University of Genova
Department of Electrical, Electronics and Telecommunication Engineering
and Naval Architecture (DITEN)
Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia

PhD course: Sciences and Technologies for Electronics
and Telecommunication Engineering
Cycle XXXIV (2018-2021)
Curriculum: Computer Vision, Pattern Recognition
and Machine Learning

Audio-Visual Learning for Scene Understanding

PhD Thesis submitted for the degree of *Doctor of Philosophy*

***PhD Candidate*: Valentina Sanguineti**

Vittorio Murino, Alessio del Bue — Supervisors
Pietro Morerio — Co-Supervisor
Maurizio Valle — Coordinator of the PhD Course

To my family, who always supported me.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified.

Valentina Sanguineti
February 2022

Abstract

Multimodal deep learning aims at combining the complementary information of different modalities. Among all modalities, audio and video are the predominant ones that humans use to explore the world. In this thesis, we decided to focus our study on audio-visual deep learning to mimic with our networks how humans perceive the world. Our research includes images, audio signals and acoustic images. The latter provide spatial audio information and are obtained from a planar array of microphones combining their raw audios with the beamforming algorithm. They better mimic human auditory systems, which cannot be replicated using just one microphone, not able alone to give spatial sound cues.

However, as microphones arrays are not so widespread, we also study how to handle the missing spatialized audio modality at test time. As a solution, we propose to distill acoustic images content to audio features during the training in order to handle their absence at test time. This is done for supervised audio classification using the generalized distillation framework, which we also extend for self-supervised learning.

Next, we devise a method for reconstructing acoustic images given a single microphone and an RGB frame. Therefore, in case we just dispose of a standard video, we are able to synthesize spatial audio, which is useful for many audio-visual tasks, including sound localization.

Lastly, as another example of restoring one modality from available ones, we inpaint degraded images providing audio features, to reconstruct the missing region not only to be visually plausible but also semantically consistent with the related sound. This includes also cross-modal generation, in the limit case of completely missing or hidden visual modality: our method naturally deals with it, being able to generate images from sound.

In summary we show how audio can help visual learning and vice versa, by transferring knowledge between the two modalities at training time, in order to distill, reconstruct, or restore the missing modality at test time.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Rationale

We, humans, perceive and interpret the world exploiting the available sensory information. Actually, multi-modal (or multi-sensory) perception is essential to interpret the world surrounding us, and sensory modalities are our primary channels of communication and sensation, such as vision or touch [13].

Designing computational systems able to emulate or surpass human capabilities in this respect is of utmost importance and constitutes a very challenging target. In fact, standard approaches typically learn a separate representation for each modality, which works well when operating within the same modality. However, the representations learnt are not aligned across modalities. Studying cross-modal representations is important for machines to understand relationships between modalities [12].

Among the different senses which both humans and machines can use to perceive the world, vision and hearing are surely the most commonly used and important, also because they are often quite correlated, temporally synchronized, and support each other for interpretation tasks [129]. More specifically, in humans, vision is supported by binaural hearing, which helps people focusing on the sound sources to better figure out what is happening around them. In fact, sound signals are received with a certain delay between the left and right ear (the so-called inter-aural time differences), as well as a slight difference in intensity (the so-called inter-aural level differences), which are critical to perceive spatial clues about the direction of provenience of the sound [124]. Besides, humans associate what they hear with what they see, and are thus able to fuse the spatial clues elaborated from their auditory system with those coming from their sight [101]. Vision and hearing are then often informative about the same structures in the world, and they are actually complementary. Even if vision

is guiding us the most, in some cases it is not reliable, and sound, since its propagation is not affected by illumination, camouflaging and occlusions, can support us in understanding scenes. Moreover, when there is no visual counterpart, some far or tiny objects, such as a plane in the sky or a gun shot in a crowded scene can be detectable only by their produced sound [177]. Therefore, sound may help to pay attention and visually focus on situations of interest and corroborate noisy or low-quality visual information, ultimately aiming at improving and fasten the interpretation of a scene. Also the opposite case can be true: humans are known to integrate audio-visual information in order to understand speech, where vision supports hearing, for example, to understand what a person is saying looking at the lips motion. Also McGurk effect [99] is well known, where a syllable is misunderstood when looking at wrong lip movements, indicating that the visual signals people receive from seeing a person speaking can influence the sound they hear.

In this thesis, we study audio-visual learning focusing mostly on three modalities: audio, optical images and acoustic images. Acoustic images constitute a spatialized audio information that, compared with monaural (single channel) audio, resemble human hearing better. In fact, in humans, audio modality contain significant spatial information. Nevertheless, video data recorded by a camera typically come with a monaural acoustic signal only. Hence, spatial cues are lost, and reliably recovering them is a difficult and only partially solved problem [41, 106]. Therefore, to mimic human hearing, usually binaural configurations are used to record audio from two microphones attached to the two ears of a dummy head approximating how humans receive sound signals [41]. However, binaural configurations are limited to the estimation of the direction of arrival only along the azimuth direction (the direction identified by the straight line joining the two microphones), and are not able to compete with the performance achieved by the human auditory system in localization tasks. Acoustic images, instead, can provide more accurate spatial audio information about acoustic frequency content and can localize sound sources on a 2-dimensional space, rather than along just one single direction as stereo audio [163]. So, acoustic images can resemble better human hearing and provide richer information as compared not only to mono, but also to stereo audio. Thus, in order to have the possibility to emulate human performance by exploiting spatially localized audio data, one needs to resort to an array of microphones positioned in special geometrical (e.g., planar) configuration, able to provide an enriched audio description of a scene. In fact, the acoustic signals gathered by a planar array of microphones (in our case, or a sonar [76] in general) can be properly combined via a beamforming algorithm [107, 149] to form an *acoustic image*. Acoustic images allow effectively to visualize the acoustic landscape of the sensed scene. We use spatialized audio information coupled to

the related visual data and design suitable multimodal deep learning models to get a better understanding of the scene, inspired by how humans learn.

We take advantage of a recent audio-visual sensor, named DualCam, composed by a microphone 2D array coupled with an off-the-shelf video camera placed at the device center, jointly calibrated, able to provide spatial acoustic data aligned with the corresponding optical image in space and time [177, 29]. The sensor is a planar array of 128 low-cost digital MEMS microphones located according to an optimized aperiodic layout. Acoustic images are obtained by combining the raw audio signals acquired by all the microphones using filter-and-sum beamforming algorithm implemented in frequency [149]. They are images where each pixel represents the spectral signature associated to the sound coming from a specific direction in space, which corresponds to a location in the optical image.

In our first work, we face audio-visual learning leveraging acoustic images considered as privileged information to be used only in training but missing at test time. In fact, as microphones arrays are expensive and not so widespread, we propose to distill acoustic images peculiar information at training time in order to be subsequently used in testing whenever they are not available, which is typically the case. Our intuition is that we can learn better audio features from acoustic images distillation, and use these improved representations during the testing phase, when single channel audio is the only available modality. Thus, we show that spatialized acoustic data allow to learn single-microphone audio models from which we can extract more discriminant and powerful features, that at test time, outperform in audio classification those built without this additional information [118]. For this setting, we use cross-modal feature learning, with the aim to learn better single modality representations given multiple modalities at training time. We achieve this knowledge transfer by distilling the spatial audio to audio features in a supervised way. The transfer is performed with the aid of the generalized distillation framework, which proposes to use the teacher-student approach from the distillation theory to extract knowledge from a privileged information source [96].

Subsequently, in our second work, we extend the distillation paradigm to the self-supervised scenario proposing a novel self-supervised distillation method. We take advantage of the alignment between modalities for a powerful source of self-supervision, which we exploit to learn a correspondence pretext task. In fact, if we supervise our models using the natural synchronization between vision and sound, we can learn good representations from unlabeled datasets [7]. The recent surge of interest in cross-modal learning from images and sound is due to the availability of virtually unlimited training material in the form of web videos that can be used to train deep networks [8]. In self-supervised learning, since no annotations are required, larger datasets can be collected and used to obtain models

generalizing better on different data. Thus, we train using correspondence task and then we test the audio and visual features obtained with our framework for downstream classification and cross-modal retrieval tasks. We show that audio-visual features are better than visual ones and that self-supervised representations are better than supervised ones [129].

Ultimately, we notice that we may need the whole missing modality at test time rather than its features only. Unfortunately, generating one modality from another one is a challenging problem, since we need to learn a mapping between their common information to make the translation across modalities be possible. It is a nontrivial task, mostly when different modalities are heterogeneous: for example, it can be difficult to relate raw pixels to audio spectrograms as we need to process them to a more abstract level to learn their correlations [109]. If we manage to extract powerful abstract features with a model associating these modalities, it is possible to recover the missing or damaged modality from the available ones based on the information shared between them.

Thus, in our last works, we study the generation of missing modalities in two cases. The former has the target to generate acoustic images from a single microphone and RGB frame. In fact, consumer-level cameras typically only record audio with a single microphone whereas microphone arrays are not so common. We hypothesize that, given the audio information and the spatial cues from video frames, it is possible to reconstruct the spatial audio modality, useful for many audio-visual tasks, such as classification, cross-modal retrieval and most noticeably sound source localization. We know that monaural audio data cannot bring any information about the spatial locations of the sound sources, but its accompanying visual frames do. Therefore, we propose to learn to generate acoustic images from a standard video, i.e., from single-microphone audio data and the visual content of the scene [128].

In a further work, we consider the opposite case: the visual modality is damaged and audio is essential to restore it. This because we address the visual inpainting problem [117, 73], aiming at filling masked images not just reconstructing a plausible visual content, but also restoring the original semantics of the scene, with the support of the associated audio features. In this way, the inpainted result is semantically consistent with the provided sound. By exploiting the common information shared between the audio data and the visual representation, it is possible to generate a realistic and semantically consistent content. Our method is also able to generate images just from sound, in the case the visual information is totally lost, in the same way as humans can imagine a scene from a sound. This is an example of cross-modal generation task, mapping from one modality space to a different, heterogeneous modality space, handling inter-sensory generation of images conditioned on sound [25].

To wrap up, the works presented in this thesis aim at the design of multimodal methods for scene understanding, able to handle missing modalities at test time with the help of multimodal learning, to enrich audio-visual representations, or to reconstruct a missing or damaged modality from the available ones. We show that, in fact, audio data can help visual learning and vice versa, by transferring knowledge between the two representations at training time, in order to distill, reconstruct or restore an unavailable or damaged modality at test time.

1.2 Contributions

Summarizing, the contributions of this thesis are the following.

1. A thorough study on a special modality representing spatial sound, acoustic images, and of the suitable methods to process it with deep learning.

2. Methods to distill acoustic images' content to audio and visual representations, in supervised and self-supervised manner.

3. Design of pipelines to handle cross-modal generation of a modality from different ones, in case we need to generate a modality not present at test time. In particular, we addressed both generation of acoustic image from available image and single-microphone sound, and restoring an image through the semantic content provided by sound and visual context.

1.3 Publications

The work presented in this thesis has produced the following publications:

- A. F. Pérez, V. Sanguineti, P. Morerio, and V. Murino. "Audio-visual model distillation using acoustic images". Winter Conference on Applications of Computer Vision (WACV) 2020

- V. Sanguineti, P. Morerio, N. Pozzetti, D. Greco, M. Cristani, and V. Murino. "Leveraging Acoustic Images for Effective Self-Supervised Audio Representation Learning". European Conference on Computer Vision (ECCV) 2020

- V. Sanguineti, P. Morerio, A. Del Bue, and V. Murino, “Audio-Visual Localization by Synthetic Acoustic Image Generation”, AAAI Conference on Artificial Intelligence 2021

1.4 Thesis Organization

This thesis is organized as follows. First of all, in chapter 2 we present the related works useful to have a background on audio-visual learning. In chapter 3 we talk about supervised distillation of acoustic and video teacher models to audio student model. Then, in chapter 4, we present self-supervised learning and self-supervised distillation using acoustic images. After that, in chapter 5, we explain how to generate acoustic images from videos collected with off-the-shelf cameras. In the opposite way, in chapter 6 we propose to restore images by using audio data. Finally, in chapter 7, we draw some conclusions.

Chapter 2

Related Works

In this chapter, we will explain basic concepts related to audio and acoustic images and the state of the art of audio-visual learning. Firstly, we explain how to compute audio features and acoustic images. Then we talk about multimodal supervised learning and self-supervised audio-visual learning. Finally, we show some examples of audio-visual applications.

2.1 Audio Features Extraction

Audio comprises a rich source of information that is complementary to visual information and often informative about the same structures in the world.

With deep learning audio models, we can attempt to learn directly from the signal in the time domain. However, it is difficult to learn the Fourier transform, which may arguably increase the model complexity. Therefore, usually some audio features are computed and then passed to nets.

Many steps are involved to compute audio features. In summary, a signal goes through a pre-emphasis filter; then gets sliced into (overlapping) frames and a window function is applied to each frame; afterwards, we do a Fourier transform on each frame (more specifically a Short-Time Fourier Transform) and calculate the power spectrum; and subsequently compute the filter banks, filtering with Mel filters and computing log of the magnitude. To obtain Mel-Frequency Cepstral Coefficients [122], a Discrete Cosine Transform (DCT) is applied to the filter banks. We retain few of the resulting coefficients while the rest are discarded[1]. We can perform this process both on single microphone audio and to compress acoustic images. We explain now every stage in detail.

[1] https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html

The first step, pre-emphasis filter on the raw signal in the time domain, has the target to amplify the high frequencies, to balance the frequency spectrum since high frequencies usually have smaller magnitudes compared to lower frequencies.

The rationale behind splitting the signal into short-time frames is that frequencies in a signal change over time, so in most cases it doesn't make sense to do the Fourier transform across the entire signal. We can safely assume that frequencies in a signal are stationary over a very short period of time: windowing is used to analyze sound samples for stable acoustic characteristics. Therefore, by doing a Fourier transform over this short-time frame, we can obtain a good approximation of the frequency of the signal. Typical frame sizes in speech processing range from 20 ms to 40 ms with 50% (+/-10%) overlap between consecutive frames. Popular settings are 25 ms for the frame size and a 10 ms stride, thus 15 ms overlap among the frames. For acoustic images, we consider short segments, 83.3 ms windows on which signal is assumed to be stationary.

After slicing the signal into frames, we apply a window function such as the Hamming window to each frame to counteract the assumption made by the FFT that the data is infinite and to reduce spectral leakage. More specifically, on each frame, we apply a Tukey window to taper the signal toward the frame boundaries. This is done to smooth the edges while taking FFT of the signal.

Short-Time Fourier Transform (STFT) defines a time-frequency distributions which specify complex amplitude versus time and frequency for any signal. The STFT is computed as a succession of FFTs : we apply FFT on each frame to calculate the frequency spectrum and then compute the power spectrum.

The final step for computing filter banks is applying triangular filters on a Mel-scale to the power spectrum to extract frequency bands. A Mel is a unit of measure based on human ears perceived frequency: the Mel-scale aims to mimic the non-linear human ear perception of sound, by being more discriminative at lower frequencies and less discriminative at higher frequencies. Each filter in the filter bank is triangular having a response of one at the center frequency and decreases linearly till it reaches the center frequencies of the two adjacent filters where the response is zero. The Mel spectrum $x(m)$ of the magnitude spectrum $S(k)$ of the signal, is obtained multiplying squared magnitude spectrum by the triangular Mel filters $H_m(k)$:

$$x(m) = \sum_k \left[|S(k)|^2 H_m(k) \right], \tag{2.1}$$

for $0 \leq m \leq M-1$, where M is the number of filters, $H_m(k)$ is the weight given to the k energy spectrum bin contributing to the output m. We notice that the Mel spectrum are

M values, one for each Mel filter, corresponding to the energy of the signal filtered by the corresponding filter. It usually is followed by a log operation, computing log Mel spectrum.

It turns out that filter bank coefficients (Mel frequency energies represented on a log scale) computed in the previous step are highly correlated, because the energy levels in adjacent bands are likely to be correlated, which could be problematic in some machine learning algorithms. Therefore, we can apply Discrete Cosine Transform (DCT) to decorrelate the Mel frequency energies represented on a log scale, a process also referred to as whitening. We obtain a compressed representation of the filter banks, the Mel-Frequency Cepstral Coefficients (MFCC). The DCT $c(n)$ (MFCC) of a signal $\log x(m)$ (log Mel filter energies) is defined as:

$$c(n) = \begin{cases} \sqrt{\frac{2}{M}} \sum_{m=0}^{M-1} \log x(m) \cos\left(\frac{\pi}{M} \cdot n \cdot (m+0.5)\right) \text{ for } n = 1, \dots C \\ \frac{1}{\sqrt{M}} \sum_{m=0}^{M-1} \log x(m), \text{ for } n = 0 \end{cases} \tag{2.2}$$

DCT also can be written as matrix- multiplication with a "Transformation Matrix" T:

$$c = \log x \cdot T, \tag{2.3}$$

where

$$T = \sqrt{\frac{2}{M}} \begin{bmatrix} \frac{1}{\sqrt{2}}, \cos\left(\frac{\pi}{M} \cdot 1 \cdot (0+0.5)\right), \dots, \cos\left(\frac{\pi}{M} \cdot C \cdot (0+0.5)\right) \\ \dots \\ \frac{1}{\sqrt{2}}, \cos\left(\frac{\pi}{M} \cdot 1 \cdot (M-1+0.5)\right), \dots, \cos\left(\frac{\pi}{M} \cdot C \cdot (M-1+0.5)\right) \end{bmatrix} \tag{2.4}$$

The system can just extract first few MFCC which have most information, so $C < M$ and Transformation Matrix is not a square matrix. In our datasets $M = 24$ is the number of employed Mel filters and $C = 12$ is the number of employed MFCC. We can discard the last MFCC coefficients because they represent fast changes in the filter bank coefficients and these fine details don't contribute to classification.

One may apply sinusoidal liftering to the MFCCs to de-emphasize higher MFCCs which has been claimed to improve speech recognition in noisy signals.

The coefficient $c(0)$ was not included in the MFCC representation of acoustic images, because it is log energy: if we substitute Eq. (2.1) in Eq. (2.2) for $n = 0$ we get

$$c(0) = \sqrt{\frac{1}{M}} \sum_{m=0}^{M-1} \log \left(\sum_k \left[|S(k)|^2 H_m(k) \right] \right), \tag{2.5}$$

This corresponds to the sum of all log energies obtained filtering the magnitude spectrum by Mel filterbank dividing by $\sqrt{M}$. Therefore we discard coefficient 0 because it is proportional to the average log-energy of the input signal, which carries little sound discriminant information [122].

2.2 Acoustic Images

Acoustic images have not caught much attention in the surveillance community, despite their several advantages. These images, resulting from acoustic beamforming applied to the signals acquired by a set of microphones, encode at each pixel the sound intensity coming from each spatial direction. It is a completely passive technology, differently from active radars, optical and infrared cameras which require light emitters. Such features make acoustic imaging devices suited to extend the functionalities of current surveillance systems in those scenarios where the target of interest is characterized also by a sound signature. In particular, the coupling together with a video camera may enable compelling applications such as the visual localisation of events that are difficult to understand only using the video signal, such as a gunshot [177].

We acquired the acoustic images using the DualCam, a prototype of acoustic-optical camera described in [177]. The optimized aperiodic microphone layout and processing parameters necessary for beamforming allow to obtain an optimal acoustic image quality in terms of spatial resolution, dynamic range and robustness to diverse environmental conditions, while keeping limited the amount of hardware and software resources. It has an on-board hybrid embedded processor in the lower right corner and a webcam (so that it doesn't have any appreciable geometric distortion) at the center of the device as depicted in Figure 2.1. So, the acoustic images are geometrically overlapped, by design, with the optical ones.

The sensor captures both audio and video data. It is a $0.45m \times 0.45m$ planar array of 128 low-cost digital MEMS microphones capable of acquiring audio data with a sampling frequency of 12 kHz in the useful bandwidth 500 Hz – 6 kHz and acoustic images and video sequences at a frame rate of 12 frames per second (fps). The device can record all frequencies from 0 up to 6 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 camera has a maximum field of view of 90° in elevation and 360° in azimuth (tunable according to the video camera field of view). The acoustic image resolution[2] provided by the sensor is of 5° at 6000 Hz. The data provided by the sensor consists in RGB video frames of 640×480 pixels,

[2] Measured at -3 dB.

Figure 2.1 DualCam acoustic-optic camera

raw audio data from 128 microphones, and $36 \times 48 \times 512$ multispectral acoustic images. This means that each acoustic pixel corresponds to 13,3 visual pixels, in fact acoustic resolution is lower than optical one.

Multispectral acoustic images are obtained from the raw audio signals of all the microphones using the frequency implementation of the filter-and-sum beamforming algorithm [149], which summarizes the audio intensity for every direction and discretized frequency bin. Full details of the algorithm can be found in [177]. In practice, a set of delays aligns the signals coming from a given point in the array so that they can be summed coherently in the beamforming procedure.

Acoustic images are volumes of size $36 \times 48 \times 512$, with 512 channels corresponding to the frequency bins discretizing frequency content, representing FFT squared magnitude spectrum. We can visualize (as heat color map) the energy of sound of an acoustic image summing the energy of all frequency bins and overlaying it on the corresponding video frame, as shown in Figure 2.2.

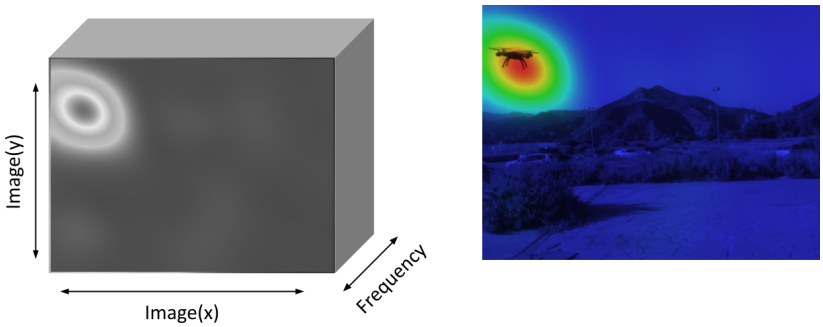

Figure 2.2 Acoustic image and its energy.

Handling input acoustic images with 512 channels is a computationally expensive task and typically the majority of information in our dataset is contained in the low frequencies. Consequently, we decided to compress the acoustic images along the frequency axis using Mel-Frequency Cepstral Coefficients (MFCC), which consider audio human perception characteristics [144]. Therefore, we compute 12 MFCC as explained in section 2.1, going from from $36 \times 48 \times 512$-D volumes to $36 \times 48 \times 12$-D volumes, retaining the most important information and reducing consistently the computational complexity and the required memory, but also resulting 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 the literature.

2.3 Multimodal Learning

We experience the world in a multimodal way - we see objects, hear sounds, feel texture, smell odors, and taste flavors, using our primary channels of communication and sensation, such as vision or touch. In a similar way, we want to make models understand the world around us to resemble how humans feel it, making them able to process and relate information from multiple modalities. For a deep neural network they can be for instance: visual signals represented with images or videos; sounds; depth; optical flow; natural language.

In Deep Learning, standard approaches typically learn a separate representation for each modality, which works well when operating within the same modality. However, the representations learned are not aligned across modalities. Multimodal learning aims at using the complementary information of different modalities, capturing correspondences between modalities and their relationship to learn useful joint representations. Fusing information from different modalities is usually non trivial due to the heterogeneity of the modalities. However, heterogeneous data modalities can improve several tasks, usually bringing more robust algorithms and better performance.

There are five problems to address in multimodal learning [13]:

1. Representation
2. Translation
3. Alignment
4. Fusion
5. Co-learning

A first fundamental challenge is learning how to represent and summarize multimodal data in a way that exploits the complementarity and redundancy of multiple modalities. [18] proposes to replace the earlier layers of the network (which are modality specific) and freeze the later layers for cross-modal alignment. First they learn a source representation that will be utilized for all modalities. They then train each specific network to categorize scenes in its modality while holding the shared higher layers fixed. Cross-modal representations can be useful, for example, in case data in one modality may be difficult to acquire for privacy, legal, or logistic reasons (eg, images in hospitals), or is of poor quality or noisy: other modalities can be more common, allowing us to train models boosting the performance [52].

A second challenge addresses how to map data from one modality to another one. Given one modality, the task is to generate a correspondent modality. One of oldest works regarding generation of one modality feeding another one is [109], which could reconstruct both audio and video from only video or audio using auto encoders. They train a deep autoencoder to reconstruct both modalities when given only one and thus discover correlations across the modalities, or pass examples that have zero values for one of the input modalities (e.g., video) and original values for the other input modality (e.g., audio), but still require the network to reconstruct both modalities (audio and video) to learn a model which is robust to inputs where a modality is absent. The Joint Multimodal Variational Autoencoder [141] generates a certain image given text modelling the joint representation of the two modalities. The joint representation of all modalities is learnt by the Multimodal Variational Autoencoder [160] too, which can generate any modality from joint latent variable. These models create a shared latent space between modalities that can be used to generate one from the other ones.

A third challenge is multimodal alignment, finding relationships and correspondences between modalities. Deep learning based approaches for alignment are becoming popular due to availability of aligned datasets. They measure a similarity metric between modalities and try to minimize the distance between them, in such a way that correspondent representations have a smaller distance between them than non-corresponding ones. For example, such models encourage the representation of the audio of barking dog and its image to have a smaller distance between them than distance between the audio of barking dog and an image of a car. When neural networks are good for this task, they can solve cross-modal retrieval problem [8]. In fact, learnt embeddings which can be used to analyze the effectiveness of learned representations. These embeddings are encouraged to have a shared space that allows them to be comparable by a similarity metric. The semantic quality of the embeddings can be analyzed: [133] conduct the sound query based video retrieval and vice versa and they report

the success ratio of semantically meaningful matches, conducting the k-nearest neighbor search by measuring the distance of audio and visual embeddings.

A fourth challenge is to join information from many modalities. First, having access to multiple modalities that observe the same phenomenon may allow for more robust predictions. Second, it might allow us to capture complementary information. Approaches can be split into early (i.e., feature-based) and late (i.e., prediction-based). On one hand, early fusion integrates features immediately after they are extracted (often by simply concatenating their representations). It learns the correlation and interactions between low level features of each modality. Late fusion, on the other hand, performs integration after each of the modalities has made a decision (e.g., classification or regression): it uses unimodal decisions and fuses them through a fusion mechanism such as averaging. A late fusion scheme tends to give better performance for most concepts, but it comes with the price of an increased learning effort [138].

A fifth challenge is to transfer knowledge between modalities. Co-learning explores how transferring information from a representation built using a data rich or clean modality can help a model representation learnt using data scarce or noisy modality to get better representations, exploiting the complementary information across modalities. This challenge is particularly relevant when one of the modalities has limited annotated data but modalities share a set of instances, e.g. audio recordings with the corresponding videos. The helper modality is used only during model training and is not used during test time. The aim is to learn better single modality representations given unlabeled data from multiple modalities. This leads to better unimodal representations, for the modality being used alone during test time. For instance, [5] hypothesises that the emotional content of speech correlates with the facial expression of the speaker to transfer annotations from the visual domain (faces) to the speech domain (voices) through cross-modal distillation. This is done because obtaining large labelled speech datasets to train models for emotion recognition is a challenging task, so very few labelled audio datasets are available for emotion recognition.

2.4 Multimodal Supervised Learning

We show in this section some example of tasks involving multimodal information which use supervised learning.

2.4.1 Multiple Streams Networks

We can use many streams, implemented as convolutional neural networks, one for each specific modality, which are then combined by late fusion.

However, even if a multi-modal network receives more information, so it should match or outperform its uni-modal counterpart, recently [156] showed the opposite: the uni-modal network often outperforms the multi-modal network. This is due to the fact that different modalities overfit and generalize at different rates. Additionally, late fusion multimodal network has nearly two times the parameters of a unimodal network, and one may suspect that the overfitting is caused by the increased number of parameters. The problem can be solved avoiding joint training by using pre-trained uni-modal features.

In fact, many two-stream networks for video classification [157, 17, 36, 136] do not train multiple modalities jointly. [136], for example, proposes a two-stream ConvNet architecture which incorporates spatial and temporal networks. The softmax scores of the two streams are combined by late fusion to capture the complementary information from still frames and motion between frames. In fact, video can naturally be decomposed into spatial and temporal components. The spatial part, in the form of individual frame appearance, carries information about scenes and objects depicted in the video. The temporal part, in the form of motion across the frames, conveys the movement of the observer (the camera) and the objects.

2.4.2 Transfer Learning

Transfer learning deals with sharing information from one task to another one. We can transfer knowledge between nets fed by the same modality to obtain a smaller network with close performances to a heavier network. [63] compresses discriminative knowledge from a well-trained complex model by distilling to a simpler model, lightweight network mimicking the ensemble of networks without loosing considerable accuracy. In fact, a very simple way to improve the performance of almost any machine learning algorithm is to train many different models on the same data and then to average their predictions, as an ensemble of networks usually performs better than a single network. Unfortunately, making predictions using a whole ensemble of models may be too computationally expensive, especially if the individual models are large neural nets. After distillation, we can use the simpler student for prediction at test time. Neural networks typically produce class probabilities by using a "softmax" output layer that converts the logit z_i computed for each class into a probability, q_i,

by comparing with the other logits.

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \tag{2.6}$$

where T is a temperature that is normally set to 1. Using a higher value for T produces a softer probability distribution over classes. [63] transfers knowledge to the distilled model by training it on a transfer set and using a soft target distribution for each sample in the transfer set that is produced by using the bigger model with a high temperature in its softmax. When the correct labels are known for the transfer set, this method can be significantly improved by also training the distilled model to produce the correct labels. They simply use a weighted average of two different objective functions.

Another approach is to transfer knowledge between networks operating on different data modalities if we have more annotated data in one modality than in a different one. For example, we can transfer knowledge from vision to other modalities, using teacher-student models leveraging the natural synchronization between modalities. [53] transfers visual supervision into depth models. [11] leverages the natural synchronization between vision and sound in unlabeled videos in order to learn a representation for sound. They make state-of-the-art networks for vision teach sound student model SoundNet to recognize scenes and objects, using unlabeled video as a bridge. They transfer from visual networks using the posterior probabilities from a teacher vision network $g_k(y_i)$ in order to train the student network $f_k(x_i)$ to recognize concepts given sound, optimizing

$$\min_{\theta} \sum_{k=1}^{K} \sum_{i=1}^{N} D_{\mathrm{KL}}\left(g_k(y_i) \,\|\, f_k(x_i;\theta)\right) \tag{2.7}$$

where

$$D_{KL}(P\|Q) = \sum_j P_j \log \frac{P_j}{Q_j} \tag{2.8}$$

is the KL divergence, transferring from both scene and object visual networks (K = 2). They chose KL-divergence because the outputs from the vision network g_k can be interpreted as a distribution of categories. They only rely on the teachers soft labels for a visual frame, which are in general less reliable than hard labels, to train a student network to recognize sound correspondent to that visual frame. Learned audio features are good for environmental sound recognition. Oppositely, [112] transfers knowledge from audio to video. They train a CNN to predict the audio features of images and cluster the audio textures. Then, they train a video network to solve a classification problem to predicts a frame's auditory cluster

assignment. They empirically evaluate the quality of the learned representation for several image recognition task such as object and scene recognition.

[12] learns through massive amounts of synchronized data cross-modal representation shared across three major natural modalities: vision, sound and language. They take advantage of discriminative visual models to teach a student model to have an aligned representation, transferring the knowledge from the teacher vision model to sound and text student models. Since the modalities are synchronized, they train the student model to predict the class probabilities from the teacher model using the KL-divergence as a loss. However, the internal representations of the students would not be aligned since each student model is disjoint. To enable an alignment to emerge in the internal representation, they constrain the upper layers of the network to have shared parameters across modalities, while the early layers are specific to the modality. For aligned representations they employ a ranking loss to push paired examples close together in representation space, and mismatched pairs further apart, up to some margin. They quantify the learned alignment by evaluating their representations at a cross-modal retrieval task.

In multimodal learning, we can have an additional modality at training time. Using the generalized distillation framework, which proposes to use the teacher-student approach from the distillation theory, we extract knowledge from a privileged information source at training time [96]. The teacher considers additional information, together with the labels, both not available at test time, to build a classifier for test time that outperforms those built on the regular features. [63] transfers the teacher f_t into

$$f_s = \arg\min_{f \in \mathscr{F}_s} \frac{1}{n} \sum_{i=1}^{n} \left[(1-\lambda) \ell \left(y_i, \sigma \left(f \left(x_i \right) \right) \right) + \lambda \ell \left(s_i, \sigma \left(f \left(x_i \right) \right) \right) \right] \tag{2.9}$$

using

$$s_i = \sigma \left(f_t \left(x_i \right) / T \right) \tag{2.10}$$

the soft predictions from f_t about the training data. The temperature parameter $T > 0$ controls how much do we want to soften or smooth the class probability predictions from f_t, and the imitation parameter $\lambda \in [0,1]$ balances the importance between imitating the soft predictions s_i and predicting the true hard labels y_i. Higher temperatures lead to softer class-probability predictions s_i. In turn, softer class-probability predictions reveal label dependencies which would be otherwise hidden as extremely large or small numbers. The soft labels (dense vectors with a real number of information per class) contain more information than hard labels (one-hot-encoding vectors with one bit of information per class) allowing for faster

learning. This information is label uncertainty. After distillation, we can use the simpler student for faster prediction at test time. [96] proposes to use distillation to extract useful knowledge from privileged information, learning the teacher function f_t using the privileged information.

[64] presents a modality hallucination architecture for training an RGB object detection model which incorporates depth side information at training time to improve test time RGB only detection models. The hallucination network learns to mimic features from a depth network. At test time images are processed jointly through the RGB and hallucination networks to produce improved detection performance. Similarly, [44] addresses action recognition by distilling knowledge from a depth network into a vision network to improve single-modality system performance. They accomplish this by training a hallucination network that learns to distill depth features with the aid of the generalized distillation framework.

We also use the generalized distillation framework in our work to extract useful knowledge from a novel privileged information, acoustic images [118]. The proposal is as follows. First, learn a teacher function by using the acoustic images. Second, compute the teacher soft labels. Third, distill teacher into audio model by using both the hard and soft labels.

2.5 Audio-Visual Self-Supervised Learning

Self-supervised learning is a subset of unsupervised learning where the data provides the supervision to train a model. It needs both a pretext task and then a downstream task. By defining a proxy task, the neural network learns representations without labels, which should be useful for the target task, such as classifying other datasets different from training one, for cross-modal retrieval and audio-visual localization.

The correlation between audio and visual modalities can be used as a supervisory signal for self-supervised learning. For the correspondence task positives are extracted from the same video, while negatives are a frame and audio extracted from different videos. Other audio-visual pretext tasks are: audio-visual synchronization, audio-visual spatial correspondence and clustering. We explain here more in detail each pretext task. Finally, we also show some examples of transformers which use these pretext tasks for being trained.

2.5.1 Audio-Visual Correspondence

A proxy task can be the correspondence between synchronized modalities. As a matter of facts, audio-visual correspondence is most widespread pretext task for audio-visual self-supervised learning. The correspondence in video clip provides strong supervisory signal for synchronous nature of data and it is free. For instance, [7] jointly trains visual and audio networks to predict whether a given pair of audio and visual examples were sampled from the same video, in an unsupervised manner, using a large number of unlabelled videos. Positives are extracted from the same video, while negatives are a frame and audio extracted from different videos. To tackle the Audio-Visual Correspondence (AVC) task, they propose a network with three distinct parts: the vision, the audio network and the fusion network which concatenates these features to produce the final decision on the alignment. They show that they are able to localize the source of the audio event in the video frame. Video and audio networks are useful to extract visual and audio features for classifying other datasets.

[8] embeds audio and visual inputs into a common space that is suitable for cross-modal retrieval. They localize the object that sounds in an image, given the audio signal. They achieve both downstream tasks by training from unlabelled videos using only audio-visual correspondence (AVC) as the objective function.

2.5.2 Audio-Visual Synchronization

Temporal synchronization is a harder problem to solve than semantic correspondence, since it requires to determine whether the audio and the visual samples are not only semantically coherent but also temporally aligned. This problem is called “Audio-Visual Temporal Synchronization” (AVTS) [82]. To ease the learning, it is beneficial to adopt a curriculum learning strategy, where harder negatives are introduced after an initial stage of learning on easier negatives. While AVC requires just an image and an audio track, here many video frames are needed, because this task requires motion analysis and is not solvable with a single frame. A positive example is obtained by extracting the audio and the visual input from a randomly chosen video so that the video frames correspond in time with the audio segment. There are two main types of negative examples. Easy negatives are those where the video frames and the sound come from two different videos. Hard negatives are those where the pair of out-of-sync audio and video segment is taken from the same video, but there is at least half a second time-gap between the audio sample and the visual clip. While easy negatives can be learnt with semantics, hard negatives require synchronization. The

curriculum learning using some harder negatives yields better results in terms of performance on the downstream tasks (audio classification and action recognition).

[110] predicts temporal misalignment in synthetically-shifted videos. The network observes raw audio and video streams some of which are aligned, and some that have been randomly shifted by a few seconds. They use this learned representation for three applications: sound source localization; audio-visual action recognition; and on/off-screen sound source separation to separate the speakers' voices by visually masking them from the video.

2.5.3 Audio-Visual Spatial Correspondence

Prior works ignore the spatial cues of audio-visual signals. Since these methods do not need to localize sound sources, they struggle to discriminate visual concepts corresponding to the same sound. In fact, the audio is a descriptor for the whole video clip, as opposed to the region containing the sounding class, for example, a car. Since cars and roads often co-occur, there is an inherent ambiguity about which of the two produces the sound. This makes it hard to learn good representations for visual concepts like "cars", distinguishable from co-occurring objects like "roads" by pure audio-visual correspondence or temporal synchronization. For example, the model of [132] activates on the road given car sound, because it obtains a good score (as it is paired). Since road has simpler appearance and typically occupies larger regions compared to diverse appearance of the car (or non-existence of the car in the frame), it is difficult for the model to discover the true causality relationship with the car without supervisory feedbacks. This ends up biasing toward a certain semantically unrelated output, as in pigeon superstition problem. [132] provides some prior knowledge with supervisory signals in the semi-supervised setting to make the algorithm learns successfully.

Another solution is that of [105], learning representations by performing audio-visual spatial alignment (AVSA) of 360° video and spatial audio as they contain strong spatial cues. They design a pretext task where audio and video clips are sampled from different viewpoints within a 360° video and spatially misaligned audio/video are treated as negatives examples for contrastive learning. They generate random rotation of either the video or audio so as to create an artificial misalignment between them. A model can then be trained to predict the applied rotation, but this is difficult to optimize. Thus, they propose to solve the audio-visual spatial alignment task in a contrastive manner. Given a 360° audio-video sample, K video and audio clips are extracted from K randomly sampled viewing angles. Audio-visual spatial alignment is then encouraged by making the model predict the correct

correspondence between the K video and the K audio signals. They show the benefits of AVSA pretext task on action recognition and video semantic segmentation.

The intrinsic temporal synchronization, but also the spatial alignment of visual and acoustic images can be exploited as a supervisory signal for audio-visual learning. In our work [129], we aligned acoustic images, images and audio in space and time to learn more powerful audio-visual representations via knowledge distillation. First, we train using the audio-visual correspondence pretext task the acoustic images' stream jointly with the RGB stream in a self-supervised way and, second, the audio stream with the RGB stream. Then, the knowledge of pre-trained acoustic image teacher is distilled, through a self-supervised learning scheme, to an audio stream, trained using the correspondence pretext task. We employ an additional triplet loss between the single-audio and the acoustic-image embeddings vectors. Such loss tries to transfer effective embeddings learned with acoustic images network to the monaural audio model. 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. We show that when training with the additional supervision of acoustic images features, audio-visual features are boosted. 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.

Another example of pretext task to learn spatial alignments is that proposed by [163], which determines whether the left and right audio channels have been flipped, forcing the architecture to reason about spatial localization across the visual and stereo audio streams. During training, they provide as input video clips where they flip the order of the channels in the audio stream with probability 0.5. They train the neural network to maximize a classification cross-entropy objective, predicting a label indicating whether or not audio is flipped with respect to visual frame, conjecturing that solving the flipping task requires understanding audio-visual spatial correspondence, as the model must match the location of objects in audio signals with the location of objects in visual signals. Understanding spatial correspondence enables models to perform better on three audio-visual tasks: sound localization, audio spatialization (upmixing a single mono audio channel to stereo binaural audio channels) and sound source separation.

2.5.4 Clustering

Correspondence audio-visual works rely on simple global correspondence. When there exist multiple sound-producers in the shown visual modality, it becomes difficult to exactly

locate the correct producer. However, the real-life acoustic environment is usually a mixture of multiple sounds. Hence, [65] proposed to jointly disentangles the audio and visual components with clustering, and establishes elaborate correspondence between multiple audio-visual objects. Therefore, they co-cluster audio-visual features into corresponding components. Finally, the model takes the similarity across modalities as the supervision for training. Downstream tasks are localization and audio-visual features classification.

Visual and audio modalities are highly correlated, yet they contain different information. Their intrinsic differences make cross-modal prediction a potentially more rewarding pretext task for self-supervised learning. [6] leverages unsupervised clustering in one modality as a supervisory signal for the other modality. This cross-modal supervision helps to utilize the semantic correlation and the differences between the two modalities. At each deep clustering iteration, they cluster the audio deep features and use their cluster assignments as pseudo-labels to train the visual encoder and viceversa. Cross-Modal Deep Clustering (XDC) yields representations that generalize better to the downstream tasks of action recognition and audio classification, compared to their within-modality counterparts. This underscores the complementarity of audio and video and the benefits of learning label-spaces across modalities.

2.5.5 Transformers

Transformers [152] were originally built for natural language processing tasks and the design of multi-head attention shows its effectiveness on modeling long-term correlation of words.

Transformers have been used in audio-visual learning because they are armed with self-attention modules that are well-known to produce powerful multimodal representations [19]. In addition, multimodal videos are abundantly available and their temporal, cross-modal content and therefore supervision, requires no human annotation. For instance, [38] tackles the tasks of caption-to-video and video-to-caption retrieval using a multimodal transformer exploiting the self-attention mechanism to leverage cross-modal and temporal information in videos effectively, with a video encoder that handles all the constituent modalities (appearance, audio, speech) jointly. Another example is the convolution-free Video Audio-Text Transformer (VATT) [4], trained with a self-supervised learning strategy from scratch, using multimodal contrastive losses. It takes raw signals as inputs and extracts multimodal representations to benefit a variety of downstream tasks: video action recognition, audio event classification, image classification, and text-to-video retrieval. Recently [71] proposes the Perceptron, a model architecture which builds upon Transformers to be more flexible:

no architectural changes are required to use the model on a diverse range of modalities. The Perceiver applies the cross-attention module and the Transformer in alternation. This corresponds to projecting the higher-dimensional byte array to a fixed-dimensional latent bottleneck before processing it with a deep Transformer.

Future works regarding acoustic images include using DEtection TRansformer, or DETR [16], which is a transformer encoder-decoder architecture. It predicts (in parallel) the final set of detections by combining a common CNN with a transformer architecture. Visualizing the attention maps of the last encoder layer of a trained model, we notice that they focus on a few points in the image. The encoder seems to separate instances, which likely simplifies object localization for the decoder.

Therefore, training a DETR for RGB images and an audio transformer, using projection heads to respectively map the video and audio transformers encoders outputs to the video-audio common space, and studying their correspondence, could help to localize sounding objects' in the space in case of multiple sound sources. In addition, we could make transformers take into account temporal content too, to disambiguate which instances are producing sound in a certain moment, using temporal-positional encodings. In fact, while encodings are typically used to encode sequence position in the context of language, they can also be used to encode spatial and temporal information. In such a way, we can exploit both spatial and a longer temporal information compared to the one we use in chapter 5 for the audio-visual localization task. Furthermore, we do not necessarily need annotations such as bounding boxes in DETR, given the fact that acoustic images can be aligned to RGB images and provide the region from where the sound is produced, so that our method can be completely self-supervised, having as supervision a data representation.

Another possible task that could be performed with DETR and a sound transformer is the acoustic images' generation. In fact, the transformers can effectively capture the long-term relationships between visual and audio information using the self-attention mechanism for sequence prediction. As [39] did, transformers can be employed in architectures for cross-modal generation tasks: while they translated musician movements to music, a more challenging task is the opposite one, that is to say, translate audio information (single channel or acoustic image) into visual information. This transformer approach can be an alternative to the method proposed in chapter 6.

2.6 Audio-Visual Applications

Audio-visual (usually self-supervised) applications enumerate speaker separation, sound separation, sound spatialization, audio-visual localization and cross-modal generation of one modality from another one.

2.6.1 Speaker Separation

Humans are remarkably capable of focusing their auditory attention on a single sound source within a noisy environment, which is known as the cocktail party effect. Research has shown that viewing a speaker's face enhances a person's capacity to resolve perceptual ambiguity in a noisy environment.

Audio-visual models, in the same way, use visual features to "focus" on the desired speaker in a scene and improve the audio separation quality [35, 1]. More specifically, [35] uses an architecture which takes detected faces and noisy audio, containing a mixture of speech and background noise, as input and outputs complex spectrogram masks. They use an off-the-shelf face detector to find faces in each frame and a pretrained face recognition model to extract one face embedding per frame for each of the detected face. The network outputs a complex spectrogram mask for each speaker, which is multiplied by the noisy input and converted back to waveforms by performing inverse STFT (ISTFT), to split the mixture into enhanced separate speech signal for each speaker, while suppressing all other interfering signals. In previous work, multiplicative masks have been observed to work better than direct prediction of spectrogram magnitudes or direct prediction of time-domain waveforms. The squared error (L2) between the power-law compressed clean spectrogram and the enhanced spectrogram is used as a loss function to train the network.

2.6.2 Sound Separation

Given a video with many sounds sources, we want to separate each sound track. [171] introduces a system that, by leveraging large amounts of unlabeled videos, learns to locate image regions which produce sounds and separate the input sound into a set of components that represent the sound from each pixel. The input audio spectrogram is passed through a U-Net whose output is K audio components feature maps. The audio synthesizer network predicts the sound by taking pixel-level visual feature and audio feature and outputs masks to be applied to the input spectrogram that will select the spectral components associated with a certain pixel. Specifically, a mask that could separate the sound of a certain pixel from the

input is estimated, and multiplied with the input spectrogram. Finally, to get the waveform of the prediction, inverse STFT is applied to the computed spectrogram to produce the final sound. They propose the Mix-and-Separate framework for source separation according to vision. The Mix-and-Separate training procedure artificially adds together the audio signals from two videos to generate an input mixture with known constituent source signals. It then solves the problem of separating and grounding sounds of interest conditioned on the visual input. The model learns to estimate the sounds in each video, given the audio mixture and the visual frames of the corresponding video. The system thus learns to separate individual sources without traditional supervision, no annotations are provided. The model accurately localizes the sounding objects.

[42] starts with the commonly adopted Mix-and-Separate framework of [171], which implicitly assumes that the original real training videos are dominated by single-source clips. Instead they learn from unlabeled videos containing multiple sound sources to return a separate sound track for each object. They propose a new co-separation network which considers pairs of training videos and, rather than simply separate their artificially mixed soundtracks, it must generate audio tracks associating consistent sounds to similar-looking objects across pairs of training videos. They enforce separation within a single video at the object level. They use a pre-trained object detector to find objects in both videos to visually guide audio source separation. Each detected object is a potential sound source. The goal is to separate the sound for each object in the mixture, conditioned on the localized object. To perform separation, they predict a spectrogram mask for each object and obtain the predicted magnitude spectrogram by soft masking the mixture spectrogram. Finally, they use the inverse short-time Fourier transform (ISTFT) to reconstruct the waveform sound for each object source. For each video, summing up the separated sounds of all objects should ideally reconstruct the audio signal for that video so that the corresponding audios for each of the pair of input videos can be reconstructed. They also introduce an object-consistency loss for each predicted audio spectrogram. The intuition is that if the sources are well-separated, the predicted "category" of the separated spectrogram should be consistent with the category of the visual object that initially guides its separation. More in detail, for the predicted spectrogram of each object, they use the cross-entropy loss to target the weak labels of the input visual objects. Ambient sounds, noise, offscreen sounds are collectively designated as having an additional audio label. The object consistency loss only knows that same object sounds should be similar after training the network, not what any given object is expected to sound like.

Using motion rather than relying on just visual appearance cues, improves performance in separating musical instrument sounds and helps sound separation of same instrument. In fact, human movements are associated to sound, and help to achieve sound separation. [40] considers to exploit the human keypoints, so human body and hand movements in the videos, while [170] uses visual appearance and optical flow.

2.6.3 Sound Spatialization

The goal of sound spatialization is to get binaural audio, or more in general, spatial audio, from a mono audio, recovering the spatial cues using the visual information. [106] upconverts a single mono recording into spatial audio guided by full 360° view video. The spatial audio is in the form of a popular encoding format called first-order ambisonics. [41] converts common monaural audio into binaural audio by guiding the process with the supervision of spatial information of accompanying visual frame. Visual frames reveal significant spatial cues that are lacking in the accompanying single-channel audio and help to predict left and right channels. They mix the two channels into a single channel so all spatial information collapses. Then, they formulate a self-supervised task to take the mixed monaural signal and its accompanying visual frame as input, and split it into two separate channels, using the original left and right channels as ground-truth during training. Instead of directly predicting the two channels, they predict two channels' difference audio signal. They obtain the complex-valued spectrogram of the difference signal by firstly generating a multiplicative mask and then by complex multiplying the input spectrogram of the mixed mono audio with the predicted complex mask. They train the network using L2 loss to minimize the distance between the ground-truth complex spectrogram and the predicted one. Finally, using ISTFT, they obtain the predicted difference signal, through which they recover the left and right channels, combining the difference with the input mono audio.

[162] proposes a pipeline that is free of binaural recordings. They leverage spherical harmonic decomposition and head-related impulse response to identify the relationship between spatial locations and received binaural audios, generating visually coherent binaural audios without accessing any recorded binaural data. Thus, mono-to-binaural networks can be trained on the created pseudo data.

In our work [128], we train a deep architecture to reconstruct acoustic images, which are spatialized audio, from the associated video frame and its corresponding single microphone audio. Therefore, we introduce a novel audio spatialization task, to predict a more spatialized

audio modality which contains the spectral signature of the sounds associated with each considered direction.

2.6.4 Audio-Visual Localization

Audio-visual localization is the problem of localizing sound sources in visual scenes. Usually, this is done in an unsupervised fashion, learning the correspondence between visual scene and the sound by observing sound and visual scene pairs.

[132, 133] propose a novel unsupervised algorithm, with attention mechanism, which is guided by sound information, with paired sound and video frame. Thus, the sound source localization can be interactive with given sound input: for instance, given a frame that contains water and people, when a water sound is given, the water area is highlighted. Similarly, the area containing people is highlighted when the sound source is from humans.

The model uses frame and sound pairs, processing each modality in its own network with a two-stream network architecture. After integrating (correlating) the information from the sound context vector and the activations of visual network, attention mechanism localizes the sound source. Spatial information is preserved in the visual feature grid, which they make interact with the sound embedding for revealing sound source location information. For each location, the attention mechanism generates a weight, by computing the simple inner product between the given sound embedding and pixel of visual feature map, normalized using the softmax. The operation measures the cosine similarity between two heterogeneous audio visual vectors i.e., correlation. They get a soft confidence score map, where each attention weight can be interpreted as the probability that the visual grid pixel is likely to be the right location related to the sound source. For the unsupervised setting, they impose that audio and visual features from two networks from the corresponding pairs (positive) are close to each other in the feature space, while non-corresponding (negative) pairs are far from each other, using the triplet loss.

Another possibility is to use acoustic based approach. However, this requires specific devices, e.g., microphone arrays, to capture phase differences of sound arrival. For example, the estimated energy from our acoustic images could be used as a ground truth for sound localization. In our work [128], we train a deep architecture to reconstruct acoustic images from the associated video frame and its corresponding single microphone audio and then we perform audio-visual localization exploiting the estimate of energy of sound. Thus, we learn sound source localization in the visual domain without any special devices but just a microphone to capture sound.

Usually state of the art audio-visual localization methods work on simple scenes with only one source. To address multi-source sound localization, [120] proposes a framework which works in 2 steps, to localize multiple sounds and objects. At the first stage, they employ a multi-task framework consisting of classification and audiovisual correspondence. Given audio and visual modalities, we can obtain the pseudo labels from pretrained models as supervision. They employ the same set of classes for both modalities. Considering that there are multiple sound sources contained in the video, multilabel binary cross entropy loss is considered for classification. Regarding audiovisual correspondence learning, it is viewed as a two-class classification problem, i.e., corresponding or not. For multi-task learning, they use the weighted sum of classification loss and correspondence loss. At the second stage, the audio and visual feature maps and classification predictions are fed into Grad-CAM module [131] to disentangle class-specific features on both modalities, then a fine-grained audiovisual alignment is performed. Given the feature map activations of the last convolutional layer and the output of classification, they calculate the class-specific Grad-CAM map [131]. Then they take class-specific map as weights to perform weighted global pooling over the feature map. They obtain the set of audio and visual class-specific feature representations for each video. They use them for fine-grained feature alignment adopting contrastive loss. To visually localize sounds by generating source-aware localization maps, the visual feature map is firstly projected into the shared embedding space, then compared with the disentangled audio features. The obtained value reveals how likely a specific region in the visual scene is the visual source of sound.

[66] localizes objects belonging to different categories when there are multiple sounding objects as well as silent ones in cocktail-party scenarios. They learn robust object representations from single source localization, and aggregate them into a dictionary for each object category. Then they exploit object knowledge for object category aware localization: they reduce the localization task into a self-supervised audiovisual matching problem.

2.6.5 Cross-Modal Generation

Cross-modal generation means recovering the missing modality from the available ones based on the common information shared between them. Audio and visual modalities own both common and complementary information respectively. Common information can make the translation possible.

Some works tried to recover sound from images. The easiest case regards musical instruments, and more in detail, piano. In fact, when playing the piano, hand gestures

are correlated with music. Both [140] and [39] translate the video frames of the keyboard and the musician hand movements playing the piano into music. [140] finds roll for each video frame, corresponding to which keys are pressed and then outputs the pseudo-MIDI signal. [39] extracts key points of the human body from video frames as intermediate visual representations, and thus can explicitly model the body movements. They align body movements to MIDI with a transformer architecture. So in both cases, piano music generation from videos can be posed as a motion to MIDI translation problem. MIDI is transformed to waveforms with an audio synthesizer.

A more difficult case is when there are many sounds in the wild environment. [175] use as input visual and optical flow, encode them to a hidden state and decode it directly to the waveform.

There is not much work from sound to visual, as some visual objects in an image are not correlated to sound. Nevertheless, when we listen to the sounds, we can imagine the visual modality. Music is correlated with visual dynamics, the motion of arms and fingers. [25, 54] can both generate body dynamics of the musicians from the played music. However, there are no works focusing on how to generate images from sounds in the wild. In Chapter 6 we perform cross modal generation with videos collected in the wild, by means of a two stage pipeline employing a PixelCNN to learn a probability distribution of pixels in space conditioned to sound.