** ± 0.0214
HearNet	self-supervised w/ transfer $\alpha = 0.9$	0.4592 ± 0.0271
L^3 Vision Subnetwork [1]	self-supervised	0.3347 ± 0.0638
Audio-visual (H)	self-supervised w/o transfer	0.2660 ± 0.0309
Audio-visual (H)	self-supervised w/ transfer $\alpha = 0.1$	0.2759 ± 0.0163
Audio-visual (H)	self-supervised w/ transfer $\alpha = 0.3$	**0.3200** ± 0.0204
Audio-visual (H)	self-supervised w/ transfer $\alpha = 0.5$	0.3070 ± 0.0294
Audio-visual (H)	self-supervised w/ transfer $\alpha = 0.7$	0.3091 ± 0.0351
Audio-visual (H)	self-supervised w/ transfer $\alpha = 0.9$	**0.3162** ± 0.0310
Audio-visual (D)	self-supervised	0.2927 ± 0.0234
DualCamNet	self-supervised	**0.5132** ± 0.0167

In detail, our models perform better than both supervised and self-supevised audio and video models of L^3net subnets [1]. Our supervised audio network HearNet does not have an improvement using distillation [28] maybe because our dataset is much more challenging than the AVIA Dataset presented in [28]. In fact, ACIVW data presents many different scenarios with different noise types and as stated by [28], the acoustic images distillation works well in cases with almost no noise.

AVIA Dataset. Features learned on ACIVW are also tested on a public multimodal dataset containing acoustic images, namely Audio-Visually Indicated Action Dataset (AVIA) [28].

We compare the result of the audio and audio-visual features extracted using this dataset in Table 3. We have a general drop in accuracy because we are testing on a different dataset, however, in this case DualCamNet has the best results, proving better generalization performance than monaural features. The improvement by the the self-supervision w/ transfer is again confirmed. Different values of the imitation parameter $\alpha \in \{0,1; 0,3; 0,5; 0,7; 0,9\}$ are investigated. In particular, we notice that $\alpha = 0.5$ for audio features and $\alpha = 0.9$ for audio-visual features are still good values of α. Self-supervised models generalize better than the supervised trained ones apart from Audio subnetwork [1]. In particular,

Table 4. Accuracy for audio models tested on DCASE 2018.

Features	**Supervision**		**Training Dataset**	**Test accuracy**
Mesaros *et al.* [22]	supervised		DCASE 2018	0.5970 ± 0.0070
L^3 Audio Subnetwork [1]	supervised		ACVIW	0.3576 ± 0.0127
HearNet w/ transfer [28]				0.2989 ± 0.0106
HearNet				0.3022 ± 0.0088
L^3 Audio Subnetwork [1]	self-supervised		ACVIW	0.3231 ± 0.0473
HearNet				0.3535 ± 0.0188
HearNet	self-supervised (w/ transfer)	$\alpha = 0.1$	ACVIW	0.3653 ± 0.0079
		$\alpha = 0.3$		**0.3757** ± 0.0094
		$\alpha = 0.5$		0.3737 ± 0.0068
		$\alpha = 0.7$		0.3696 ± 0.0098
		$\alpha = 0.9$		0.3638 ± 0.0072

HearNet self-supervised is more general than the one trained with distillation [28].

DCASE 2018. In Table 4 we report classification accuracies (KNN) for DCASE 2018 using it for testing the generalization and transfer capabilities of the learned features. In other words, in our setup, DCASE was used for testing only. Specifically, we used the test set (development dataset) for the acoustic scene classification task for device A [22]. Classification is carried out by running KNN on both supervised and self-supervised features extracted from models pre-trained on ACIVW Dataset with supervised and self-supervised training.

We do not use DCASE training data for learning any model. For this reason, the reported accuracies are below that in [22], which is reported just for reference.

Self-supervised learned representations provide a better accuracy than supervised models, showing that learning from concurrence of two modalities can lead to better generalization than learning from labels and with supervised distillation [28]. Transferring is useful to obtain more general features and the best result is that of $\alpha = 0.3$. For [1] this does not happen. However, even if the result of supervised case is better than the self-supervised submodule, it has a lower accuracy than our audio models self-supervised with acoustic image transfer.

6 Conclusions

In this paper, we have investigated the potential of acoustic images in a novel self-supervised learning framework and with the aid of a new multimodal dataset, specifically acquired for this purpose. Evaluating the trained models on classification and cross-modal retrieval downstream tasks, we have shown that acoustic images are a powerful source of self-supervision and their information can be distilled into monaural audio and audio-visual representation to make them more robust and versatile. Moreover, features learned with the proposed method can generalize better to other datasets than representations learned in a supervised setting.

References

1. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
2. Arandjelovic, R., Zisserman, A.: Objects that sound. In: The European Conference on Computer Vision (ECCV) (September 2018)
3. 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
4. Aytar, Y., Vondrick, C., Torralba, A.: See, hear, and read: Deep aligned representations. CoRR **abs/1706.00932** (2017), http://arxiv.org/abs/1706.00932
5. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: ICML (2009)
6. 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)
7. Crocco, M., Trucco, A.: Design of superdirective planar arrays with sparse aperiodic layouts for processing broadband signals via 3-d beamforming. IEEE/ACM Transactions on Audio, Speech, and Language Processing **22**(4), 800–815 (April 2014). https://doi.org/10.1109/TASLP.2014.2304635
8. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W., Rubinstein, M.: Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. ACM Transactions on Graphics **37** (04 2018)
9. Gao, R., Grauman, K.: 2.5d visual sound. CVPR 2019 arXiv:1812.04204 (2019)
10. Garcia, N.C., Morerio, P., Murino, V.: Learning with privileged information via adversarial discriminative modality distillation. CoRR **abs/1810.08437** (2018)
11. Garcia, N.C., Morerio, P., Murino, V.: Modality distillation with multiple stream networks for action recognition. In: The European Conference on Computer Vision (ECCV) (September 2018)
12. Harwath, D., Recasens, A., Suris, D., Chuang, G., Torralba, A., Glass, J.: Jointly discovering visual objects and spoken words from raw sensory input. In: The European Conference on Computer Vision (ECCV) (September 2018)
13. Harwath, D., Torralba, A., Glass, J.: Unsupervised learning of spoken language with visual context. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29, pp. 1858–1866. Curran Associates, Inc. (2016), http://papers.nips.cc/paper/6186-unsupervised-learning-of-spoken-language-with-visual-context.pdf
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (June 2016)
15. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. NIPS 2014 Deep Learning Workshop **abs/1503.02531** (2015)
16. Hoffman, J., Gupta, S., Darrell, T.: Learning with side information through modality hallucination. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 826–834 (June 2016). https://doi.org/10.1109/CVPR.2016.96
17. Hu, D., Nie, F., Li, X.: Deep multimodal clustering for unsupervised audiovisual learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)

18. Jing, L., Tian, Y.: Self-supervised visual feature learning with deep neural networks: A survey. CoRR **abs/1902.06162** (2019)
19. Korbar, B., Tran, D., Torresani, L.: Cooperative learning of audio and video models from self-supervised synchronization. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31, pp. 7774–7785. Curran Associates, Inc. (2018), http://papers.nips.cc/paper/8002-cooperative-learning-of-audio-and-video-models-from-self-supervised-synchronization.pdf
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems (2012)
21. Lopez-Paz, D., Bottou, L., Schölkopf, B., Vapnik, V.: Unifying distillation and privileged information. ICLR 2016 **abs/1511.03643** (2016)
22. Mesaros, A., Heittola, T., Virtanen, T.: A multi-device dataset for urban acoustic scene classification. DCASE 2018 Workshop (2018)
23. Morgado, P., Vasconcelos, N., Langlois, T., Wang, O.: Self-supervised generation of spatial audio for 360 video. In: Proceedings of the 32Nd International Conference on Neural Information Processing Systems. pp. 360–370. NIPS'18, Curran Associates Inc., USA (2018)
24. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th International Conference on International Conference on Machine Learning. pp. 689–696. ICML'11, Omnipress, USA (2011), http://dl.acm.org/citation.cfm?id=3104482.3104569
25. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: The European Conference on Computer Vision (ECCV) (September 2018)
26. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2405–2413 (2016)
27. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Learning sight from sound: Ambient sound provides supervision for visual learning. International Journal of Computer Vision **126**(10), 1120–1137 (Oct 2018). https://doi.org/10.1007/s11263-018-1083-5, https://doi.org/10.1007/s11263-018-1083-5
28. 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)
29. Ramaswamy, J., Das, S.: See the sound, hear the pixels. 2020 IEEE Winter Conference on Applications of Computer Vision (WACV) pp. 2959–2968 (2020)
30. 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)
31. Senocak, A., Oh, T.H., Kim, J., Yang, M.H., So Kweon, I.: Learning to localize sound source in visual scenes. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
32. Terasawa, H., Slaney, M., Berger, J.: A statistical model of timbre perception. In: SAPA@INTERSPEECH (2006)
33. Van Trees, H.: Detection, Estimation, and Modulation Theory, Optimum Array Processing. Wiley (2002)

34. Vapnik, V., Vashist, A.: A new learning paradigm: Learning using privileged information. Neural networks **22**(5-6), 544–557 (2009)
35. Yang, K., Russell, B., Salamon, J.: Telling left from right: Learning spatial correspondence of sight and sound. CVPR (2020)
36. Zunino, A., Crocco, M., Martelli, S., Trucco, A., Bue, A.D., Murino, V.: Seeing the sound: A new multimodal imaging device for computer vision. In: 2015 IEEE International Conference on Computer Vision Workshop (ICCVW). pp. 693–701 (Dec 2015). https://doi.org/10.1109/ICCVW.2015.95