# Audio-Visual Localization by Synthetic Acoustic Image Generation

**Valentina Sanguineti,**[1, 2] **Pietro Morerio,** [1] **Alessio Del Bue,** [3] **Vittorio Murino** [1,4,5]
[1] Pattern Analysis & Computer Vision, Istituto Italiano di Tecnologia, Genoa, Italy
[2] University of Genova, Genoa, Italy
[3] Visual Geometry and Modelling, Istituto Italiano di Tecnologia, Genoa, Italy
[4] University of Verona, Verona, Italy
[5] Huawei Technologies Ltd., Ireland Research Center, Dublin, Ireland
{valentina.sanguineti,pietro.morerio,alessio.delbue,vittorio.murino}@iit.it

## Abstract

Acoustic images constitute an emergent data modality for multimodal scene understanding. Such images have the peculiarity to distinguish the spectral signature of sounds coming from different directions in space, thus providing richer information than the one derived from mono and binaural microphones. However, acoustic images are typically generated by cumbersome microphone arrays, which are not as widespread as ordinary microphones mounted on optical cameras. To exploit this empowered modality while using standard microphones and cameras we propose to leverage the generation of synthetic acoustic images from common audio-video data for the task of audio-visual localization. The generation of synthetic acoustic images is obtained by a novel deep architecture, based on Variational Autoencoder and U-Net models, which is trained to reconstruct the ground truth spatialized audio data collected by a microphone array, from the associated video and its corresponding monaural audio signal. Namely, the model learns how to mimic what an array of microphones can produce in the same conditions. We assess the quality of the generated synthetic acoustic images on the task of unsupervised sound source localization in a qualitative and quantitative manner, while also considering standard generation metrics. Our model is evaluated by considering both multimodal datasets containing acoustic images, used for the training, and unseen datasets containing just monaural audio signals and RGB frames, showing to reach more accurate localization results as compared to the state of the art.

## Introduction

Humans interpret the world by means of several senses, being vision and hearing the most crucial ones. More specifically, for interacting with the surrounding environment vision is supported by binaural hearing, which helps people focusing on the sources of sound to better understand 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 (Rayleigh 1875). 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.



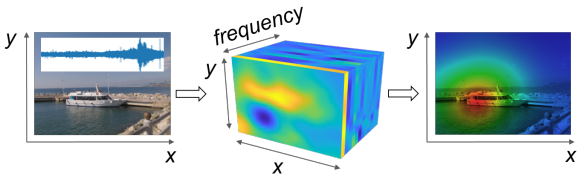



Figure 1: We achieve unsupervised sound source localization in videos through the generation of a spatialized audio frequency map, called *acoustic image*: starting from an RGB frame and the corresponding monaural audio (left), we synthesize the frequency distribution for each direction in space and associate it to a pixel in the acoustic image (center). Localization is then obtained by extracting the energy of sound (right).

Binaural microphone configurations have been lately investigated in audiovisual learning (Vasudevan, Dai, and Van Gool 2020; Yang, Russell, and Salamon 2020; Gao and Grauman 2019a; Gan et al. 2019). However, binaural configurations are limited to the estimation of the direction of arrival only along the azimuth direction and are not able to compete with the performance achieved by the human auditory system in localization tasks (May, van de Par, and Kohlrausch 2012).

In this work, we exploit instead the data gathered by a planar array of microphones, which can produce more accurate spatial audio information and can better help to localize sound sources than stereo audio. In fact, the acoustic signals acquired by an array can be combined via a filter-and-sum beamforming algorithm (Van Trees 2002) to produce an *acoustic image* allowing to localize sound sources on a 2-dimensional space, as we can see in Figure 1, rather than along just one single direction (Yang, Russell, and Salamon 2020). Each pixel of such images contains the audio spectral properties of the sound coming from the corresponding direction in space. Acoustic images have already been proven useful to learn good representations when used both in supervised (Pérez et al. 2020) and self-supervised learning (Sanguineti et al. 2020). Moreover, their spatial content can

be distilled to audio models in order to get more robust features, which generalize better to new datasets than monaural features. They have also been used to face audio tracking (Zunino et al. 2015; Crocco et al. 2018), where acoustic information plays a fundamental role when visual counterpart is cluttered or unreliable.

Unfortunately, microphone arrays are not so common and they are very expensive devices. Therefore, we introduce a novel audio spatialization task, consisting in learning how to generate acoustic images from a standard video, i.e., from single-microphone audio enriched with the visual content of the scene. In this way, even without an array of microphones, we can synthesize acoustic images, which allow to better tackle different audio-visual tasks.

To solve this problem, we propose a novel architecture, which is a hybrid of a Variational Autoencoder (VAE) (Kingma and Welling 2014) and a U-Net (Ronneberger, Fischer, and Brox 2015) models, in order to exploit the upsides of both: VAEs are very effective generation tools and have a principled mathematical formulation, but they show limitation when the size of the output is too large; on the contrary, U-Nets are reconstruction tools which can effectively deal with the small details of large images.

We evaluate the synthetic samples generated by the proposed model by common reconstruction metrics, showing that they are similar to real acoustic images in terms of structure and semantic content. More importantly, we validate them on the task of audio-visual localization, by extracting the energy of the spatialized sound. We show that this energy is very close to that of real acoustic images, which is useful to accurately locate the region originating the sound. We show that our model transfers well on new datasets and performs better than previous state-of-the-art models on unseen data.

In summary, the contributions of our paper can be summarized as follows:

1. We propose to carry out audio-visual localization in a novel fashion, namely through the generation of acoustic images and by estimating the energy of the synthesized spatialized sound.
2. To this end, we propose a new multimodal learning architecture, trained in a self-supervised way, to generate synthetic acoustic images, by jointly processing monaural audio signals and associated RGB images[1].
3. We present a set of experiments to evaluate the quality of the reconstruction of spatialized audio in terms of classification and localization performance. Moreover, ground truth acoustic images allow for a fair evaluation of the sound source localization task, as they are bias-free from human annotations. We also verify that our method generalizes better to datasets never seen in training and it is more accurate than the previous state of the art.

The rest of the paper is organized as follows. We review related works, then we present in Method section the architecture utilized for the generation of the acoustic images and its training strategy. Experiments section reports the experiments and, finally, we draw conclusions.

[1]Code available at https://github.com/IIT-PAVIS/Acoustic-Image-Generation

## Related Works

In this section, we review works regarding topics related to our proposed method: acoustic imaging, audio-visual localization, sound separation and spatialization, cross-modal VAE.

**Acoustic Images.** Acoustic images are generated from the raw audio signals of the microphones of a planar array combining them with the filter-and-sum beamforming algorithm (Van Trees 2002) and summarize the per-direction audio information in the frequency domain. They are volumes, with depth channels corresponding to frequency bins. Handling input data with many frequency bins is a computationally expensive task and typically the majority of information in audio is contained in the low frequencies. Consequently, the acoustic images' channels were compressed to Mel-Frequency Cepstral Coefficients (MFCC) representation (Pérez et al. 2020), which consider audio human perception characteristics (Terasawa, Slaney, and Berger 2006), reducing consistently the computational complexity. So far, deep learning has been applied to the field of acoustic imaging in two works only: (Pérez et al. 2020) proposed an architecture able to classify acoustic images in a supervised way and showed how to distill acoustic images' information to audio models, still in a supervised way; instead, (Sanguineti et al. 2020) proposes a self-supervised learning approach of audiovisual representations. The model we propose is instead intended to *generate* acoustic images starting from RGB images and audio signals.

**Audio-Visual Localization.** The earliest works about sound source localization are (Hershey and Movellan 1999; Monaci, Vandergheynst, and Sommer 2009), which grounded on the natural synchrony between audio and visual signals. Many recent approaches for sound localization exploit two-stream deep network architectures to find the correlation between the sound and visual feature representations (Hu et al. 2020; Hu, Nie, and Li 2019; Hu et al. 2020a; Parekh et al. 2020; Morgado, Li, and Vasconcelos 2020; Harwath et al. 2018; Senocak et al. 2018; Arandjelović and Zisserman 2018). Other works also aggregate temporal information with LSTMs (Ramaswamy and Das 2020; Tian et al. 2018) or optical flow (Afouras et al. 2020). Some works (Qian et al. 2020; Owens and Efros 2018; Owens et al. 2018a) apply the class activation map (CAM) (Zhou et al. 2016), to achieve sound localization.

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

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 by subsequently extracting their energy.

**Sound Separation and Spatialization.** The task of sound *separation* aims at separating sounding objects making noise simultaneously in the scene. (Zhao et al. 2018) proposed the

Mix-and-Separate framework for source separation. They mixed sounds from different videos to generate a complex audio input signal and then trained a model with the objective to separate a sound source from the input (conditioned on the visual input) by estimating a mask for the input spectrogram. (Gao and Grauman 2019b; Zhao et al. 2019) draw inspiration from the previous method; (Gan et al. 2020a) instead of purely conditioning on the visual features proposes to exploit body dynamics in the videos. The goal of sound *spatialization* is instead to separate binaural audio from mono audio (Yang, Russell, and Salamon 2020; Gao and Grauman 2019a). We draw inspiration from methods used for the previous two tasks to propose a novel and challenging different audio spatialization task: predict a more spatialized audio modality, different from the input single-microphone audio. We therefore reconstruct 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.

**Cross-Modal VAE.** One of the first works addressing the problem of generating one modality from another one is (Ngiam et al. 2011) which reconstructs both audio and video from video or audio only by means of an autoencoder. (Suzuki, Nakayama, and Matsuo 2017; Wu and Goodman 2018) instead model the joint representation of all the modalities and they can generate one modality from a joint latent variable. Other works (Chaudhury et al. 2017; Spurr et al. 2018) also find a common shared latent space from single modalities latent variables. Finally, (Jo et al. 2020) proposes to use different VAEs for each modality and translates the latent variable of one modality into the latent variable of another one using an "associator".

Our work differs from the above in two respects: i) our VAE has a U-Net structure to better deal with details: reconstruction is thus performed not only based on the VAE latent variable but also based on intermediate feature maps which retain spatial cues; ii) our latent space is not constructed from all modalities, but only from those available at test time (RGB and monaural audio).

# Method

## Architecture

We generate acoustic images resembling those produced by combining the signals provided by a planar microphone array, starting from monaural audio samples and the corresponding video frame, in order to provide spatial cues which are missing in omnidirectional microphones. The proposed architecture is a VAE with skip connections as in U-Net model, and the concatenation of visual features, as shown in Figure 2. We name this model U-VAE. Acoustic images, as stated in Related Works, contain the frequency information for each acoustic pixel represented with MFCC. Therefore, to simplify the generation task, we feed the MFCC of a single microphone tiled along spatial dimensions, rather than raw waveforms or spectrograms. Similarly, works about audio spatialization and separation (Gao and Grauman 2019a,b) use as input mixed spectrograms and reconstruct spectrograms of interest, in order to have homogeneous input-output. Visual features are extracted using ResNet50 (He et al. 2016), pretrained on ImageNet (Krizhevsky, Sutskever, and Hinton 2012), modified with the removal of global average pooling and the addition of a 2D convolution layer in order to get a visual feature map that preserves spatial cues. We train the last ResNet50 layer only in order to focus on the specific regions producing sound in the considered training datasets. The visual feature map is then concatenated with the last feature map produced by the audio encoder. The two feature maps come from two different streams and can have different ranges of values, thus we normalized them before concatenation.

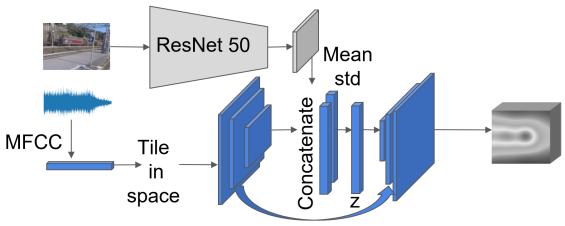



Figure 2: We propose an architecture based on VAE and U-Net (U-VAE) to generate acoustic images. The inputs are monaural audio samples and the corresponding video frame: we compress audio samples to MFCC. ResNet50 visual features are concatenated to audio encoder features.

The network is trained to reconstruct acoustic images for the time interval $1/12\,\mathrm{s}$ as the ground truth acoustic images and RGB images frame rates are 12 frames/s. Therefore, we provide in input MFCC corresponding to $1/12\,\mathrm{s}$ of sound and relative RGB frame. This allows to have almost a real-time estimate of the directional sound, whereas previous works considered from $1\,\mathrm{s}$ (Tian et al. 2018) up to $20\,\mathrm{s}$ (Senocak et al. 2018) of audio to visually localize the sound. Furthermore, considering one frame only for a long audio track can lead to miss important cues about synchronization.

## U-VAE

We introduce U-VAE, a VAE-based architecture able to generate an acoustic image from a single microphone audio signal and the corresponding image. VAEs can in fact improve reconstruction with respect to using a simple autoencoder because the latent loss acts as a regularizer (Asperti 2020).

Assuming that the latent variable distribution $p(\mathbf{z})$ is centered isotropic multi-variate Gaussian and that the inferred posterior distribution is a multi-variate Gaussian with diagonal covariance, means 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 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 reparameterization trick (Kingma and Welling 2014): first sampling a random vector $\mathbf{u}$ from a unit Gaussian $\mathcal{N}(0, \mathbf{I})$, and then multiplying it by the standard deviation $\boldsymbol{\sigma}(\mathbf{x})$ and adding the mean $\boldsymbol{\mu}(\mathbf{x})$:

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

where $\odot$ is the element-wise product. Maximizing the Evidence Lower Bound (ELBO) maximizes the log probability of likelihood of generating data similar to real ones $p(\mathbf{x})$.

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

where $\beta = 1$. The opposite of the first addendum of ELBO in Eq. 2 is often interpreted as a reconstruction loss: we employ not only MSE loss, which is commonly used, but also Huber Loss which provides better generalization. The Kullback-Leibler term KL is the latent loss. As proposed by (Higgins et al. 2017), $\beta$ can be an adjustable hyperparameter that balances the two terms as one regards latent independence constraints and the other one reconstruction accuracy. They propose to consider $\beta > 1$ for good disentangled representations. Instead, we are more interested in obtaining good reconstruction than good latent variables space, therefore we weight latent loss using $\beta < 1$. We tuned the weight of latent loss until the reconstruction loss and latent loss had the same order of magnitude and the network achieved a good reconstruction, keeping such weight fixed for all iterations. As an alternative, some works suggested a KL annealing during the first steps and then to weigh equally both losses (Bowman et al. 2016). However, reconstruction was unsatisfactory as the VAE focused only on minimizing latent loss.

The majority of papers dealing with VAE show results on very small images, (Chaudhury et al. 2017; Suzuki, Nakayama, and Matsuo 2017; Wu and Goodman 2018; Jo et al. 2020; Devaraj et al. 2020). There are few works considering bigger images, such as (Razavi, van den Oord, and Vinyals 2019), which employs VQ VAE with different resolution feature maps from the decoder in order to synthesize high resolution images. Similarly to the last work, we are using intermediate feature maps employing a U-Net architecture and we verify that skip connections can improve the quality of reconstruction.

### Dealing with Silence

We trained our architecture using correspondent audio and video only. However, in real scenarios it is possible that object sound is not always present and we hear instead just background noise. Therefore, we introduce additional input couples, synthesized by using a low pass filter on the original audio. This would like to simulate the cases when the object is producing no sound and 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: since the object we see is producing no sound, we would like a nearly uniform map as the output sound distribution. In fact, in real scenarios if we point an array of microphones to a scene in order to reconstruct the relative spatialized audio, either some objects are producing sound or there is just random background noise.

### Energy of Sound Approximation Method

We propose a novel way to perform localization with a more precise supervision, consisting in the estimation of the spatial sound distribution provided by the acoustic images. We use the energy of the synthesized acoustic images to localize the sound sources, which can be computed from the MFCC representations. Then, we verify that the energy estimates for real and synthetic acoustic images look similar.

To understand how the sound energy is computed, we must remind that MFCC coefficients' extraction from an audio signal requires applying a Discrete Cosine Transform (DCT) to log Mel filters' energies. Actually, summing up Mel filters' energies is a good approximation of the sound energy. Thus, we compute the inverse DCT (IDCT) of MFCC coefficients to recover log Mel filter energies. However, our estimate is not precise because IDCT is performed without the first MFCC coefficient, not included in AVIA and ACIVW acoustic images, because it carries little sound discriminant information (Rao and Vuppala 2014). So, we performed the exponential of estimated log Mel filter coefficients to recover original energies and then we summed them for each acoustic pixel. Now, due to the absence of the first coefficient, we get a map inversely proportional to the real energy. Simply computing its reciprocal allows obtaining an estimation of the energy, allowing to localize the sound sources present in the scene.

## Experiments

This section first describes the employed datasets. After that, we assess the quality of reconstruction of our U-VAE, considering acoustic images ground truth. Our metrics include standard reconstruction error (MSE) and two additional metrics which measure faithfulness and diversity of generated samples by means of classification. Finally, we evaluate both quantitatively and qualitatively audio-visual localization, which is performed on datasets containing acoustic images and on datasets containing videos collected from the Internet.

### Datasets

We consider the following datasets:

- ACIVW (Sanguineti et al. 2020) is a multimodal dataset including acoustic images containing 5 hours of videos acquired in the wild, containing 10 classes: drone, shopping cart, traffic, train, boat, fountain, drill, razor, hairdryer, vacuum cleaner.
- AVIA (Pérez et al. 2020) is a multimodal dataset including acoustic 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 (Aytar, Vondrick, and Torralba 2016) employed by (Senocak et al. 2018) includes sounds sources positions annotated by three subjects, which facilitates quantitative evaluation. We are considering just the testing data, which includes 250 pairs of frames and their corresponding sound.
- VGGSound (Chen et al. 2020) 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 in testing to evaluate the generalization capability of our U-VAE on unseen domains.

## Evaluation of Reconstruction

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:

- Mean square error (MSE) - to measure the reconstruction error for each acoustic pixel;
- GAN-test (Shmelkov, Schmid, and Alahari 2018) - to measure the accuracy of a classifier trained on real acoustic images but evaluated on generated images (we evaluate also on real ones) to quantify semantic similarity to real samples;
- GAN-train (Shmelkov, Schmid, and Alahari 2018) - to measure the accuracy of a classifier trained on generated data and evaluated on real test images (we evaluate also on generated ones). GAN-train metric captures the diversity of generated samples.

We consider the DualCamNet network introduced by (Pérez et al. 2020) for classifying acoustic images, employed for training both on real and then on generated acoustic images to measure GAN-test and GAN-train. We and add one fully connected layer of size 1000 before the last one which has number of classes size to the modified version in (Sanguineti et al. 2020). We classify 1 s, which means 12 acoustic images provided the rate is 12 frames/s.

In Table 1 we evaluate reconstruction of acoustic images for both the test sets of ACIVW and AVIA datasets as they include ground truth acoustic images.

**GAN Test.** We train DualCamNet on real acoustic images from the training sets and we compare its accuracy when tested on real acoustic images and generated ones. We see that training on ACIVW dataset we have only 1% drop when testing on generated acoustic images. AVIA dataset has a bigger drop, 16%, as its acoustic images were collected in noisy scenarios and contain periodic sounds. We also test on synthetic acoustic images created by replicating single-microphone MFCC along the 2 spatial dimensions. We see that the drop in accuracy is huge: 30% for ACIVW and 63% for AVIA, showing that our architecture is essential to generate different MFCC for each acoustic pixel, namely to modulate sound in space.

**GAN Train.** We train DualCamNet on generated acoustic images and compare its accuracy when testing on generated and on real acoustic images. The best result is obtained on ACIVW dataset, where we have only 9% drop when testing on real samples. When testing on AVIA instead the drop is 14%. Nevertheless, we notice that on generated data we have good results for both datasets.

Last, we train DualCamNet on uniform acoustic images artificially created by replicating MFCC from a single microphone and testing on real acoustic images. This experiment is designed to show how our U-VAE is actually modulating MFCC for each spatial direction. Furthermore, when both training and testing on replicated single microphone MFCC we get worse performance than when training and testing from acoustic images (GAN-test on real), proving that spatialized audio allows increasing classification accuracy.

| | Test | ACIVW | AVIA |
|---|---|---|---|
| MSE | - | 1.1426±0.0053 | 0.9483±0.0026 |
| GAN-test | real | 0.8497±0.0014 | 0.8383±0.0022 |
| | gen. | 0.8342±0.0093 | 0.6700±0.0009 |
| | MFCC | 0.5410±0.0175 | 0.2091±0.0027 |
| GAN-train (on gen.) | gen. | 0.8512±0.0089 | 0.7871±0.0039 |
| | real | 0.7661±0.0065 | 0.6456±0.0100 |
| GAN-train (on MFCC) | MFCC | 0.7323±0.0072 | 0.6614±0.0038 |
| | real | 0.4270±0.0186 | 0.1307±0.0119 |

Table 1: Reconstruction metrics for AVIA and ACIVW models. MSE values are multiplied by $10^{-2}$. We specify considered test modalities: real acoustic images, generated acoustic images, tiled MFCC from a single microphone.

## Audio-Visual Localization

We evaluate now localization results both quantitatively and qualitatively firstly on ACIVW, AVIA using intersection over union (IoU) and area under the curve (AUC), then on Flickr-SoundNet using consensus IoU and AUC, whereas we have no ground truth for VGGSound dataset for which can only show some qualitative results.

### Results for ACIVW and AVIA

**1. Quantitative Results** Given synthetic energy $g$ and true energy $\alpha$, 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, $\tau$ is the threshold (in our case mean of energy for that sample) chosen respectively for true energy and reconstructed energy. We used 0.5 for IoU threshold. AUC measures the area under the plot for different IoU thresholds varying from 0 to 1 with a step of 0.1.

The results are shown on the top of Table 2. We see that training and testing on ACIVW dataset we have a better result than when training and testing on AVIA because we have more data and less noise. However, as AVIA is a more difficult dataset, when testing on AVIA the model trained on ACIVW we get worse results than when testing on ACIVW the model trained on AVIA.

**2. Qualitative Results** As regards ACIVW dataset, the energy of our reconstruction is very similar to the energy of acoustic images if we compare Figure 3b to real test samples in Figure 3a. As it can be noticed from the first row of Figure 3b, the reconstructed image is sometimes even less noisy than the ground truth one.

The AVIA dataset is a more challenging benchmark not only because of noise present in some scenarios, but also due to the periodicity of considered sounds: in some frames we do not have any sound but only background noise, such as in the second row of Figure 3c so that is difficult to match sound and video. On the contrary, ACIVW dataset contains continuous sound and energy is always mapping with visual objects. As we can see in the first row of Figure 3d, in the

| Train | Test | AUC | IoU |
|---|---|---|---|
| ACIVW | ACIVW | 59.7±0.2 | 76.8±0.2 |
| | AVIA | 36.3±0.1 | 19.3±1.1 |
| AVIA | AVIA | 51.2±0.3 | 54.4±0.7 |
| | ACIVW | 40.2±0.3 | 26.3±0.6 |

| Train | Test | AUC | cIoU |
|---|---|---|---|
| Senocak 1 | Flickr-SoundNet (subset) | 44.9 | 43.6 |
| Senocak 2 | | 51.2 | 52.4 |
| Senocak 3 | | 55.8 | 66.0 |
| ACIVW | | 50.3±0.5 | 53.1±1.9 |
| AVIA | | 37.2±1.8 | 20.1±3.0 |
| Hu, Nie, and Li 1 | | 45.2 | 41.6 |
| Hu et al. | | 49.2 | 50.0 |
| Qian et al. | | 49.6 | 52.2 |
| Hu, Nie, and Li 2 | | 56.8 | 67.1 |

Table 2: We evaluate on audio-visual localization ACIVW and AVIA models and compare with other benchmarks. Senocak 1: Unsupervised 10k, Senocak 2: Unsupervised 144k ReLU, Senocak 3: Unsupervised 144k, Hu, Nie, and Li 1: Unsupervised 20k AudioSet, Hu, Nie, and Li 2: Unsupervised 400k Flickr-SoundNet.

anechoic chamber the sound localization is very precise because the sound is present and there is little noise (compare to real energy in Figure 3c). In the last row of Figure 3d we see that also in AVIA we can sometimes improve sound localization when there is noise in the original acoustic image.

**Results for Subset of Flickr-SoundNet**

**1. Quantitative Results** We tested ACIVW and AVIA models on Flickr-SoundNet test set of (Senocak et al. 2018) even if the considered classes are different. This dataset includes ground truth bounding boxes to have an objective evaluation of localization. To compare with (Senocak et al. 2018), we evaluated the energy estimate using their metric, which is consensus IoU (cIoU), based on a consensus map $g$ between different annotators.

Given such map $g$ and the energy map $\alpha$,

$$\mathrm{cIoU}(\tau) = \frac{\sum_{i \in \mathcal{A}(\tau)} g_i}{\sum_i g_i + \sum_{i \in \mathcal{A}(\tau) - \mathcal{G}} 1}, \tag{3}$$

where $i$ is the pixel, $\tau$ is the threshold (in our case mean of computed energy for that sample), $\mathcal{A}(\tau) = \{i \mid \alpha_i > \tau\}$, and $\mathcal{G} = \{i \mid g_i > 0\}$. We used 0.5 for cIoU threshold. In this case the AUC measures the area under the plot for different cIoU thresholds varying from 0 to 1 with a step of 0.1. We compare our self-supervised model with other unsupervised models in Table 2 (bottom). We cannot beat the model of (Hu, Nie, and Li 2019) 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 the testing. However, our ACIVW model can obtain results that perform better than the recent (Qian et al. 2020), trained on 10k soundtracks-frames pairs of Flickr-SoundNet and than 2 of the models proposed by (Senocak et al. 2018), trained with 10k and 144k Flickr-SoundNet couples. The number of videos in AVIA is 378 (around 136k frames), whereas there are 268 videos in ACIVW (around 220k frames). (Senocak et al. 2018) employed just one frame for each video even if they consider the entire soundtrack. Thus, we train with a similar number of frames, but fewer videos and fewer seconds of audio as we only consider $1/12\,\mathrm{s}$ of audio for each frame. To be compatible with them, for the test we consider each frame with the whole corresponding audiotrack to compute MFCC. Our ACIVW model performs better also than (Hu et al. 2020; Hu, Nie, and Li 2019), trained on 20k pairs of AudioSet-Balanced-Train, which includes many more videos and classes than ACIVW dataset. So our model is more efficient as, when using a similar amount of data, we obtain higher performances, even if training using a dataset different from the testing including only 10 classes. This shows the effectiveness of the proposed method, which can generalize well to new datasets. Besides, we notice that when training and testing on ACIVW, our model has a higher performance than the best models of (Senocak et al. 2018) and (Hu, Nie, and Li 2019). AVIA contains data collected in noisy conditions, so its model has a lower accuracy than that we obtain from ACIVW model.

**2. Qualitative Results** We see some results of estimated energy by ACIVW model and AVIA model in Figure 3e and Figure 3f. Considering the third row, we can see that ACIVW model can understand that sound of the train is coming from wheels rolling on the rail rather than the train itself. Considering the second row, we see an example in which AVIA model performs better as it is trained considering actions accomplished by people.

**Results for VGGSound Dataset**

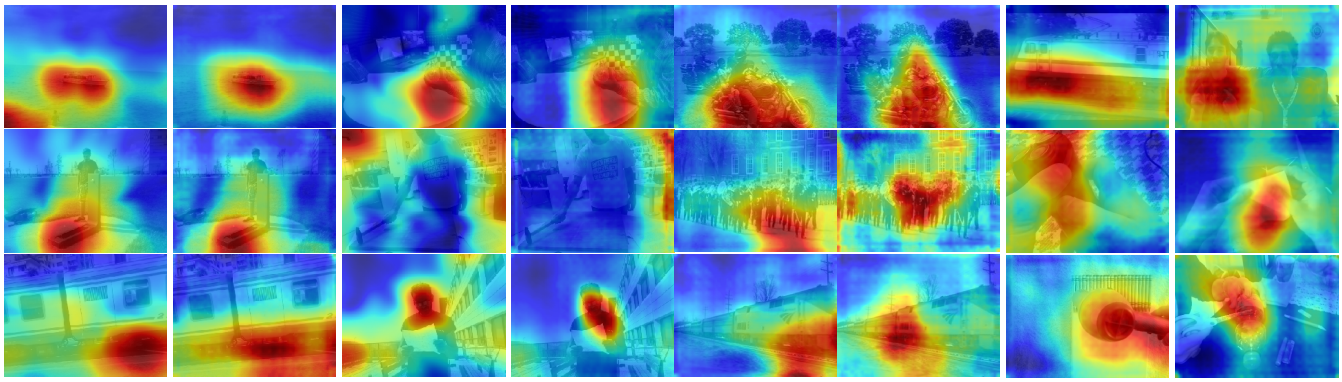
**Qualitative Results** To evaluate ACIVW and AVIA models on real videos we test them on VGGSound dataset. We report some qualitative samples as no ground truths are provided. We consider a subset of VGGSound choosing classes similar to those considered at training time, depending on training dataset. We see examples from ACIVW model in Figure 3g and from AVIA model in Figure 3h. The estimated energy maps are very realistic even if belonging to a completely different dataset never seen during training. Best results are obtained using model trained on ACIVW.

## Ablation Study

We show here some baselines and ablate on:

- Skip connections and training losses
- Audio or video only sound localization
- Sound localization training using also background noise

**Skip Connections and Training Losses.** We show results for 0, 1, 2 skip connections in Table 3. 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 discriminative when introducing skip connections. In fact, the highest accuracy we obtain is without using skip connections, whereas using skip connections we have a drop of more than 10% in the performance. Naive autoencoder and architectures using only MSE loss or only Huber loss



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 3: Qualitative results for audio-visual localization. (a) True energy of ACIVW. (b) Synthetic energy of ACIVW. (c) True energy of AVIA. (d) Synthetic energy of AVIA. (e) Synthetic energy of Flickr-SoundNet using ACIVW model. (f) Synthetic energy of Flickr-SoundNet using AVIA model. (g) Synthetic energy of VGGSound using ACIVW model: train, razor and hairdryer. (h) Synthetic energy of VGGSound using AVIA model: finger snapping, ripping paper and plastic bottle crushing.

| | **VAE** | **VAE bn** | **AE** | **VAE mse** | **VAE huber** | **VAE 2s** | **VAE 0s** |
|---|---|---|---|---|---|---|---|
| MSE | 1.1426±0.0053 | 1.1375±0.0061 | 1.1285±0.0076 | 1.1392±0.0032 | 1.1190±0.0143 | 1.1249±0.0071 | 1.2099±0.0125 |
| KNN | 0.6713±0.0253 | 0.7150±0.0097 | 0.6296±0.0084 | 0.6813±0.0101 | 0.6286±0.0189 | 0.6849±0.0021 | 0.7818±0.0086 |
| GANtest | 0.8342±0.0093 | 0.8317±0.0096 | 0.8231±0.0040 | 0.8329±0.0063 | 0.8280±0.0140 | 0.8259±0.0104 | 0.8187±0.0105 |
| GANtr1 | 0.8512±0.0089 | 0.8441±0.0014 | 0.8256±0.0039 | 0.8451±0.0157 | 0.8442±0.0102 | 0.8427±0.0169 | 0.8126±0.0042 |
| GANtr2 | 0.7661±0.0065 | 0.7968±0.0118 | 0.7636±0.0092 | 0.8022±0.0073 | 0.7854±0.0128 | 0.7826±0.0190 | 0.7772±0.0081 |
| IoU | 0.7676±0.0019 | 0.7597±0.0030 | 0.7620±0.0095 | 0.7463±0.0040 | 0.7758±0.0127 | 0.7550±0.0036 | 0.6937±0.0054 |
| AUC | 0.5973±0.0019 | 0.5955±0.0026 | 0.5916±0.0039 | 0.5910±0.0047 | 0.5971±0.0048 | 0.5944±0.0021 | 0.5705±0.0027 |
| IoUFli. | 0.5307±0.0191 | 0.5200±0.0247 | 0.4613±0.0556 | 0.4587±0.0038 | 0.4360±0.0173 | 0.4680±0.0113 | 0.3853±0.0241 |
| AUCFli. | 0.5027±0.0046 | 0.5052±0.0084 | 0.4889±0.0128 | 0.4827±0.0015 | 0.4749±0.0067 | 0.4879±0.0021 | 0.4619±0.0095 |

Table 3: Ablation study on ACIVW model. MSE values are multiplied by $10^{-2}$. KNN are classification accuracies of latent variables or embedding (autoencoder). GANtest is accuracy testing 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, trained using MSE and Huber losses. VAE bn: adding background noise samples. AE: autoencoder. VAE mse: using only MSE in reconstruction loss. VAE huber: using only Huber loss in reconstruction loss. VAE 2s: 2 skip connections. VAE 0s: 0 skip connections.

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

**Audio or Video Only Sound Localization.** In Table 3 we see that providing audio and RGB frames to U-VAE, IoU is 0.7676. IoU feeding audio and a black image is 0.2337, in which case the network is probably exploiting the mean position for every sound class. IoU with video and background noise is 0.7168, only a bit lower than the original one, showing that visual modality is essential to add spatial clues missing in omnidirectional audio to reconstruct acoustic images, but also audio can give cues to what was sounding object.

**Sound Localization Training Using Also Background Noise.** Training without using background noise samples the network is highlighting the most important object in the scene even if we feed just background noise. Therefore, providing background noise synthetic training data is useful when training with just correspondent audio-video pairs to get a flat energy visualization in such cases. In the case of periodic sounds, such as in AVIA dataset (Pérez et al. 2020), we have background noise samples between sound samples, therefore we do not train with this second 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 the noise coming from outside, they focus on the borders of images randomly, making the training task more complex.

## Conclusions

In this work, we proposed an architecture to reconstruct acoustic images from standard videos, without the use of an array of microphones. These synthetic samples allow performing audio-visual localization in a novel way, exploiting estimated energy of sound, providing a more accurate localization than recent methods based on the correlation between audio and video data. We evaluated quantitatively and qualitatively reconstruction quality and sound source localization both on datasets including acoustic images and on natural videos, showing the soundness of our method.

## References


Afouras, T.; Owens, A.; Chung, J. S.; and Zisserman, A. 2020. Self-Supervised Learning of Audio-Visual Objects from Video. In *Computer Vision – ECCV 2020*. Springer International Publishing.

Arandjelović, R.; and Zisserman, A. 2018. Objects that Sound. In *Computer Vision – ECCV 2018*, 451–466. Springer International Publishing.

Asperti, A. 2020. Variance Loss in Variational Autoencoders. *arXiv preprint arXiv:2002.09860* .

Aytar, Y.; Vondrick, C.; and Torralba, A. 2016. SoundNet: Learning Sound Representations from Unlabeled Video. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, 892–900. USA: Curran Associates Inc. ISBN 978-1-5108-3881-9.

Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A. M.; Józefowicz, R.; and Bengio, S. 2016. Generating Sentences from a Continuous Space. In *CoNLL*.

Chaudhury, S.; Dasgupta, S.; Munawar, A.; Salam Khan, M. A.; and Tachibana, R. 2017. Text to image generative model using constrained embedding space mapping. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6.

Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020. VGGSound: A Large-scale Audio-Visual Dataset. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Crocco, M.; Martelli, S.; Trucco, A.; Zunino, A.; and Murino, V. 2018. Audio Tracking in Noisy Environments by Acoustic Map and Spectral Signature. *IEEE Transactions on Cybernetics* 48: 1619–1632.

Devaraj, C.; Chowdhury, A.; Jain, A.; Kubricht, J. R.; Tu, P.; and Santamaria-Pang, A. 2020. From Symbols to Signals: Symbolic Variational Autoencoders. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, 3317–3321.

Gan, C.; Huang, D.; Zhao, H.; Tenenbaum, J. B.; and Torralba, A. 2020a. Music Gesture for Visual Sound Separation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10475–10484. doi: 10.1109/CVPR42600.2020.01049.

Gan, C.; Zhao, H.; Chen, P.; Cox, D.; and Torralba, A. 2019. Self-Supervised Moving Vehicle Tracking With Stereo Sound. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* 7052–7061.

Gao, R.; and Grauman, K. 2019a. 2.5D Visual Sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gao, R.; and Grauman, K. 2019b. Co-Separating Sounds of Visual Objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Harwath, D.; Recasens, A.; Surís, D.; Chuang, G.; Torralba, A.; and Glass, J. 2018. Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input. In *Computer Vision – ECCV 2018*, 659–677. Springer International Publishing. ISBN 978-3-030-01231-1.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hershey, J.; and Movellan, J. 1999. Audio-vision: Using Audio-visual Synchrony to Locate Sounds. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, 813–819. Cambridge, MA, USA: MIT Press.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M. M.; Mohamed, S.; and Lerchner, A. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *ICLR*.

Hu, D.; Nie, F.; and Li, X. 2019. Deep Multimodal Clustering for Unsupervised Audiovisual Learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9240–9249. doi:10.1109/CVPR.2019.00947.

Hu, D.; Qian, R.; Jiang, M.; Tan, X.; Wen, S.; Ding, E.; Lin, W.; and Dou, D. 2020a. Discriminative Sounding Objects Localization via Self-supervised Audiovisual Matching. *NIPS* .

Hu, D.; Wang, Z.; Xiong, H.; Wang, D.; Nie, F.; and Dou, D. 2020. Curriculum Audiovisual Learning. *ArXiv* abs/2001.09414.

Jo, D. U.; Lee, B.; Choi, J.; Yoo, H.; and Choi, J. 2020. Associative Variational Auto-Encoder with Distributed Latent Spaces and Associators. *Proceedings of the AAAI Conference on Artificial Intelligence* 34: 11197–11204. doi: 10.1609/aaai.v34i07.6778.

Kingma, D.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *ICLR 2014*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25, 1097–1105. Curran Associates, Inc.

May, T.; van de Par, S.; and Kohlrausch, A. 2012. A Binaural Scene Analyzer for Joint Localization and Recognition of Speakers in the Presence of Interfering Noise Sources and Reverberation. *IEEE Transactions on Audio, Speech, and Language Processing* 20(7): 2016–2030.

Monaci, G.; Vandergheynst, P.; and Sommer, F. T. 2009. Learning Bimodal Structure in AudioVisual Data. *IEEE Transactions on Neural Networks* 20(12): 1898–1910. ISSN 1045-9227. doi:10.1109/TNN.2009.2032182.

Morgado, P.; Li, Y.; and Vasconcelos, N. 2020. Learning Representations from Audio-Visual Spatial Alignment. *NIPS* .

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal Deep Learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, 689–696. USA: Omnipress. ISBN 978-1-4503-0619-5.

Owens, A.; and Efros, A. A. 2018. Audio-Visual Scene Analysis with Self-Supervised Multisensory Features. In *Computer Vision – ECCV 2018*, 639–658. Springer International Publishing. ISBN 978-3-030-01231-1.

Owens, A.; Wu, J.; McDermott, J. H.; Freeman, W. T.; and Torralba, A. 2018a. Learning Sight from Sound: Ambient Sound Provides Supervision for Visual Learning. *International Journal of Computer Vision* 126(10): 1120–1137. ISSN 1573-1405. doi:10.1007/s11263-018-1083-5.

Parekh, S.; Essid, S.; Ozerov, A.; Duong, N. Q. K.; Pérez, P.; and Richard, G. 2020. Weakly Supervised Representation Learning for Audio-Visual Scene Analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28: 416–428.

Pérez, A. F.; Sanguineti, V.; Morerio, P.; and Murino, V. 2020. Audio-Visual Model Distillation Using Acoustic Images. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2843–2852. doi:10.1109/WACV45572.2020.9093307.

Qian, R.; Hu, D.; Dinkel, H.; Wu, M.; Xu, N.; and Lin, W. 2020. Multiple Sound Sources Localization from Coarse to Fine. In *Computer Vision – ECCV 2020*, 292–308. Springer International Publishing. ISBN 978-3-030-58565-5.

Ramaswamy, J.; and Das, S. 2020. See the Sound, Hear the Pixels. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2959–2968. doi:10.1109/WACV45572.2020.9093616.

Rao, K.; and Vuppala, A. 2014. *Speech Processing in Mobile Environments*. Springer International Publishing. ISBN 978-3-319-03115-6. doi:10.1007/978-3-319-03116-3.

Rayleigh, L. 1875. On Our Perception of the Direction of a Source of Sound. *Proceedings of the Musical Association* 2: 75–84.

Razavi, A.; van den Oord, A.; and Vinyals, O. 2019. Generating Diverse High-Fidelity Images with VQ-VAE-2. In *Advances in Neural Information Processing Systems*, volume 32, 14866–14876. Curran Associates, Inc.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 234–241. Springer International Publishing. ISBN 978-3-319-24574-4.

Sanguineti, V.; Morerio, P.; Pozzetti, N.; Greco, D.; Cristani, M.; and Murino, V. 2020. Leveraging Acoustic Images for Effective Self-supervised Audio Representation Learning. In *Computer Vision – ECCV 2020*, 119–135. Springer International Publishing. ISBN 978-3-030-58542-6.

Senocak, A.; Oh, T.-H.; Kim, J.; Yang, M.-H.; and So Kweon, I. 2018. Learning to Localize Sound Source in Visual Scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shmelkov, K.; Schmid, C.; and Alahari, K. 2018. How Good Is My GAN? In *Computer Vision – ECCV 2018*, 218–234. Springer International Publishing. ISBN 978-3-030-01216-8.

Spurr, A.; Song, J.; Park, S.; and Hilliges, O. 2018. Cross-Modal Deep Variational Hand Pose Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Suzuki, M.; Nakayama, K.; and Matsuo, Y. 2017. Joint Multimodal Learning with Deep Generative Models. In *ICLR*.

Terasawa, H.; Slaney, M.; and Berger, J. 2006. A statistical model of timbre perception. In *SAPA@INTERSPEECH*.

Tian, Y.; Shi, J.; Li, B.; Duan, Z.; and Xu, C. 2018. Audio-Visual Event Localization in Unconstrained Videos. In *Computer Vision – ECCV 2018*, 252–268. Springer International Publishing. ISBN 978-3-030-01216-8.

Van Trees, H. 2002. *Detection, Estimation, and Modulation Theory, Optimum Array Processing*. Wiley. ISBN 9780471221104. doi:10.1002/0471221104.

Vasudevan, A. B.; Dai, D.; and Van Gool, L. 2020. Semantic Object Prediction and Spatial Sound Super-Resolution with Binaural Sounds. In *Computer Vision – ECCV 2020*, 638–655. Springer International Publishing. ISBN 978-3-030-58548-8.

Wu, M.; and Goodman, N. 2018. Multimodal Generative Models for Scalable Weakly-Supervised Learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS, 55805590. Red Hook, NY, USA: Curran Associates Inc.

Yang, K.; Russell, B.; and Salamon, J. 2020. Telling Left From Right: Learning Spatial Correspondence of Sight and Sound. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhao, H.; Gan, C.; Ma, W.-C.; and Torralba, A. 2019. The Sound of Motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhao, H.; Gan, C.; Rouditchenko, A.; Vondrick, C.; McDermott, J.; and Torralba, A. 2018. The Sound of Pixels. In *Computer Vision – ECCV 2018*, 587–604. Springer International Publishing. ISBN 978-3-030-01246-5.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning Deep Features for Discriminative Localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2921–2929. doi:10.1109/CVPR.2016.319.

Zunino, A.; Crocco, M.; Martelli, S.; Trucco, A.; Del Bue, A.; and Murino, V. 2015. Seeing the Sound: A New Multimodal Imaging Device for Computer Vision. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 693–701. doi:10.1109/ICCVW.2015.95.