Unsupervised Synthetic Acoustic Image Generation for Audio-Visual Scene Understanding

Valentina Sanguineti, Pietro Morerio, *Member, IEEE*, Alessio Del Bue, *Member, IEEE*, and Vittorio Murino, *Fellow, IEEE*

***Abstract*— Acoustic images are an emergent data modality for multimodal scene understanding. Such images have the peculiarity of distinguishing the spectral signature of the sound coming from different directions in space, thus providing a richer information as compared to that derived from single or binaural microphones. However, acoustic images are typically generated by cumbersome and costly microphone arrays which are not as widespread as ordinary microphones. This paper shows that it is still possible to generate acoustic images from off-the-shelf cameras equipped with only a single microphone and how they can be exploited for audio-visual scene understanding. We propose three architectures inspired by Variational Autoencoder, U-Net and adversarial models, and we assess their advantages and drawbacks. Such models are trained to generate spatialized audio by conditioning them to the associated video sequence and its corresponding monaural audio track. Our models are trained using the data collected by a microphone array as ground truth. Thus they learn to mimic the output of an array of microphones in the very same conditions. We assess the quality of the generated acoustic images considering standard generation metrics and different downstream tasks (classification, cross-modal retrieval and sound localization). We also evaluate our proposed models by considering multimodal datasets containing acoustic images, as well as datasets containing just monaural audio signals and RGB video frames. In all of the addressed downstream tasks we obtain notable performances using the generated acoustic data, when compared to the state of the art and to the results obtained using real acoustic images as input.**

***Index Terms*— Deep learning, self-supervised learning, audio-visual systems, spatial audio.**

I. Introduction

Vision and hearing are the most important senses human beings use to explore the world. They are complementary: even if vision is guiding us the most, sound supports us as its propagation is not affected by illumination, camouflaging and occlusions. Moreover, when there is not enough visual evidence, tiny or very far objects such as a plane in the sky or a gunshot in a crowded scene can be detectable only by their sound signature.

More specifically, while interacting with the surrounding environment, vision is supported by binaural hearing, which helps people focus on the sound sources to figure out what is happening around them. Sound signals are received with a certain delay between the left and right ear, as well as a slight difference in intensity, which are critical to perceive spatial cues about the direction of provenience of the sound [1]. Besides, humans associate what they hear with what they see, hence being able to fuse the spatial cues elaborated by their auditory system with those coming from their sight [2].

Binaural microphone configurations have been lately investigated to replicate human auditory capabilities [3], [4], [5], [6]. Similarly, spatial audio was considered to improve robot interactions with the physical environment, as auditory perception can be used to localize sound-emitting targets [7], especially when visual modality is not reliable. Finally, when fused to other multi-sensory data, spatial sound can contribute to the understanding of the physical spaces, as it embeds geometry information. Stereo audio provides a richer informative content about the location of provenience of sound than monaural audio. However, binaural microphone configurations cannot compete with the performance achieved by the human auditory system in localization tasks and are also limited to estimating the direction of sound arrival only along the azimuth direction [8].

This work proposes to reconstruct spatial audio information in the 2D image plane starting from standard monaural audio and video frames and shows how this data can support scene understanding. For this purpose, we exploit the data gathered by a planar array of microphones for training our models. Such device provides more accurate spatial audio information and can localize sound sources much better than stereo audio. In fact, the acoustic signals acquired by an array of microphones can be combined via a beamforming algorithm [9], [10] to produce an *acoustic image*, which is employed to provide information about the sound sources' locations and the acoustic frequency content in a 2-dimensional space (see Figure 1), rather than just along a single direction as binaural audio [4].

Manuscript received 19 May 2021; revised 19 April 2022 and 7 October 2022; accepted 15 October 2022. Date of publication 8 November 2022; date of current version 14 November 2022. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hairong Qi. *(Corresponding author: Pietro Morerio.)*

Valentina Sanguineti is with the Department of Pattern Analysis & Computer Vision (PAVIS), Istituto Italiano di Tecnologia, 16163 Genova, Italy, and also with the with DITEN, University of Genova, 16126 Genova, Italy (e-mail: valentina.sanguineti@fastwebnet.it).

Pietro Morerio and Alessio Del Bue are with the Department of Pattern Analysis & Computer Vision (PAVIS), Istituto Italiano di Tecnologia, 16163 Genova, Italy (e-mail: pietro.morerio@iit.it; alessio.delbue@iit.it).

Vittorio Murino is with the Department of Pattern Analysis & Computer Vision (PAVIS), Istituto Italiano di Tecnologia, 16163 Genova, Italy, and also with the Department of Computer Science, University of Verona, 37129 Verona, Italy (e-mail: vittorio.murino@iit.it).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2022.3219228, provided by the authors.

Digital Object Identifier 10.1109/TIP.2022.3219228

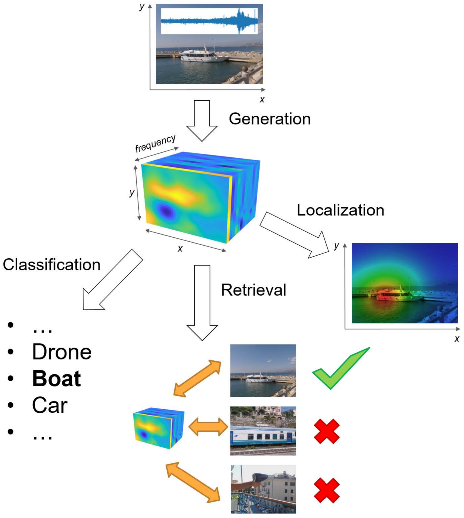

Fig. 1. We generate a spatialized audio frequency map called *acoustic image*. Starting from an RGB frame and the corresponding monaural audio (top), we synthesize the sound frequency distribution in each direction in space and associate it to each pixel in the acoustic image (middle). Then, we use it for different downstream tasks: classification, cross-modal retrieval and unsupervised sound source localization (bottom, left to right).

Acoustic images have proven their usefulness for scene understanding when used in both supervised [11] and self-supervised learning [12] scenarios because of their peculiar spatial distribution. Moreover, this spatial content can be distilled to audio models to get more robust features which generalize better to new domains (i.e., datasets) than monaural features [11], [12]. They have also been used to tackle audio tracking applications [13], [14], where acoustic information plays a fundamental role when visual counterpart is cluttered, occluded, or noisy.

Unfortunately, microphone arrays are needed to acquire acoustic images, which are expensive, cumbersome, and not so commonly available. Therefore, we propose to generate such spatialized audio data from a standard video sequence, i.e., from a single-microphone audio signal enriched with the visual content of the scene. In this way, even without an array of microphones, we can synthesize acoustic images, which are then evaluated on many different audio-visual tasks, such as classification, cross-modal retrieval and audio-visual localization.

To this end, we introduce three generative architectures, specifically designed for acoustic image reconstruction conditioned on certain single microphone audio and specific video frame, while exploring their peculiarities. The first model is a hybrid architecture based on Variational Autoencoder (VAE) [15] and U-Net [16], which is designed to exploit the upsides of both [17]. VAEs are very effective generative architectures and have a principled mathematical grounding, but they show limitations when the size of the output is too large. On the contrary, U-Nets are reconstruction models that can effectively deal with the details of high-resolution images, but they do not include any stochastic element which can regularize them. The second proposed model is a variant of the former, exploring its adversarial counterpart by adding a discriminator on top. Finally, the third proposed model is a multimodal conditional generative adversarial network (GAN) [18], which is suitable for reconstructing acoustic images conditioned on multimodal input. An advantage of adversarial models is their ability to focus more on the overall semantic content rather than on local areas. All generators are conditioned to a multimodal input made up of an audio signal and an optical image, corresponding to the acoustic image to be generated, in order to preserve not only the sound information, but also the spatial correspondence between the generated acoustic image and the visual frame.

We evaluate the synthetic data generated by the proposed models by common generation metrics, showing that they are similar to real acoustic images in terms of mean square error and semantic content. Our generated acoustic images obtain remarkable results on downstream classification and cross-modal retrieval, and also have a good transfer learning capability. In fact, our models perform better than previous state-of-the-art networks in localization task on new datasets.

To recap, the contributions of our work can be summarized as follows:

1) Driven by the evidence that spatially distributed acoustic data constitutes a richer information source than monaural audio signals, we propose three different multimodal models based on Variational Autoencoder, U-Net and GAN to tackle a new audio spatialization task, the reconstruction of acoustic images from RGB images and single microphone audio *without* the use of an array of microphones at test time.
2) Our new spatialization task allows us to carry out audio-visual localization in a novel fashion, i.e., by estimating the energy of the synthetic spatialized sound. Moreover, using ground-truth acoustic images, we do not need human annotations because their energy is sufficient to show the actual regions where the sound is coming from. Interestingly, our method demonstrates good transfer learning capabilities, i.e., it generalizes better than previous works in localization tasks on datasets never seen in training.
3) We present a set of experiments to evaluate the quality of the generated data in terms of reconstruction error, classification, cross-modal retrieval and localization performances.

This journal extends the previous conference publication [17] where a single architecture was proposed for audio spatialization and audio-visual localization.

The rest of the paper is organized as follows. Section II reviews the related works and highlights the original aspects of our proposed approach. Section III presents the generative architectures and their training strategies. Section IV shows the different downstream tasks used to evaluate the generated synthetic acoustic images. Section V reports the experiments and ablation studies and, finally, Section VI draws conclusions and future work.

II. Related Works

In this section, we review the previous works related to the main topics addressed by our paper, i.e. audio spatialization, image generation/translation models, and audio-visual representation learning for different downstream tasks such as audio-visual localization, cross-modal retrieval and classification.

A. Sound Spatialization

The goal of video-based *sound spatialization* is to upgrade a single mono audio recording into spatial audio, for example, binaural audio [4], [5] or first order ambisonics audio [19] guided by the visual information that provides spatial cues that are missing in the single channel audio. More in detail, [19] generates audio for the full viewing sphere given 360° video, while [4], [5] use standard video. These methods employ for the audio stream a U-Net architecture conditioned to the feature maps coming from a visual stream to predict a set of spectrogram complex masks, one for each channel. They multiply each of the masks by the input mono audio spectrogram to reconstruct spatial audio, using the original spatial audio as target during the training. We draw inspiration from these works to propose a novel and challenging audio spatialization task: reconstructing the spectral signature of the sounds associated with each considered direction, namely each acoustic pixel in the acoustic image. We are not aware of any work trying to recover such spatialized sound information from monaural microphone and RGB frames only.

B. Image Generation and Translation Models

Image generation and translation tasks are usually solved using VAE and GAN models as they provide different upsides: a VAE is trained with a reconstruction loss, thus being suitable for minimizing pixel-by-pixel loss, while GAN objective preserves better the overall semantic meaning of the sample. We employ both VAE and GAN architectures to analyse how their peculiarities impact the performance.

1) Cross-Modal VAE: One of the first works addressing the problem of translating one modality to another one is [20], which reconstructs both audio and video from video or audio respectively, only using an autoencoder. [21], [22] instead model the joint representation of all the modalities and can generate one modality from a joint latent variable. Other works [23], [24] also find a common shared latent space from single-modality latent variables. Finally, [25] proposed to use different VAEs for each modality and translated the latent variable of one modality into the latent variable of another one using an "associator".

Our proposed VAE model differs from the above in two ways:

1) it has a U-Net structure to better deal with details so that reconstruction is performed not only based on the VAE latent variable, but also on the intermediate feature maps that retain spatial cues;
2) the latent space is not constructed from all modalities, but only from those available at test time (RGB and monaural audio).

In [17], we introduced U-VAE, a baseline to which we compare our proposed adversarial models. In addition, while in [17] we tackled only audio-visual localization, we now evaluate the generated samples on different downstream tasks. More specifically, we show that the synthesized samples from a dataset never seen during training (downloaded from the Internet) are performing well for such tasks. We perform a broader experimental validation and extend the analysis by including the impact of losses and of different visual model.

2) Image Transformation: This task typically aims at transforming an input image into an output image with different properties. Examples include denoising, super-resolution [26], [27], style transfer [28], [29], [30], colorization [31], [32], semantic segmentation [33], depth estimation [34], domain adaptation [35], and multimodal generation (e.g., from RGB to depth) [36].

For image translation, [37] proposed to use a conditional GAN made up of U-Net and a discriminator that learns to perform binary classification between fake and real input-output tuples. In conditional generative adversarial networks (cGANs), both generator and discriminator are conditioned to additional information in the generation process [38], rather than on the noise vector only. Such conditioning can be based on any auxiliary information such as on labels, on some part of data (e.g. for inpainting), or even on data from a different modality, such as in [39] and in our work. Image-to-image translation uses a training set of aligned image pairs. When paired training data is not available, cycle-GAN [40] can be employed to train a GAN using cycle consistency loss.

Aligned audio signals, video frames and acoustic images are available in our training dataset. Thus, we draw inspiration from [37] for our adversarial models. Differently from these works, our task is more difficult as our input is multimodal (made up of a single microphone and visual features) and a different output modality, which is the target acoustic image.

C. Audio-Visual Learning and Downstream Tasks

We validate our generative models considering three downstream tasks where we can employ synthetic acoustic images to learn good audio-visual representations: audio-visual localization, classification, and cross-modal retrieval.

1) Audio-Visual Learning: Lately, there has been an increased interest in multimodal learning of auditory and visual signals to obtain better representations than those learned by single modalities. Early approaches trained a double-stream neural network by using the prediction of the stream of one modality as a supervisory signal for the other one [41], [42], [43], [44], [45]. Other works [46], [47], [48], [49], [50] train both visual and audio networks aiming at learning multimodal representations useful for many applications, such as cross-modal retrieval, sound source localization and on/off-screen audio source separation. [46], [47] learn aligned audio-visual representations using an audio-visual correspondence pretext task. The common factor in all these audio-visual works is that they exploit the natural *temporal* synchronization between auditory signal and visual images as a rich supervisory self-training signal to train the several models using self-supervised learning. Some works instead explore the *spatial alignment* between stereo auditory signal and visual images [4]. Furthermore, the intrinsic *temporal synchronization*, but also the *spatial alignment* of visual and acoustic images can be exploited as a supervisory signal. By using knowledge distillation and acoustic images, [12] forces audio-visual agreement between feature maps to find aligned shared representations.

2) Audio-Visual Localization: Some of the earliest works about sound source localization are grounded on the natural synchrony between audio and visual signals [51], [52]. Many recent approaches for sound localization exploit two-stream deep network architectures to find the correlation between the sound and visual feature representations [47], [50], [57], [58], [53], [54], [55], [56]. Other works also aggregate temporal information with LSTMs [59], [60] or optical flow [61]. Other methods [45], [49], [62] apply the class activation map (CAM) [63] for sound localization.

Audio-visual localization works usually leverage the temporal correspondence between a visual object and the corresponding sound. This supervision might not be so reliable as not focusing solely on the region where the sound was originated from, but often from the entire object or the whole image.

Differently, we propose to perform sound source localization by first generating acoustic images using a model trained with the richer supervision provided by the real acoustic images and subsequently, by extracting their energy. Our models are not highlighting the whole object but, more accurately, just the regions from which the sound is originating. This is due to the ground-truth, real acoustic images, which contain the actual distribution of the sound energy in the 2D space.

3) Cross-Modal Retrieval: In [12], a self-supervised learning approach for audio-visual representations with acoustic images is proposed, which is evaluated in the cross-modal retrieval task. We propose here another method for audio-visual self-supervised learning with acoustic images, showing to obtain better results in cross-modal retrieval with respect to [12], also when using generated samples. We use separate acoustic and video features rather than using audio and audio-visual features and perform both audio-to-video and video-to-audio retrieval.

4) Acoustic Image Classification: Supervised deep learning has been applied to the field of acoustic imaging by [11], which proposed an architecture able to classify acoustic images in a supervised way. Here, we consider the same network [11] to evaluate our generated samples on the downstream classification task by comparing their classification accuracy with the one obtained by real samples.

Our work presents a novel audio spatialization task, more challenging than those present in the state of the art [4], [5]. For solving this task, we designed a VAE architecture that exploits a U-Net framework, differently from cross-modal VAEs addressed in the literature. We also propose a cGAN that gets inspiration from [37] with the main difference that it can be conditioned to a multimodal input (single microphone audio and video frame), different in modality from the output target acoustic image.

We show that the generated samples allow to reach performance close to real samples on different downstream tasks outperforming the results obtained by the state of the art on the same tasks.

III. The Method

In this section, we first present the type of input to feed our proposed models, followed by their design and the related training strategies.

A. Input Data

The monaural audio track and the associated video compose the input data to our models, while supervision is provided by the acoustic images.

The latter are generated from a planar array of microphones and are volumes of size $36 \times 48 \times 512$, where 36×48 is the image size and 512 is the number of the frequency bins (see Supplementary Materials for a detailed description). The acoustic images are compressed along the frequency axis using 13 Mel-Frequency Cepstral Coefficients (MFCC), which consider human audio perception characteristics [64]. Here, we discard the first MFCC coefficient, which represents average log energy of the input signal, as it carries little sound discriminant information [65]. Thus, they have just 12 channels, preserving most of the information but consistently reducing the computational complexity and the required memory. When we reconstruct acoustic images, we consider the time interval $1/12s$ for all modalities (RGB frames, spatialized audio, and single microphone audio), as the ground-truth acoustic image and the RGB image frame rate is just 12 frames/s (fps). Starting from audio samples correspondent to $1/12s$ we compute a vector of 12 MFCC. To simplify the generation task, rather than using raw waveforms or spectrograms, we consider homogeneous input-output, feeding the models with MFCC of a single microphone tiled along spatial dimensions to obtain a $36 \times 48 \times 12$ map, which has the same dimensions of an acoustic image. The inputs to the network are then the tiled MFCC and an RGB frame, while the output consists in the reconstruction of the acoustic image.

This allows a quasi-real time estimate of the directional sound, every $1/12s$. Previous works, instead, considered audio signals lasting from 1 s [60] to 20 s [50], and only one frame sampled from the video clip in the middle of that soundtrack to visually localize the sound. However, this can cause to miss important synchronization cues between audio and video, which instead we are taking into account in our work.

B. Reconstruction Models

We propose three different architectures specifically designed to generate acoustic images resembling those produced by combining the signals acquired by a planar array of microphones. We consider as input single-microphone features and condition our networks on visual information in order to provide spatial cues which are missing in the omnidirectional microphones. The architectures, shown in Figure 2, are:

1) A VAE with skip connections as in U-Net model [16], named U-VAE, whose loss reconstructs each pixel separately, which is useful for localization.

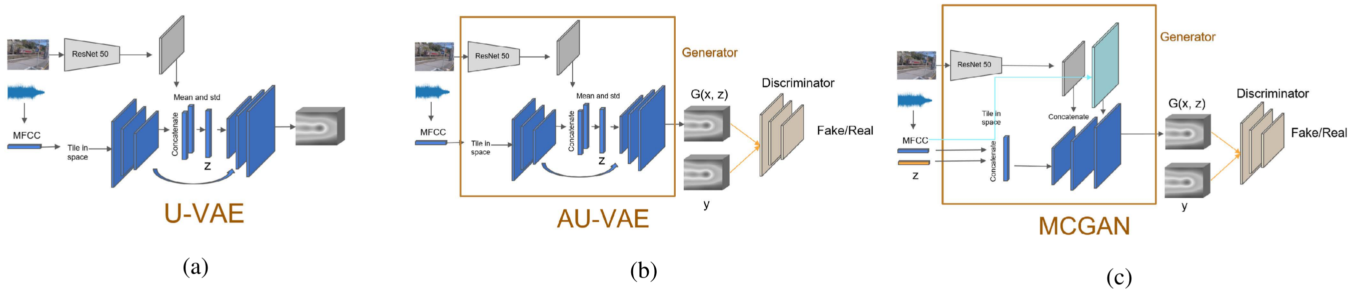

Fig. 2. The three proposed architectures. (a) U-VAE is based on VAE and U-Net to generate acoustic images. (b) AU-VAE (adversarial U-VAE) uses U-VAE as a generator and adds a discriminator on top. (c) MCGAN (Multimodal Conditional GAN) is so named since it has a multimodal conditional input suitable for image translation. For all the models, input data are monaural audio samples (represented as compressed MFCC coefficients) and ResNet50 visual features.

2) A model based on the previous network by adding a discriminator on top, named AU-VAE (Adversarial U-VAE).
3) A network named MCGAN (Multimodal Conditional GAN), conditioned on multimodal input and which generates new samples from a random noise vector z to obtain more (random) variability in the generated samples.

We hypothesize that introducing an adversarial loss helps preserving the overall semantic content of spatial audio for downstream tasks such as classification and cross-modal retrieval.

Visual features are extracted using ResNet50 [66] pre-trained on ImageNet [67] and modified with the removal of global average pooling and the addition of a 2D convolution layer to get a feature map. We train the last ResNet50 layer only to focus on the specific regions producing sound in the considered training datasets. The visual feature map is then concatenated to the generator's feature maps as shown in Figure 2. The audio and visual feature maps come from two different streams and can have different ranges of values, thus we normalized them before concatenation. More details about the proposed models and training hyperparameters are reported in the Supplementary Materials.

C. U-VAE Model

The U-VAE architecture is depicted in Figure 2a. It is composed by encoder and decoder modules: the former is taking as input the audio MFCC features and the video feature map, while the latter is fed by the produced bottleneck and aims at reconstructing the final acoustic image with the aid of skip connections. We chose a VAE as a base model rather than a simple autoencoder because it can improve reconstruction due to the latent loss regularization [68].

Assuming that the latent variable distribution $p(\mathbf{z})$ is a centered isotropic multivariate Gaussian, and that the inferred posterior distribution is a multivariate Gaussian with diagonal covariance, we have that $p(\mathbf{z}) = \mathcal{N}(0, \mathbf{I})$ and $q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}(\mathbf{x}), \mathrm{diag}(\boldsymbol{\sigma}^2(\mathbf{x}))$, where $\mathbf{x}$ is the input. The encoder outputs two vectors, mean and standard deviation $\boldsymbol{\mu}(\mathbf{x})$, $\boldsymbol{\sigma}(\mathbf{x}) \in \mathbb{R}^d$, where d is the dimensionality of the latent space.

We can sample $\mathbf{z}$ from $q(\mathbf{z}|\mathbf{x})$ with the re-parameterization trick [15]: first, we sample a random vector $\mathbf{u}$ from a unit Gaussian $\mathcal{N}(0, \mathbf{I})$, and then we multiply it by the standard deviation $\boldsymbol{\sigma}(\mathbf{x})$ while adding the mean $\boldsymbol{\mu}(\mathbf{x})$ giving:

$$\mathbf{z} = \boldsymbol{\mu}(\mathbf{x}) + \boldsymbol{\sigma}(\mathbf{x}) \odot \mathbf{u}, \qquad (1)$$

where $\odot$ denotes the element-wise product. VAEs are trained to maximize the log probability of the likelihood of generating data similar to real ones $p(\mathbf{x})$, maximizing the Evidence Lower Bound (ELBO), given by:

$$ELBO = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] - \beta KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})). \qquad (2)$$

The first addendum of ELBO can be interpreted as a reconstruction loss (with a minus sign). The second term is the positive Kullback-Leibler divergence and represents the latent loss. In standard VAE formulation, the hyperparameter β in Eq. (2) is fixed to one. However, [69] proposed that β can be tuned as an adjustable hyperparameter to balance the two terms, as the first term represents the reconstruction accuracy and the second one regards latent independence constraint. In particular, [69] proposed to consider $\beta > 1$ for good disentangled representations. Instead, we are more interested in obtaining good reconstruction than good latent variables, so we weigh the latent loss using $\beta < 1$. Specifically, we set β to get the same order of magnitude of reconstruction and latent losses so that the network can achieve good reconstruction quality.

The choice of a hybrid model based on VAE and U-Net is derived from the fact that VAE is a very effective generative tool, but its outcome is usually not so finely detailed, while U-Nets are models that can effectively deal with the small details thanks to their skip connections. Our ablation study reported later in Subsection V-F will confirm such findings.

D. AU-VAE and MCGAN Models

Past works [37], [70] proposed to translate an input image into a corresponding output image mapping pixels to pixels using as loss a weighted sum of GAN objective and regression losses. In fact, a regression loss (such as in VAEs and autoencoders) minimizes the mean pixel-wise error, but it results in a blurry averaged image as each output pixel is independent from all the others. Instead, adding an adversarial loss penalizes the joint configuration of the output, obtaining sharp and realistic images while making the output indistinguishable from the real one.

In this way the generative model has to generate output near the ground-truth target and then to fool the discriminator by optimizing the following objective function:

$$G^* = \arg\min_G \max_D \lambda_{adv}\mathcal{L}_{cGAN}(G, D) + \lambda_{rec}\mathcal{L}_{L2}(G). \quad (3)$$

We propose two adversarial models: a standard GAN architecture (called MCGAN) and one which employs a U-VAE generator (AU-VAE). The choice of the latter architecture is adequate to carry the low-level information shared between the input and output across the network through the skip connections [37]. Both models employ the same network architecture for the discriminator, DualCamNet [11], designed for acoustic image classification, while determining whether the generated acoustic image is true or fake.

More in detail, we use conditional GANs (cGANs). Differently from standard GANs [18], which learn a mapping $G : z \rightarrow y$ from a noise vector z to an image y, cGANS can be obtained by adding x as input to both G and D and learning a mapping from z and input x to output y, i.e. $G : \{z, x\} \rightarrow y$. The input x to which we condition the two generators has a multimodal nature as it is composed of audio-visual features used at different stages of the models. We condition only the generators to the input while the discriminator observes just real and fake outputs and must distinguish them. The loss of our conditional GANs $\mathcal{L}_{cGAN}$ is given by:

$$\mathcal{L}_{\text{cGAN}}(G, D) = \mathbb{E}_y[\log D(y)] + \mathbb{E}_{x,z}[\log(1D(G(x, z))], \quad (4)$$

which should be minimized by G against an adversarial D that maximizes it, i.e., $G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D)$. To implement this loss, we train D to maximize the probability of classifying the real image as label 1 and the generated image as 0 using binary cross entropy loss, while the generator G is trying to fool D to classify the generated image as real (with label 1).

IV. Downstream Task

Using the proposed architectures we address three different downstream tasks to show the goodness of our approach in several scenarios: synthetic acoustic image classification, cross-modal retrieval and sound source localization.

A. Classification of Synthetic Acoustic Images

We evaluated the classification of generated samples on both acoustic image datasets and an unseen dataset containing standard videos (i.e., visual data and associated monaural audio streams).

More specifically, we trained our proposed generators on datasets containing acoustic images, and then we used them to synthesize acoustic images from standard videos both for the training and the test set. Then, we trained a classifier on synthetic acoustic images belonging to the training set and we tested it using synthetic acoustic images of test set.

B. Cross-Modal Retrieval

We evaluate whether generated acoustic images samples are useful for audio-video retrieval. For this purpose, we consider ResNet50 as feature extractor for the visual modality stream, DualCamNet for extracting acoustic image features and VGGish [71] as audio model for processing single-microphone audio data. The input data are $1s$ of audio signal and the associated acoustic images in that interval while considering one visual frame still taken in that interval.

[12] performed cross-modal retrieval using acoustic images for the first time, from acoustic to visual modality through audio and audio-visual features. It employed acoustic and visual stream networks trained with triplet loss. The visual stream model was ResNet18, trained from scratch. Since such video modality is quite hard to classify in the considered datasets it was employed an audio attention mechanism to improve the visual features. Hence, the output of the method were audio and audio-visual $12 \times 16 \times 128$ volumes, which were reduced to 128D vectors performing a sum along the spatial dimensions in order to feed the triplet loss with vectors of a reasonable dimension.

Differently, we use here separate video and audio features, making it possible to do symmetric retrieval (also from visual to acoustic modality), getting rid of the audio attention mechanism. In fact, another difference is that we consider ResNet50 pre-trained on ImageNet. This allows us to get more powerful video features that do not need audio guidance. As in [12], we train video and acoustic image stream networks with the help of the triplet loss. However, in our case, we consider $12 \times 16 \times 12$ feature map volumes, limiting the number of channels to 12, while [12] employed 128 channels. Therefore, we do not need a spatial sum to reduce the dimensionality of the feature map and we just flatten the feature volumes to vectors to be used in the triplet loss, hence preserving in a better way the spatial content even if we reduce the channels. Since our cross-modal retrieval has a better performance, we speculate that spatial information is more important than that provided by using more channels.

We cannot have a corresponding spatial feature map for the VGGish model as a single microphone does not contain spatial information, but we consider an embedding from this network with the same size as the flattened feature maps for video and acoustic samples.

We train audio with video stream nets in pairs, to compare retrieval results using different acoustic input data: spectrograms, real acoustic images, generated acoustic images, single-microphone MFCC coefficients, as displayed in Figure 3.

To train our models for cross-modal retrieval, we minimize a triplet loss [72]:

$$\begin{aligned}\mathcal{L}_{XY}^{triplet}(x_i^a, y_i^p, y_i^n) &= \sum_{i=1}^{N}[||f_X(x_i^a) - f_Y(y_i^p)||_2^2 \\ &+ \quad -||f_X(x_i^a) - f_Y(y_i^n)||_2^2 + m]_+, \quad (5)\end{aligned}$$

where $[h(x)]_+$ represents $\max(h(x), 0)$, m is a margin between positive and negative pairs fixed at $m = 1.0$, $f(x_i)$ are the normalized feature vectors, x_i is an audio sample, y_i is a video sample, f_X is the audio model, f_Y is the video model, and a, p, n stand for anchor, positive and negative, respectively. The triplet loss aims to separate the positive pairs from the negative ones by a distance margin. It minimizes the distance

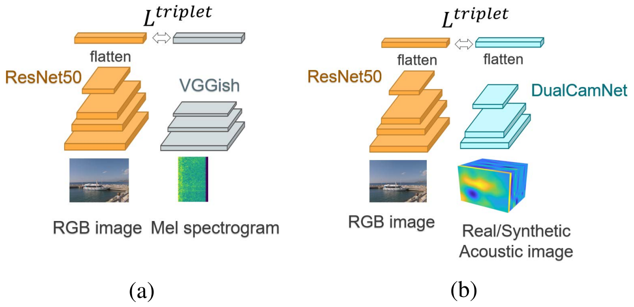

Fig. 3. We perform cross-modal retrieval from audio to video and vice versa using video and: (a) Mel spectrogram, (b) real acoustic image or synthetic acoustic image (or replicated MFCC).

between an anchor a and a positive, p, both with the same identity, and maximizes 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 a video embeddings $f_Y(y_i^p)$ coming from the same video clip, and a large one from video embeddings $f_Y(y_i^n)$ obtained from a different clip.

C. Energy of Sound Approximation for Localization

Differently from previous works that considered audio-visual correlation as supervision, we perform localization with more precise supervision, consisting in the spatial sound distribution provided by the acoustic images. The energy of the synthesized acoustic images implicitly localizes the sound sources, which can then be computed from the MFCC representations. Then, we verify that the energy estimates for real and synthetic acoustic images are similar.

To understand how the sound energy is computed, we must remind that the extraction of MFCC coefficients from an audio signal is performed by a Discrete Cosine Transform (DCT) [73] applied to log Mel filters' [74] energies.

Actually, summing up Mel filters' energies is a good approximation of the sound energy and that is why we compute the inverse DCT (IDCT) of MFCC coefficients to recover such energy. Then, we performed the exponential of the estimated log Mel filters' coefficients to recover original energies and we summed them up for each acoustic pixel. However, due to the absence of the first MFCC coefficient, our estimate is not precise and we obtain a map inversely proportional to the real energy. Computing its reciprocal allows to obtain an estimation of the energy distribution over the scene and to localize the sound sources present therein.

V. Experiments

In this section, we validate the proposed models in terms of reconstruction quality and of the three downstream tasks, namely classification, cross-modal retrieval and localization. We first describe the datasets and after that, we assess the quality of reconstruction of our proposed networks by comparing their outputs to the ground-truth acoustic images. We adopt the standard reconstruction error (Mean Square Error, MSE) and also include two additional metrics measuring the faithfulness and diversity of the generated samples (through classification). We then consider synthesized acoustic images for classification and cross-modal retrieval tasks evaluating them on a dataset with acoustic images and on unseen data. Finally, we evaluate audio-visual localization performance, both quantitatively and qualitatively, on datasets containing acoustic images and videos collected from the Internet.

The different metrics we employ verify both structure and meaning of the synthetic samples. On the one hand, MSE and audio-visual localization measure how good the pixel-by-pixel reconstruction of the acoustic image is. On the other hand, we evaluate semantic meaning through classification and cross-modal retrieval tasks. In this way, we assess whether the generated samples are similar to real ones and preserve their semantic information.

A. Datasets

We consider the following four datasets to test our models:

- ACIVW [12] is an audio-visual dataset including acoustic images, single microphone audio and images containing 5 hours of videos acquired in the wild, containing 10 classes.
- AVIA [11] is an audio-visual dataset including acoustic images, single microphone audio and images with 14 different actions producing a characteristic sound performed by 9 people in 3 different scenarios with increasing and varying noise conditions.
- A random subset of Flickr-SoundNet [41] employed by [50], which includes sounds sources positions annotated by three subjects, facilitating quantitative evaluation. We consider just the testing data, which includes 250 pairs of frames and their corresponding sound.
- VGGSound [75] is a dataset with over 200k 10s video clips containing an object making sound for 300 audio classes from YouTube videos.

We use the first two datasets for both training (since they contain acoustic images needed as ground truth) and testing. The remaining two are instead used for testing only, to evaluate the generalization capability of our generative models on unseen domains. The Flickr-SoundNet dataset does not include any class-specific information, but provides the locations of sound sources (bounding boxes). We thus employ it only for the audio-visual localization task. VGGSound, on the other hand, does not contain any annotation of sound location, but provides class labels: we employed it for qualitative evaluation of audio-visual localization and for quantitative evaluation of classification and cross-modal retrieval of generated acoustic images. In particular, we considered a subset of VGGSound, i.e., the samples with the classes most similar to those included in ACIVW.

B. Generation Metrics

Differently from the evaluation of synthetic RGB images, we cannot visually assess the quality of acoustic images. We instead need to evaluate if they preserve their frequency content. This is done by using the following metrics:

TABLE I

Generation Metrics for ACIVW and AVIA Models. MSE Values Are Multiplied by 10^{-2}. Test Data Refers To: Real Acoustic Images, Generated Acoustic Images, Tiled MFCC From Single Microphone. For the Description of GAN-Train and GAN-Test, Refer to Sec. V-B

	Test data	**ACIVW**			**AVIA**		
		U-VAE	AU-VAE	MCGAN	U-VAE	AU-VAE	MCGAN
MSE	-	1.1426±0.0053	1.2243±0.0076	1.3463±0.0142	0.9483±0.0026	1.1248±0.0118	1.1754±0.0040
GAN-test	real	0.8497±0.0014			0.8383±0.0022		
	generated	0.8342±0.0093	0.8415±0.0166	0.8068±0.0025	0.6700±0.0009	0.7576±0.0079	0.7133±0.0046
	MFCC	0.5410±0.0175			0.2091±0.0027		
GAN-train (on gen.)	generated	0.8512±0.0089	0.8495±0.0150	0.8175±0.0419	0.7871±0.0039	0.8766±0.0087	0.9096±0.0087
	real	0.7661±0.0065	0.8196±0.0185	0.8147±0.0238	0.6456±0.0100	0.6982±0.0038	0.6117±0.0090
GAN-train (on MFCC)	MFCC	0.7323±0.0072			0.6614±0.0038		
	real	0.4270±0.0186			0.1307±0.0119		

- Mean square error (MSE), to measure the reconstruction error for each acoustic pixel.
- GAN-test [76], to measure the accuracy of a classifier trained on real acoustic images but evaluated on generated images (and possibly on real ones) to quantify semantic similarity to real samples.
- GAN-train [76], to measure the accuracy of a classifier trained on generated data and evaluated on real test images (and possibily on generated ones). The GAN-train metric captures the diversity of generated samples.

For GAN-test and GAN-train, we consider the DualCamNet network introduced in [11] as a standard classifier for acoustic images. We classify sequences of 1s, which means 12 acoustic images at a time, provided that the rate is 12 fps.

Using the aforementioned metrics, we evaluate the reconstruction quality for all the three proposed generation models on both the test sets of ACIVW and AVIA datasets as they include ground truth acoustic images. Results are provided in Table I.

1) MSE: Regarding MSE, we see that the lowest pixel-by-pixel error is obtained without using a GAN loss. GANs in general are more focused on the content of the entire sample rather than on pixel reconstruction. AU-VAE has lower MSE than MCGAN because of its U-Net architecture, which helps in reconstructing the details. We observe a similar increase in MSE using adversarial losses on both AVIA and ACIVW datasets.

2) GAN-Test: We train the DualCamNet classifier on real acoustic images from the training sets of either AVIA or ACIVW and we compare its accuracy when tested on both real acoustic images and generated ones. For both datasets we observe a reasonable drop (when testing on generated acoustic images), however we see that training on the ACIVW dataset gives a smaller drop than using AVIA as its acoustic images were collected in more noisy (i.e., realistic) scenarios and contain periodic sounds. While the GAN test on ACIVW is similar for the three models (the biggest gap is 3%), we notice that on AVIA, MCGAN increases the performance of 4% and AU-VAE of 9% with respect to U-VAE. **MFCC** refers to using a synthetic acoustic image generated by replicating single-microphone MFCC along the 2 spatial dimensions. This is done to assess the quality of a single-microphone representation against the generated acoustic images, where MFCC coefficients are modulated by the visual image content through the proposed architectures, instead of being merely replicated. We notice that the drop in accuracy is considerable: 30% for ACIVW and 63% for AVIA, showing that our architectures are essential to generate different MFCC for each acoustic pixel, namely to modulate sound in space.

3) GAN-Train: We train DualCamNet on synthetic acoustic images and compare its accuracy when testing on generated and on real acoustic images. On the ACIVW dataset we have a smaller drop when testing on real samples than when testing on AVIA. More precisely on ACIVW the lower drops are obtained with adversarial models, which verifies our assumption that they capture better than the U-VAE the semantic meaning of the overall sample. Testing on generated data (GAN-train tested on generated) we have good classification accuracy for both datasets. On AVIA the highest results on generated data are obtained by adversarial models. Last, we train DualCamNet on uniform acoustic images artificially created by replicating MFCC from a single microphone and test it on real acoustic images. This experiment is designed to show how our generators are actually modulating MFCC for each spatial direction. Furthermore, when both training and testing on replicated single microphone MFCC (GAN-train on MFCC tested on MFCC), we get worse performance than when training and testing with acoustic images (GAN-test on real), proving that spatialized audio allows increasing classification accuracy.

C. Classification

In Table II, we show results for acoustic image classification on VGGSound. This dataset does not include acoustic images, which are indeed generated by using the models we trained on ACIVW.

VGGSound videos are collected from YouTube and can be noisy both regarding aural and visual modality. In fact, using pretrained models we gain 25% accuracy for both audio classification (we finetune VGGish pretrained on YouTube-8M) and video classification (finetuning ResNet50 pretrained on ImageNet).

If we consider audio classification, by training VGGish and DualCamNet from scratch to classify respectively spectrograms and tiled MFCC, we get lower accuracy on

TABLE II

CLASSIFICATION ACCURACY FOR VGGSOUND. WE PROVIDE RESULTS FOR ALL DIFFERENT MODALITIES, INCLUDING GENERATED ACOUSTIC IMAGES

Modality	Model	Accuracy
Audio	VGGish	0.3817±0.0593
	VGGish pretrained	0.6381±0.0003
	MFCC	0.3125±0.0011
Video	ResNet50 pretrained	0.7321±0.0034
	ResNet50	0.4864 ±0.0244
Acoustic Images	U-VAE	0.3540±0.0157
	AU-VAE	0.3762±0.0032
	MCGAN	0.4234±0.0156
	U-VAE resnet triplet	0.5418±0.0126
	AU-VAE resnet triplet	0.5163±0.0099
	MCGAN resnet triplet	0.6040±0.0128
	U-VAE resnet class	0.5903±0.0190
	AU-VAE resnet class	0.5560±0.0164
	MCGAN resnet class	0.6767±0.0049

TABLE III

CROSS-MODAL RETRIEVAL FROM ACOUSTIC MODALITY TO VIDEO (A-V) AND FROM VIDEO TO ACOUSTIC MODALITY (V-A). ON THE ROWS WE LIST ALL THE CONSIDERED ACOUSTIC MODALITIES

	ACIVW		VGG Sound	
	A-V	V-A	A-V	V-A
A. Im. [12]	33.41±3.65	-	-	-
A. Im.	74.35±0.96	66.67±0.92	-	-
U-VAE	71.85±1.55	68.75±0.29	32.77±0.21	35.20±0.26
AU-VAE	71.68±2.02	70.28±1.40	31.48±0.78	31.90±0.75
MCGAN	71.42±0.22	71.74±2.52	41.45±0.85	49.52±0.99
MFCC	58.27±2.31	41.39±1.59	28.30±0.34	22.99±0.61
Audio	69.11±1.61	63.76±2.57	43.52±0.29	46.70±0.55
Audio [12]	28.95±2.15	-	-	-
U-VAE tr	-	-	35.77±0.90	33.21±1.57
AU-VAE tr	-	-	35.05±1.65	33.41±1.52
MCGAN tr	-	-	61.54±1.15	67.82±1.31
U-VAE cl	-	-	31.98±1.00	30.35±1.41
AU-VAE cl	-	-	32.86±1.87	31.14±1.69
MCGAN cl	-	-	60.71±0.44	69.62±0.70

MFCC because they are much more compressed than a log Mel spectrogram. Even if acoustic images are generated from compressed MFCC, we obtain a higher performance than single microphone MFCC, which suggests that our generators are able to recover the spatial cues missed by omnidirectional audio. In addition, we notice that the best accuracy is obtained by the MCGAN model, which is preserving at best the semantic information: it is improving U-VAE performance by 7%, while AU-VAE shows just a 2% increase.

However, there is a gap in performance for synthesized samples due to visual domain differences. This can be mitigated by replacing the visual stream of the generator with one trained on the same data in a self-supervised or supervised way. In this way, synthetic acoustic images perform better than spectrograms. Thus, we substituted the original concatenated visual features with those of ResNet50 finetuned on VGGSound, resulting in an increase in accuracy of about 20%. Specifically, we consider both ResNet50 finetuned on classification task (‘resnet class’) or self-supervised learning of audio-visual correspondence pretext task (‘resnet triplet’) (see Table II).

D. Cross-Modal Retrieval

The goal of cross-modal retrieval consists in choosing one sample from one modality and searching for the corresponding samples of the same class from the other modality. We perform cross-modal retrieval from audio to visual images and vice versa on ACIVW and VGGsound datasets. In our experiments, audio samples can come in the form of an acoustic image, tiled MFCC, or spectrogram.

As an evaluation metric, we employ the Cumulative Matching Characteristic (CMC). Given an audio sample, corresponding visual embeddings are ranked based on their distance in the common feature space. Rank K retrieval performance indicates if at least one sample of the correct class falls in the top K neighbors. We consider CMC scores for the most difficult case, $K = 1$.

Generated data using ACIVW, compared to real data, obtain very similar results in cross-modal retrieval, showing similar semantic content. The synthetic samples also obtain much better results than using single microphone MFCC replicated for all acoustic pixels, showing that our model is useful to generate proper spatialized audio information. Furthermore, even if being compressed with MFCC, real and generated acoustic images have better accuracy than using spectrograms, showing the usefulness of spatial audio for this task.

As compared to [12], we obtain a better retrieval, furthermore [12] cannot perform retrieval from video to audio, but just the opposite one.

VGGSound dataset is not including real acoustic samples, so we evaluate just synthetic data. The best results are obtained from MCGAN: this model showed to preserve better semantic information than U-VAE generators. On this dataset too we have an increase in CMC score compared to MFCC tiled in space proving once more the beneficial effect of spatial audio produced by generators. Nevertheless, spectrograms are more effective for the domain gap between the audio-visual training data used for the generator and test data employed for cross-modal retrieval. Again, to mitigate the visual domain gap, we substitute ResNet50 used in generation with that pretrained on classification (cl) or for spectrogram-to-video retrieval (tr) (see Table III). This is useful to make MCGAN acoustic images perform better than spectrograms, but it is not the case for AU-VAE and U-VAE.

E. Audio-Visual Localization

Here, we first evaluate the localization results both quantitatively and qualitatively on ACIVW and AVIA using intersection over union (IoU) and area under the curve (AUC). Second, we test on Flickr-SoundNet using consensus IoU and AUC. Since we have no ground truth for VGGSound dataset we can only show some qualitative results for this data.

1) Results for ACIVW and AVIA:

a) Quantitative results: Given synthetic energy g and true energy α, we evaluate quantitatively our results using IoU and AUC considering the binary maps $\mathcal{A}(\tau_1) = \{i \mid \alpha_i > \tau_1\}$, and $\mathcal{G}(\tau_2) = \{i \mid g_i > \tau_2\}$,

where i is the pixel index of the map, τ is the threshold (the average energy for that sample) chosen for true and

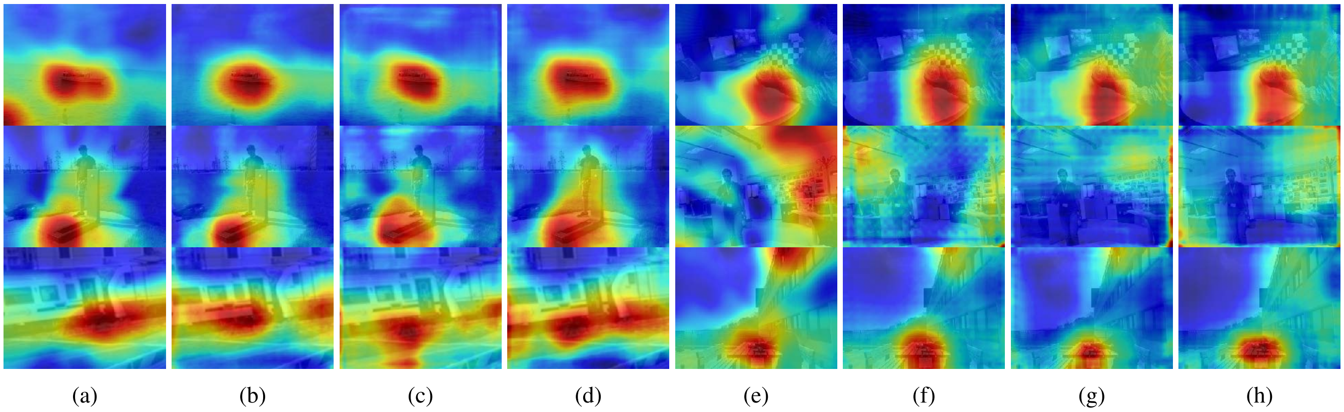

Fig. 4. Qualitative results for audio-visual localization: (a)-(d) for ACIVW and (e)-(h) for AVIA. (a) True energy. Synthetic energy of: (b) U-VAE, (c) AU-VAE, (d) MCGAN. (e) True energy. Synthetic energy of: (f) U-VAE, (g) AU-VAE, (h) MCGAN.

reconstructed energy, respectively. We used 0.5 as IoU threshold. AUC measures the area under the curve for different IoU thresholds varying from 0 to 1 with a step of 0.1.

The results are on the top of Table IV, where we see that training and testing on the ACIVW dataset provides better results than training and testing on AVIA because we have more data and less noise.

b) Qualitative result: The energy of our reconstruction is very similar to the energy of acoustic images for the ACIVW dataset. By comparing Figures 4b, 4c, 4d to real test samples in Figure 4a, it can be noticed from the first row that the reconstructed image is even less noisy than the ground truth one in some examples.

The AVIA dataset is a more challenging benchmark not only because of the noise present in some scenarios, but also due to the periodic nature of the considered sounds. Moreover, in some frames we do not have any sound but only background noise, such as in the second row of Figure 4e, so it is difficult to match sound and video. On the contrary, the ACIVW dataset contains continuous sound and energy is always mapped with visual objects. As we can see in the first row of Figures 4f, 4g, 4h, in the anechoic chamber the sound localization is very precise because sound is present and there is very little noise (as compared to real energy in Figure 4e). Using AVIA ground-truth acoustic images for training we can sometimes improve sound localization removing noise in original energy as shown in last row of Figures 4f, 4g, 4h.

2) Results for Subset of Flickr-SoundNet:

a) Quantitative results: We tested ACIVW and AVIA models on the Flickr-SoundNet test set of [50] even if the considered classes are different. This dataset includes ground-truth bounding boxes to evaluate localization objectively. For comparisons, we evaluated energy estimate using the same metric used in [50], consensus IoU (cIoU), which is based on a consensus map g between different annotators. Such metric takes into account multiple annotations, constructing a weight map assigning a score according to the consensus of many annotations [50]. Given such map g and the energy map α, $cIoU$ is given by:

$$\mathrm{cIoU}(\tau) = \frac{\sum_{i\in\mathcal{A}(\tau)} g_i}{\sum_i g_i + \sum_{i\in\mathcal{A}(\tau)-\mathcal{G}} 1}, \tag{6}$$

where i is the pixel index of the map, τ is the threshold (the computed average energy for that sample), $\mathcal{A}(\tau) = \{i \mid \alpha_i > \tau\}$, and $\mathcal{G} = \{i \mid g_i > 0\}$. We used 0.5 as cIoU threshold. In this case, the AUC measures the area under the curve for different cIoU thresholds varying from 0 to 1 with steps of 0.1.

TABLE IV

WE EVALUATE AUDIO-VISUAL LOCALIZATION ACIVW AND AVIA MODELS AND COMPARE THEM WITH OTHER BENCHMARKS. SENOCAK 1: UNSUPERVISED 10K, SENOCAK 2: UNSUPERVISED 144K RELU, SENOCAK 3: UNSUPERVISED 144K, HU 2019 [54] 1: UNSUPERVISED 20K AUDIOSET, HU 2019 [54] 2: UNSUPERVISED 400K FLICKR-SOUNDNET

Model	**Dataset**	**AUC**	**IoU**
U-VAE	ACIVW	59.7±0.2	76.8±0.2
AU-VAE		66.5±1.9	75.8±1.3
MCGAN		59.3±0.4	77.2±0.5
U-VAE	AVIA	51.2±0.3	54.4±0.7
MCGAN		49.7±0.5	50.2±0.7
AU-VAE		49.6±0.2	50.3±0.1
Train	**Test**	**AUC**	**cIoU**
Senocak 1 [50]	Flickr-SoundNet (subset)	44.9	43.6
Senocak 2 [50]		51.2	52.4
Senocak 3 [50]		55.8	66.0
ACIVW U-VAE		50.3±0.5	53.1±1.9
ACIVW AU-VAE		49.8±0.4	49.6±2.9
ACIVW MCGAN		52.6±0.5	58.9±1.8
AVIA U-VAE		37.2±1.8	20.1±3.0
AVIA AU-VAE		38.6±1.1	20.4±1.6
AVIA MCGAN		44.2±2.5	36.5±6.6
Hu 2019 [54] 1		45.2	41.6
Hu 2020 [53]		49.2	50.0
Qian 2020 [62]		49.6	52.2
Hu 2019 [54] 2		56.8	67.1

We compare our self-supervised model with other unsupervised models in Table IV (bottom). Senocak 1 and 2-3 networks [50] were trained with 10k and 144k Flickr-SoundNet soundtracks-frames respectively. The number of videos in AVIA is 378 (around 136k frames), whereas there are 268 videos in ACIVW (around 220k frames). [50] employed just one frame for each video even if they consider the entire soundtrack. We train our models with a similar number of frames but with fewer videos and smaller temporal duration of the audio signals as we only consider 1/12s for each frame. For the test, we consider each frame with the whole corresponding audiotrack to compute MFCC, in order to carry out a fair comparison with [50]. We get a higher cIoU than two of the

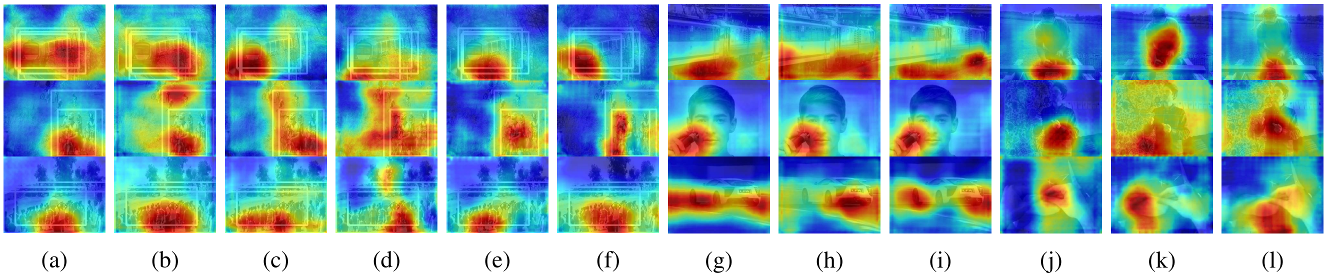

Fig. 5. Qualitative results for audio-visual localization: (a)-(f) for Flickr-SoundNet and (g)-(l) for VGGSound. Synthetic energy of: (a) ACIVW U-VAE, (b) ACIVW AU-VAE, (c) ACIVW MCGAN, (d) AVIA U-VAE, (e) AVIA AU-VAE, (f) AVIA MCGAN, (g) ACIVW U-VAE, (h) ACIVW AU-VAE, (i) ACIVW MCGAN, (j) AVIA U-VAE, (k) AVIA AU-VAE, (l) AVIA MCGAN.

models proposed by [50] using U-VAE and MCGAN trained on ACIVW. We cannot outperform instead the model proposed in Hu et al. [54] trained on 400k videos of Flickr-SoundNet, which is far more data than what we used in training, and above all the same dataset used for testing. However, our models can obtain better results than the recent [62], which was trained on 10k pairs of Flickr-SoundNet, and than [53] and [54], which were trained on 20k pairs of AudioSet-Balanced-Train (this dataset includes many more videos and classes than ACIVW dataset). So, our models result more efficient when using a similar amount of data, as we obtain higher performance even if the training is done with a dataset different from the testing one, including only 10 classes. This shows the effectiveness of the proposed methods, which are able to generalize well to new datasets.

AVIA contains data collected in noisy conditions, so its models' accuracies are lower than the ones we get from ACIVW models. Moreover, we notice that the MCGAN model outperforms U-VAE and AU-VAE models on Flickr-SoundNet localization. In addition, when training and testing on ACIVW, our models have a higher performance than the best models of [50] and [54], trained on 400k videos of Flickr-SoundNet.

b) Qualitative results: We see some results of estimated energy by the ACIVW model and AVIA model in Figure 5(a)-(f). Models perform well in general and capture different possible sources of sound.

3) Results for VGGSound Dataset:

a) Qualitative results: To evaluate ACIVW and AVIA models on real videos we test them on the VGGSound dataset. We only report some qualitative samples as no ground truth data is provided. We consider a subset of VGGSound, choosing classes similar to those considered at training time, depending on the training dataset.

We see examples from the ACIVW model and AVIA model in Figure 5 (g)-(l). The estimated energy maps are very realistic even if belonging to a completely different dataset never seen in training. We obtain the best results with the model trained on ACIVW. We provide more examples for all datasets in the Supplementary Material.

F. Ablation Study

We show here some baselines and ablate on:

- training using also background noise;
- audio or video only sound localization;
- ResNet50 impact on localization;
- skip connections.

TABLE V

ACCURACY OF OUR ARCHITECTURES TRAINED ON ACIVW ON AUDIO-VISUAL LOCALIZATION TASK TESTING BOTH AUDIO AND VIDEO (AV), VIDEO AND BACKGROUND NOISE (V), AUDIO AND NO VIDEO (A)

Model	IoU AV	IoU V	IoU A
VAE	0.7676±0.0019	0.7168±0.0219	0.2337±0.0441
MCGAN	0.7715±0.0052	0.6930±0.0093	0.1410±0.0669
AU-VAE	0.7576±0.0128	0.6659±0.0113	0.2977±0.0440

TABLE VI

ACCURACY OF AUDIO-VISUAL LOCALIZATION USING ENERGY OR RESNET50 FEATURES

Features	ACIVW	Flickr-SoundNet
Self-supervised ResNet50 U-VAE		
Energy	0.7676±0.0019	0.5307±0.0191
12 × 16	0.1219±0.0145	0.4907±0.0147
14 × 19	0.0084±0.0000	0.2160±0.0000
Self-supervised ResNet50 AU-VAE		
Energy	0.7576±0.0128	0.4960±0.0290
12 × 16	0.0768±0.0132	0.4827±0.0245
14 × 19	0.0084±0.0000	0.2160±0.0000
Self-supervised ResNet50 GAN		
Energy	0.7715±0.0052	0.5893±0.0180
12 × 16	0.1497±0.0389	0.4160±0.0753
14 × 19	0.0084±0.0000	0.2160±0.0000
Supervised ResNet50		
CAM[63]	0.1649±0.0094	0.2000±0.0173
12 × 16	0.1228±0.0075	0.2733±0.0161
14 × 19	0.0084±0.0000	0.2160±0.0000

1) Training With Background Noise Samples: Using the ACIVW dataset we train our architecture using correspondent audio and video only. However, in real scenarios, it is possible that an object is not always producing sound and there is instead just background noise. Therefore, we introduce additional input pairs, synthesized by using a low-pass filter on the original audio to simulate the cases when only background noise is audible. In this case, instead of reconstructing a real acoustic image, we train the network to reconstruct the same map given as input, obtained by tiling the filtered audio MFCC vector. In fact, the object we see is producing no sound, so we want to get a uniform energy map. Adding synthetic background noise samples during training, we correctly get a flat energy visualization in such cases, without affecting the other metrics on the ACIVW test set. This solves an issue we experienced in such cases: when training without using background noise samples the network was highlighting the most important object in the scene even if we feed

TABLE VII

ABLATION STUDY ON ACIVW MODEL. MSE VALUES ARE MULTIPLIED BY 10^{-2}. KNN ARE CLASSIFICATION ACCURACIES OF LATENT VARIABLES OR EMBEDDING (AUTOENCODER). GANTEST IS THE ACCURACY TESTING ON GENERATED ACOUSTIC IMAGES. FOR GANTRAIN WE TRAIN ON RECONSTRUCTED ACOUSTIC IMAGES. GANTRAIN1 IS THE ACCURACY ON GENERATED SAMPLES, GANTRAIN2 IS ACCURACY ON REAL ONES. VAE: VAE WITH 1 SKIP CONNECTION. AE: AUTOENCODER. VAE 2S: 2 SKIP CONNECTIONS. VAE 0S: 0 SKIP CONNECTIONS. MCGAN: MULTIMODAL CONDITIONAL GAN. AU-VAE: ADVERSARAL U-VAE. BN: ADDING BACKGROUND NOISE SAMPLES

	VAE	**VAE bn**	**AE**	**VAE 2s**	**VAE 0s**	**MCGAN**	**MCGANbn**	**AU-VAE**	**AU-VAE bn**
MSE	1.143±0.005	1.138±0.006	1.129±0.008	1.125±0.007	1.210±0.013	1.346±0.014	1.331±0.009	1.224±0.008	1.242±0.013
KNN	0.671±0.025	0.715±0.010	0.630±0.008	0.685±0.002	0.782±0.009	-	-	0.665±0.019	0.696±0.020
GANtest	0.834±0.009	0.832±0.010	0.823±0.004	0.826±0.010	0.819±0.011	0.807±0.003	0.834±0.008	0.842±0.017	0.839±0.004
GANtr1	0.851±0.009	0.844±0.001	0.826±0.004	0.843±0.017	0.813±0.004	0.818±0.042	0.836±0.009	0.850±0.015	0.852±0.007
GANtr2	0.766±0.007	0.797±0.012	0.764±0.009	0.783±0.019	0.777±0.008	0.815±0.024	0.819±0.022	0.820±0.019	0.829±0.013
IoU	0.768±0.002	0.760±0.003	0.762±0.010	0.755±0.004	0.694±0.005	0.772±0.005	0.759±0.022	0.758±0.013	0.734±0.012
AUC	0.597±0.002	0.596±0.003	0.592±0.004	0.594±0.002	0.571±0.003	0.593±0.004	0.594±0.008	0.581±0.004	0.576±0.001
IoUFli.	0.531±0.019	0.520±0.025	0.461±0.056	0.468±0.011	0.385±0.024	0.589±0.018	0.579±0.044	0.496±0.029	0.517±0.011
AUCFli.	0.503±0.005	0.505±0.008	0.489±0.013	0.488±0.002	0.462±0.010	0.526±0.005	0.527±0.012	0.498±0.004	0.493±0.001

just background noise. More details are reported in the Supplementary Material.

In the case of periodic sounds, such as in the AVIA dataset [11], since we have background noise samples in between other sound samples, we do not train using the latter strategy on this dataset. Specifically, the lower performances on AVIA come out from the fact that real acoustic images in case of background noise are not flat. Due to background noise, they focus on the border of images randomly, making the training task more complex.

2) Audio- or Video-Only Sound Localization: In Table V we notice that IoU based on audio only (with a black image as visual input) is much lower than when providing both audio and RGB frames to our generators. In fact, the network is probably exploiting the mean position for every sound class. as can be noted from left image of the second row of

We computed also the IoU with video only (and background noise to replace audio input). Results show that the visual modality is essential to add spatial cues which would be missing in omnidirectional audio.

3) ResNet50 Impact on Localization: Localization performance on our dataset is only partially due to ResNet50 being pretrained on ImageNet. The comparison between energy IoU for different ResNet50 feature maps (trained in either a self-supervised or supervised way) is reported in Table VI. First, we notice that 14×19 ResNet50 map pretrained on ImageNet in all four models has the same accuracy because it has been frozen. It has very bad results on Flickr-SoundNet and even worse on our dataset.

The 12×16 map obtained just adding one layer and training with the ACIVW dataset from scratch improves accuracy on Flickr-SoundNet. The improvement is small for ResNet50 trained in a supervised way because it does not receive any supervision from acoustic images, which are useful for localizing correctly sound energy much more than just using video and labels. Furthermore, using the 12×16 map we can improve localization on the ACIVW dataset compared to just using pretrained model. CAM on ResNet50 has similar accuracy to the 12×16 map because it is computed from it and using last layer weights.

In conclusion, pretrained ResNet50 is not enough for localizing correctly. Adding a convolutional layer trained with the supervision of acoustic images is essential to improve localization, despite not being optimal as our complete model, which is indeed needed for the best localization.

4) Skip Connections: We show the results for 0, 1, 2 skip connections in U-VAE model in Table VII. We see that MSE and IoU are much worse without skip connections. We also evaluated the quality of latent features classifying them with a kNN using $k = 15$, discovering that latent variables become less discriminant when introducing skip connections. In fact, the highest accuracy we obtain is without using skip connections, whereas we have a drop of more than 10% in the performance when using them.

Furthermore, even if the MCGAN model has no skip connections as VAE 0 skip connections, it has better results for GAN-train and much higher for localization on ACIVW and Flickr-SoundNet, showing the efficiency of the proposed adversarial model.

Naive autoencoder and the network with 2 skip connections have all metrics similar to the proposed U-VAE. Nevertheless, they generalize less effectively to different datasets.

VI. CONCLUSION

In this work, we have proposed three multimodal architectures based on VAE and GAN, specifically designed to reconstruct acoustic images from standard videos without the use of an array of microphones. We show that the VAE model is better at preserving pixel to pixel correspondence, while GAN models are better at preserving semantic information.

We tested the generated samples on different downstream tasks: classification, retrieval and audio-visual localization, on both datasets including acoustic images and standard videos taken from the Internet never seen during training. Generated data have significant performance when compared to real data in those tasks. Audio-visual localization was performed by exploiting the estimated energy of sound of reconstructed acoustic images, which provides a more accurate sound source localization than recent methods based on the correlation between audio and video data.

REFERENCES

[1] L. Rayleigh, "On our perception of the direction of a source of sound," in *Proc. Musical Assoc.*, vol. 2, 1875, pp. 75–84.

[2] X. Min, G. Zhai, J. Zhou, X. P. Zhang, X. Yang, and X. Guan, "A multimodal saliency model for videos with high audio-visual correspondence," *IEEE Trans. Image Process.*, vol. 29, pp. 3805–3819, 2020.

[3] A. B. Vasudevan, D. Dai, and L. Van Gool, "Semantic object prediction and spatial sound super-resolution with binaural sounds," in *Computer Vision—ECCV 2020*. Springer, 2020, pp. 638–655.

[4] K. Yang, B. Russell, and J. Salamon, "Telling left from right: Learning spatial correspondence of sight and sound," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9932–9941.

[5] R. Gao and K. Grauman, "2.5D visual sound," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 324–333.

[6] C. Gan, H. Zhao, P. Chen, D. Cox, and A. Torralba, "Self-supervised moving vehicle tracking with stereo sound," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7052–7061.

[7] C. Chen et al., "Soundspaces: Audio-visual navigation in 3D environments," in *Computer Vision—ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 17–36.

[8] T. May, S. van de Par, and A. Kohlrausch, "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 7, pp. 2016–2030, Sep. 2012.

[9] V. Murino and A. Trucco, "A confidence-based approach to enhancing underwater acoustic image formation," *IEEE Trans. Image Process.*, vol. 8, no. 2, pp. 270–285, Feb. 1999.

[10] H. V. Trees, *Detection, Estimation, and Modulation Theory, Optimum Array Processing*. Hoboken, NJ, USA: Wiley, 2002.

[11] A. F. Perez, V. Sanguineti, P. Morerio, and V. Murino, "Audio-visual model distillation using acoustic images," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2843–2852.

[12] V. Sanguineti, P. Morerio, N. Pozzetti, D. Greco, M. Cristani, and V. Murino, "Leveraging acoustic images for effective self-supervised audio representation learning," in *Computer Vision—ECCV 2020*. Springer, 2020, pp. 119–135.

[13] A. Zunino, M. Crocco, S. Martelli, A. Trucco, A. D. Bue, and V. Murino, "Seeing the sound: A new multimodal imaging device for computer vision," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 693–701.

[14] M. Crocco, S. Martelli, A. Trucco, A. Zunino, and V. Murino, "Audio tracking in noisy environments by acoustic map and spectral signature," *IEEE Trans. Cybern.*, vol. 48, no. 5, pp. 1619–1632, May 2018.

[15] D. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. ICLR*, 2014, pp. 1–14.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Munich, Germany: Springer, 2015, pp. 234–241.

[17] V. Sanguineti, P. Morerio, A. Del Bue, and V. Murino, "Audio-visual localization by synthetic acoustic image generation," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 2523–2531.

[18] I. Goodfellow et al., "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2014, pp. 2672–2680.

[19] P. Morgado, N. Vasconcelos, T. R. Langlois, and O. Wang, "Self-supervised generation of spatial audio for 360° video," in *Proc. NeurIPS*, 2018, pp. 360–370.

[20] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.* Madison, WI, USA: Omni Press, 2011, pp. 689–696.

[21] M. Suzuki, K. Nakayama, and Y. Matsuo, "Joint multimodal learning with deep generative models," in *Proc. ICLR*, 2017, pp. 1–12.

[22] M. Wu and N. Goodman, "Multimodal generative models for scalable weakly-supervised learning," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2018, pp. 5580–5590.

[23] S. Chaudhury, S. Dasgupta, A. Munawar, M. A. S. Khan, and R. Tachibana, "Text to image generative model using constrained embedding space mapping," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2017, pp. 1–6.

[24] A. Spurr, J. Song, S. Park, and O. Hilliges, "Cross-modal deep variational hand pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 89–98.

[25] D. U. Jo, B. Lee, J. Choi, H. Yoo, and J. Choi, "Associative variational auto-encoder with distributed latent spaces and associators," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2020, vol. 34, no. 4, pp. 11197–11204.

[26] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[27] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, May 2017, pp. 105–114.

[28] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," 2015, *arXiv:1508.06576*.

[29] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

[30] X. Wang and A. Gupta, "Generative image modeling using style and structure adversarial networks," in *Computer Vision—ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Amsterdam, The Netherlands: Springer, 2016, pp. 318–335.

[31] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Computer Vision—ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Amsterdam, The Netherlands: Springer, 2016, pp. 649–666.

[32] Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 415–423.

[33] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[34] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3827–3837.

[35] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 95–104.

[36] K. G. Lore, K. Reddy, M. Giering, and E. A. Bernal, "Generative adversarial networks for depth map estimation from RGB video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 12–1258.

[37] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[38] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[39] P. Chen, Y. Zhang, M. Tan, H. Xiao, D. Huang, and C. Gan, "Generating visually aligned sound from videos," *IEEE Trans. Image Process.*, vol. 29, pp. 8292–8302, 2020.

[40] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[41] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning sound representations from unlabeled video," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2016, pp. 892–900.

[42] D. Harwath, A. Torralba, and J. Glass, "Unsupervised learning of spoken language with visual context," in *Proc. Adv. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2016, pp. 1858–1866.

[43] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," in *Computer Vision—ECCV 2016*. Amsterdam, The Netherlands: Springer, 2016, pp. 801–816.

[44] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2405–2413.

[45] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, "Learning sight from sound: Ambient sound provides supervision for visual learning," *Int. J. Comput. Vis.*, vol. 126, no. 10, pp. 1120–1137, Oct. 2018.

[46] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 609–617.

[47] R. Arandjelović and A. Zisserman, "Objects that sound," in *Computer Vision—ECCV 2018*. Munich, Germany: Springer, 2018, pp. 451–466.

[48] A. Ephrat et al., "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–11, 2018.

[49] A. Owens and A. A. Efros, “Audio-visual scene analysis with self-supervised multisensory features,” in *Computer Vision—ECCV 2018*. Munich, Germany: Springer, 2018, pp. 639–658.

[50] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. S. Kweon, “Learning to localize sound source in visual scenes,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4358–4366.

[51] J. Hershey and J. Movellan, “Audio-vision: Using audio-visual synchrony to locate sounds,” in *Proc. 12th Int. Conf. Neural Inf. Process. Syst.* Cambridge, MA, USA: MIT Press, 1999, pp. 813–819.

[52] G. Monaci, P. Vandergheynst, and F. T. Sommer, “Learning bimodal structure in audio–visual data,” *IEEE Trans. Neural Netw.*, vol. 20, no. 12, pp. 1898–1910, Dec. 2009.

[53] D. Hu, Z. Wang, H. Xiong, D. Wang, F. Nie, and D. Dou, “Curriculum audiovisual learning,” 2020, *arXiv:2001.09414*.

[54] D. Hu, F. Nie, and X. Li, “Deep multimodal clustering for unsupervised audiovisual learning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9240–9249.

[55] D. Hu et al., “Discriminative sounding objects localization via self-supervised audiovisual matching,” in *Proc. NIPS*, 2020, pp. 10077-10087.

[56] S. Parekh, S. Essid, A. Ozerov, N. Q. K. Duong, P. Perez, and G. Richard, “Weakly supervised representation learning for audio-visual scene analysis,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 416–428, 2020.

[57] P. Morgado, Y. Li, and N. Vasconcelos, “Learning representations from audio-visual spatial alignment,” in *Proc. NIPS*, 2020, pp. 4733–4744.

[58] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, “Jointly discovering visual objects and spoken words from raw sensory input,” in *Computer Vision—ECCV 2018*. Munich, Germany: Springer, 2018, pp. 659–677.

[59] J. Ramaswamy and S. Das, “See the sound, hear the pixels,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2959–2968.

[60] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, “Audio-visual event localization in unconstrained videos,” in *Computer Vision—ECCV 2018*. Munich, Germany: Springer, 2018, pp. 252–268.

[61] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman, “Self-supervised learning of audio-visual objects from video,” in *Computer Vision—ECCV 2020*. Springer, 2020.

[62] R. Qian, D. Hu, H. Dinkel, M. Wu, N. Xu, and W. Lin, “Multiple sound sources localization from coarse to fine,” in *Computer Vision—ECCV 2020*. Springer, 2020, pp. 292–308.

[63] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proc. 29th IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2921–2929.

[64] H. Terasawa, M. Slaney, and J. Berger, “A statistical model of timbre perception,” in *Proc. SAPA@INTERSPEECH*, 2006, pp. 1–5.

[65] K. Rao and A. Vuppala, *Speech Processing in Mobile Environments*. Springer, 2014.

[66] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Feb. 2016, pp. 770–778.

[67] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, vol. 25. Red Hook, NY, USA: Curran Associates, 2012, pp. 1097–1105.

[68] A. Asperti, “Variance loss in variational autoencoders,” 2020, *arXiv:2002.09860*.

[69] I. Higgins et al., “beta-VAE: Learning basic visual concepts with a constrained variational framework,” in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*. Toulon, France: OpenReview.net, Apr. 2017. [Online]. Available: https://openreview.net/forum?id=Sy2fzU9gl

[70] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[71] S. Hershey et al., “CNN architectures for large-scale audio classification,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 131–135.

[72] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[73] N. Ahmed, T. Natarajan, and K. R. Rao, “Discrete cosine transform,” *IEEE Trans. Comput.*, vol. C-23, no. 1, pp. 90–93, Jan. 1974.

[74] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[75] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, “Vggsound: A large-scale audio-visual dataset,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 721–725.

[76] K. Shmelkov, C. Schmid, and K. Alahari, “How good is my GAN?” in *Computer Vision—ECCV 2018*. Munich, Germany: Springer, 2018, pp. 218–234.

Valentina Sanguineti received the B.Sc. degree *(summa cum laude)* in electronic engineering and information technology and the M.Sc. degree *(summa cum laude)* in internet and multimedia engineering from the University of Genoa, Genoa, Italy, in 2016 and 2018, respectively, and the Ph.D. degree in computer vision, pattern recognition and machine learning in 2022, which was done in collaboration with the University of Genoa and Istituto Italiano di Tecnologia (IIT), Genoa. During her Ph.D. degree, she has been involved in research on audio and video processing using deep neural networks with IIT.

Pietro Morerio (Member, IEEE) received the B.Sc. and M.Sc. degrees *(summa cum laude)* in physics from the University of Milan, Italy, in 2007 and 2010, respectively, and the Ph.D. degree in computational intelligence from the University of Genoa, Italy. He was a Research Fellow in video analysis for interactive cognitive environments at the University of Genoa, from 2011 to 2012. From 2016 to 2021, he was a Postdoctoral Researcher at the Italian Institute of Technology (IIT), Genoa, Italy, where he is currently a Technologist at the Pattern Analyisis and computer VISion (PAVIS) Research Line. His research focuses on machine learning, deep learning, and computer vision.

Alessio Del Bue (Member, IEEE) is currently a Tenured Senior Researcher leading the Pattern Analysis and computer VISion (PAVIS) Research Line of the Italian Institute of Technology (IIT), Genoa, Italy. He is the coauthor of more than 100 scientific publications in refereed journals and international conferences on computer vision and machine learning topics. His current research interests include 3D scene understanding from multi-modal input (images, depth, and audio) to support the development of assistive artificial intelligence systems. He is a member of the technical committees of major computer vision conferences (CVPR, ICCV, ECCV, and BMVC). He serves as an Associate Editor for *Pattern Recognition* and *Computer Vision and Image Understanding* journals. He is a member of ELLIS.

Vittorio Murino (Fellow, IEEE) received the Laurea degree in electronic engineering and the Ph.D. degree in electronic engineering and computer science from the University of Genova, Genoa, Italy, in 1989 and 1993, respectively. From 2009 to 2019, he worked at the Istituto Italiano di Tecnologia, Italy, as the Director of the Pattern Analysis and Computer Vision (PAVIS) Department. From 2019 to 2021, he worked as a Senior Video Intelligence Expert at the Ireland Research Centre of Huawei Technologies (Ireland) Company Ltd., Dublin. He is currently a Full Professor at the University of Verona, Italy, and a Visiting Scientist at the PAVIS Department, Istituto Italiano di Tecnologia. He is the coauthor of more than 400 papers published in refereed journals and international conferences. His main research interests include computer vision and machine learning, nowadays focusing on deep learning approaches, domain adaptation and generalisation, and multimodal learning for (human) behavior analysis and related applications, such as video surveillance and biomedical imaging. He is a fellow of IAPR and ELLIS, a member of the technical committees of important conferences (CVPR, ICCV, ECCV, ICPR, and ICIP), and a guest coeditor of special issues in relevant scientific journals. He is a member of the Editorial Board of *Computer Vision and Image Understanding* and *Machine Vision and Applications* journals.