Telling Left from Right: Learning Spatial Correspondence of Sight and Sound

Karren Yang*
MIT

Bryan Russell
Adobe Research

Justin Salamon
Adobe Research

http://karreny.github.io/telling-left-from-right

Abstract

Self-supervised audio-visual learning aims to capture useful representations of video by leveraging correspondences between visual and audio inputs. Existing approaches have focused primarily on matching semantic information between the sensory streams. We propose a novel self-supervised task to leverage an orthogonal principle: matching spatial information in the audio stream to the positions of sound sources in the visual stream. Our approach is simple yet effective. We train a model to determine whether the left and right audio channels have been flipped, forcing it to reason about spatial localization across the visual and audio streams. To train and evaluate our method, we introduce a large-scale video dataset, YouTube-ASMR-300K, with spatial audio comprising over 900 hours of footage. We demonstrate that understanding spatial correspondence enables models to perform better on three audio-visual tasks, achieving quantitative gains over supervised and self-supervised baselines that do not leverage spatial audio cues. We also show how to extend our self-supervised approach to 360 degree videos with ambisonic audio.

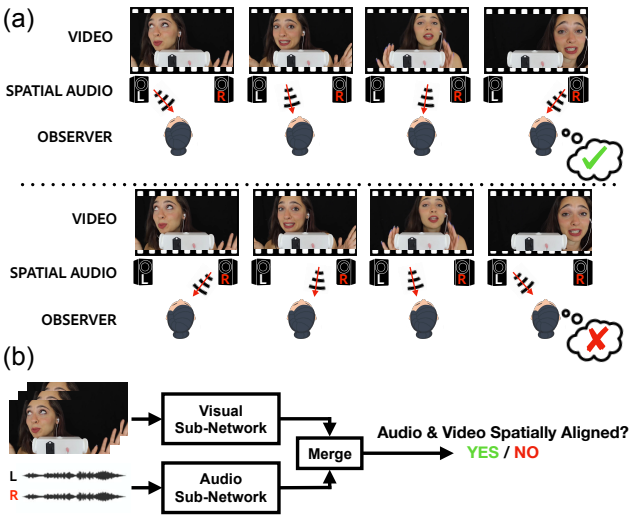

Figure 1. (a) Based on the perceived location of the speaker using spatial audio cues, we can determine when the left/right positions of the sound sources based on sight and sound are aligned (top row) or flipped (bottom row). (b) We teach a model to understand audio-visual spatial correspondence by training it to classify whether a video's left-right audio channels have been flipped.

1. Introduction

Consider Figure 1(a). Here, we illustrate two example videos with the perceived locations of a depicted speaker over time based on the spatial audio[1]. In the first video, notice that our spatial perception of the speaker moving from left to right is consistent between the visual and auditory streams, while in the second video, there is an obvious discrepancy between the two modalities (the left and right audio channels have been flipped, so the sound comes from the wrong direction). This effect is due to the spatial audio signal in these two videos: the audio emulates our real-world auditory experience by using separate left and right (*i.e.*, stereo) channels to deliver binaural cues influencing spatial perception [30]. Because humans have the ability to establish spatial correspondences between our visual and auditory senses, we can immediately notice that the visual and audio streams are consistent in the first video and flipped in the second video. Our ability to establish audio-visual spatial correspondences enables us to interpret and navigate the world more effectively (*e.g.*, a loud clatter draws our visual attention telling us where to look; when interacting with a group of people, we leverage spatial cues to help us disambiguate different speakers). In turn, understanding audio-visual spatial correspondence could enable machines to interact more seamlessly with the real world, improving performance on audio-visual tasks such as video understanding and robot navigation.

Learning useful representations over visual and audio streams in video is challenging. While strong features have been learned using strongly supervised training data [18], large quantities of annotations are difficult to obtain. To

*Work done at Adobe Research during KY's summer internship.

[1]These videos are provided in the Supplementary Materials. We encourage you to watch and listen to the videos wearing headphones.

overcome this challenge, many self-supervised learning approaches have recently been proposed to exploit the audio-visual correspondence in video as a free source of labels [3, 10, 13, 15, 22, 27, 32, 36, 37]. These approaches learn audio and visual representations primarily by matching semantic information or temporal information in audio to the presence or motion of sound sources in the visual stream, without leveraging the audio-visual spatial relation. In contrast, we seek to explicitly focus on spatial correspondence and explore a completely orthogonal approach to audio-visual feature learning based on matching spatial cues in the audio stream to the positions of sound sources in the visual stream. At a time when videos with spatial audio are rapidly proliferating (*e.g.*, due to advances in cellphone mics and AR/VR technology), understanding how to leverage these data to learn strong audio and visual representations is of significant scientific and practical interest.

In this work, we investigate a simple, yet effective, way to teach machines to understand audio-visual spatial correspondence – learn to classify whether a video's left-right audio channels have been flipped, as illustrated in Figure 1(b). We conjecture that a model needs to establish spatial correspondence between audio and visual inputs in order to solve this task. The primary contribution of our work, therefore, is to study the extent to which spatial understanding is useful by (i) proposing a novel self-supervised pretext task for teaching audio-visual spatial correspondence and (ii) evaluating the learned features on an array of downstream audio-visual tasks that could potentially benefit from a strong multimodal spatial representation. Critical to the evaluation of our task is the ability to train on a large video dataset with spatial audio. As part of our contribution, we introduce a new video dataset of ASMR videos from YouTube ("YouTube-ASMR-300K"), the largest reported video dataset with spatial audio comprising over 900 hours of footage. We demonstrate that machines improve on audio-visual tasks through spatial understanding. Over three downstream tasks – (1) sound localization, (2) audio spatialization (upmixing a single mono audio channel to stereo binaural audio channels), and (3) sound source separation – we achieve a quantitative improvement over prior self-supervised audio-visual correspondence tasks and over strongly supervised visual features alone. We also extend the left-right pretext task to 360 degree videos with ambisonic audio and apply the learned embeddings to 360 degree sound localization.

2. Related Work

Self-supervised learning in videos. Due to the challenges in obtaining large-scale annotated data for supervised training, many works have proposed leveraging audio-visual correspondences for self-supervised learning [3, 8, 22, 26, 27, 28, 29]. These tasks learn audio-visual representations either by leveraging the shared semantic information [3, 29] or by exploiting temporal correlation [22, 26, 27, 28]. However, they do not exploit spatial correspondence for self-supervision, and most only take mono audio as input.

Inspiring our effort is prior work on audio-visual correspondence tasks for predicting whether visual and audio signals come from the same video [3], and whether visual and audio signals are temporally aligned [22, 27]. These tasks learn audio and visual representations based on correspondence in semantic or temporal information. In contrast, our correspondence task is designed to teach a model to match spatial audio cues to positions of sound sources in the video and exploits the spatial relation.

Audio-visual source separation. Audio-visual source separation utilizes visual information to aid in the separation of sound mixtures. Many self-supervised approaches have been proposed to solve this task [10, 13, 15, 27, 32, 36], including mix-and-separate frameworks that combine audio tracks from videos and then train models to separate them using visual information [10, 27, 36, 37]. In order to leverage the visual cues, these separation models learn an audio representation that captures the semantics [37] or temporal patterns [10, 36] of sound in order to match them to the visual frames respectively. Strategies to separate audio without explicit mixtures [13, 15] or to co-segment video at the same time [32] have also been proposed. However, all of these approaches still learn an audio representation based on matching semantics or temporal patterns and do not leverage spatial cues.

Sound localization. Estimating the direction-of-arrival of a sound using multiple microphones has been traditionally tackled using beamforming algorithms such as steered power response [6] that do not learn an audio representation and cannot easily handle the concurrency of multiple sound sources. More recently, methods based on neural networks have been proposed for direction of arrival estimation [1, 19, 23]. These models learn spatial audio representations, but they are trained through strong supervision, while we propose to use self-supervision to learn spatial audio cues by leveraging audio-visual spatial correspondence.

A separate stream of research has focused on localizing sound sources in videos [5, 11, 17, 21], including in self-supervised ways [4, 27, 33]. However, these methods do not exploit spatial audio. Recently, Gan *et al*. [12] proposed learning a spatial audio representation that localizes vehicles in video using a pretrained vision network as a teacher. In contrast, our representation is learned by leveraging audio-visual spatial correspondence without explicitly modeling the target locations of a teacher.

Audio spatialization. Several self-supervised approaches have recently been proposed for audio spatialization, which is the task of converting mono audio to spatial audio using a concurrent visual stream to inject the audio with spa-

Figure 2. Examples of videos from our YouTube-ASMR dataset.

tial cues [14, 25, 24]. Similar to us, these approaches use spatial audio as a self-supervisory signal. For example, Gao *et al*. [14] propose a U-Net architecture with a visual stream to convert a mono audio input (generated by downmixing stereo audio) to a stereo output, using the original stereo audio as the target during training. In their model, the visual features provide complementary spatial information that is missing from the mono audio to produce stereo audio. In contrast, our audio representation learns spatial cues directly from stereo audio in order to match the perceived localization of a sound with its position in the video. Also similar to us, Lu *et al*. [24] proposed a spatial correspondence classifier, but they apply it as an adversarial loss for aiding audio spatialization whereas we propose the spatial audio-visual correspondence task for self-supervised feature learning.

3. YouTube-ASMR Dataset

Learning from videos with spatial audio is a relatively new domain. While the amount of spatial audio content is increasing, currently there are few video datasets with spatial audio in which the visual content is spatially aligned with the audio content. We therefore introduce a new large-scale dataset of ASMR videos collected from YouTube that contains stereo audio. ASMR (autonomous sensory meridian response) videos are readily available online and typically feature an individual actor or "ASMRtist" making different sounds while facing towards a camera set up with stereo/binaural or paired microphones. Screenshots from our dataset are shown in Figure 2. The audio in these videos contains binaural cues that, when listened to with headphones, create a highly immersive experience in which listeners perceive the sounds as if they were happening around them. Thus there is strong correspondence between the visual and spatial audio content in these videos.

Our full dataset, YouTube-ASMR-300K, consists of approximately 300K 10-second video clips with spatial audio. From this full dataset, we also manually curated a subset of 30K clips from 30 ASMR channels that feature more sound events moving spatially for training our models. We call this curated dataset YouTube-ASMR. We split the video clips into training, validation, and test sets in an 80-10-10 ratio. Compared to the existing datasets, YouTube-ASMR-300K is (1) larger by at least 8X (Table 1), (2) collected in-the-wild, and (3) contains sound sources in motion (*e.g*., a user waves a tuning fork across the field of view), which is important for training models on diverse spatial cues. The dataset URLs are available on the project website listed on the title page.

Dataset	# Unique Videos	Duration (Hrs)
Lu *et al*. [24]	N/R	9.3
FAIR-Play [14]	N/R *	5.2
YouTube-360 [25]	1146	114
YouTube-ASMR	3520	96
YouTube-ASMR-300K	**33725**	**904**

Table 1. Current large-scale video datasets with spatial audio. Our YouTube-ASMR-300K dataset has the most unique clips and the longest total duration. *2000 10-second clips in total.

4. Learning to Tell Left from Right

Problem Formulation. Spatial audio enables listeners to infer the locations of sound sources. In the case of stereo audio, binaural cues such as differences in time of left and right signals arriving (inter-aural time differences) and the difference in levels of the left and right signals (inter-aural level differences) contribute to the perception of sound sources being localized to the left or right [30]. We hypothesize that these binaural cues can be used to learn useful multimodal representations by teaching a model to recognize when the video and audio are not spatially aligned. During training, we provide as input video clips where we flip the order of the channels in the audio stream with probability 0.5, *i.e*., if the original audio is given by $a(t) = (a_l(t), a_r(t))$, where $a_l(t)$ and $a_r(t)$ are the left and right channels as a function of time t, the flipped audio is $\tilde{a}(t) = (a_r(t), a_l(t))$. This transformation switches the inter-aural time and level differences between the audio channels.

Formally, let $D = \{(v, a, y)_i\}_{i=1}^{N}$ be our video dataset where v is a visual stream, a is an audio stream, and y indicates whether or not a is flipped with respect to v. We train the neural network $f_w(v, a)$ with parameters w to maximize a classification cross-entropy objective given by the log-likelihood,

$$\sum_{(v,a,y)\in D} y \log f_w(v, a) + (1 - y) \log(1 - f_w(v, a)). \quad (1)$$

We conjecture that solving the flipping task requires understanding audio-visual spatial correspondence, as the model must match the location of objects in audio signals with the location of objects visual signals.

Spatial Alignment Network. We illustrate our network in Figure 3(a). The network comprises two streams – a visual stream and an audio stream – which are fused by concatenating features along the time dimension, followed by additional convolutional layers to yield an output representation prior to classification. For vision, our base model uses the public PyTorch implementation of ResNet-18 [16]

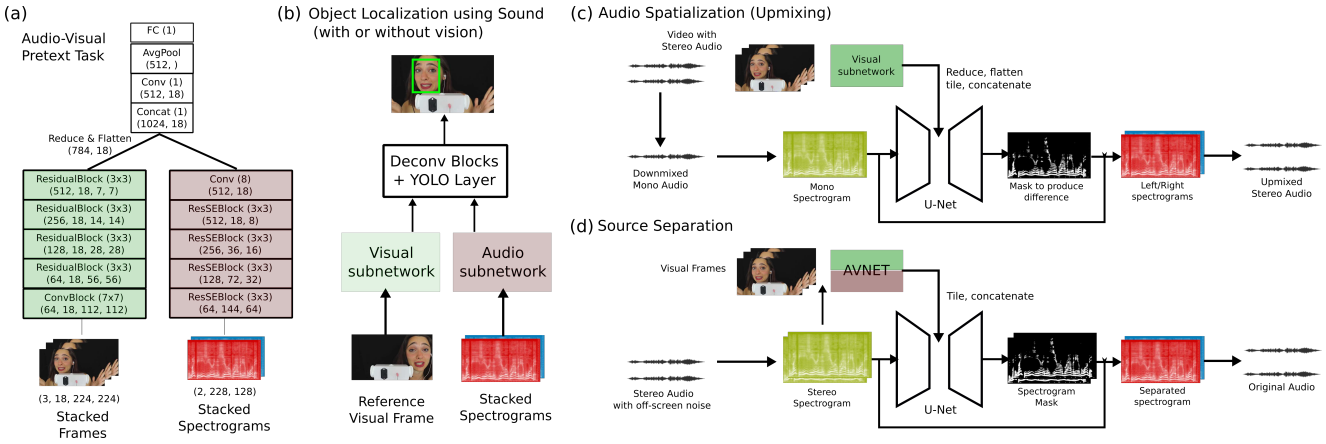

Figure 3. (a) The model architecture for our spatial audio-visual correspondence task. See main text for details. (b-d) Network architectures leveraging our pretrained features for downstream tasks: (b) localization/tracking, (c) audio spatialization, and (d) source separation.

(note that we only apply spatial and not temporal convolutions). We use frames sampled at 6 Hz and resized to 256 x 256 as input. For audio, our base model uses stacked residual blocks with S&E [20], matching the output temporal dimension to that of the vision network. We use the log-scaled mel-spectrogram of audio sampled at 16 kHz and stack the left and right stereo channels as input. Our two-stream network is comparable to the architectures of previous audio-visual correspondence tasks [3, 22]. However, a key distinction is that our model requires the positions of sound sources detected by the visual sub-network. While spatial pooling of the features from the visual sub-network prior to fusion is suitable for other correspondence tasks, we found it necessary for our task to flatten the visual features along the spatial dimensions without pooling prior to fusion with the audio. Models used for audio spatialization tasks, which also require knowledge of the position of sound sources in the visual frame, process visual features in a similar way [14, 25]. For applications to downstream tasks, we can use features from the audio sub-network, visual sub-network, or the fused representation.

Training. We train and evaluate our models on both our YouTube-ASMR dataset and the FAIR-Play dataset [14]. The latter dataset consists of approximately 2K 10-second video clips of people playing instruments. While this dataset is smaller than YouTube-ASMR, we use it to demonstrate the generality of our approach. For both datasets, we use 3-second clips sampled from full clips, introducing flipped audio examples with probability 0.5. We apply a random crop and shift the color/contrast of the frames for data augmentation. To account for possible left-right biases in the audio and visual information, we apply random left-right flipping of both the video and audio channels (note: flipping both at the same time maintains the audio-visual spatial alignment or lack thereof, and thus does not change the prediction target). For optimization, we used SGD (momentum=0.9) with a learning rate of 0.01 on up to 7M samples for YouTube-ASMR and 800K samples for FAIR-Play using 1-4 GPUs.

Baselines and Ablations. Our base model is trained from scratch using ResNet-18 as the visual sub-network architecture as shown in Figure 3(a). To determine if our model can obtain effective visual features with self supervision, we compared against a baseline using ResNet-18 pretrained on ImageNet classification and finetuned on our task ("Pretrained on ImageNet"). To assess the importance of motion features, we trained a model from scratch using MCx as the visual sub-network [35], which uses 3D spatiotemporal convolutions and is designed for video classification ("Motion"). To determine if semantic audio features improve the task performance, we force the model to learn semantic audio features in a third separate stream that only takes mono audio input ("+Mono audio"). This third stream uses the same audio sub-network as stereo audio (with one input channel instead of two) and is similarly fused by concatenating features along the time dimension. Finally, to determine whether the learned audio features additionally benefit from having traditionally computed spatial audio cues, we integrated features computed using Generalized Cross Correlation with Phase Transform [7] to the audio stream by introducing three additional channels in the stereo audio input ("+GCC-Phat").

Results. We report the test set classification accuracy of the audio-visual spatial correspondence model trained on YouTube-ASMR and FAIR-Play in Table 2. Our models perform well and comparably to the supervised baseline; in fact, the performance on the YouTube-ASMR dataset is comparable to human performance on a sub-sample of 200 clips (about 80%, N=2 subjects). This result indicates that our model architecture is well-suited for matching spatial

Model	YouTube-ASMR	FAIR-Play
Pretrained on ImageNet	80.4	92.6
Ours	80.1	93.6
Motion	80.4	71.3
+Mono Audio	**81.3**	**96.3**
+GCC-Phat	80.1	94.1

Table 2. Test set classification accuracy of the pretext task trained on the YouTube-ASMR and FAIR-Play datasets. Our base models trained from scratch perform comparably or outperform a model that uses supervised features from ImageNet classification.

audio cues with sound source locations in the video frames, and also that the spatial audio cues in both datasets are sufficiently rich for learning the visual features of sound sources. We did not obtain gains using the MCx network, suggesting that 3D spatiotemporal convolutional features may not be integral to the task on these datasets. We observed gains using the dual audio model ("+Mono Audio"), indicating that using semantic audio features could potentially aid performance on the pretext task (*e.g.*, when there are multiple sound sources coming from different directions). Finally, GCC-Phat features did not improve the model performance, which suggests that our audio sub-network learned, in a fully self-supervised manner, the spatial localization cues that are traditionally computed using beamforming algorithms. In the downstream task analysis, we use our base model with the ResNet-18 visual sub-network trained from scratch to focus on evaluating audio-visual features learned using spatial audio cues rather than semantic cues.

5. Analysis on Downstream Tasks

5.1. Sound Localization

Does our spatial correspondence task learn an effective representation for mapping spatial audio cues to the positions of sound sources in the visual stream? To answer this question, we first evaluate whether audio embeddings extracted from held-out data contain spatial information. Specifically, we compute the correlation between learned audio features and the sound source's approximate spatial location, which we determine based on the log-energy difference between the two audio channels. We find that the audio features are strongly correlated with the location of the sound source ($R = 0.790$, see Supplemental). In Figure 4(a), we show sound sources automatically tracked over time using our audio embeddings, visualized by a vertical yellow bar in each frame. We generate the visualizations by binning the values of the first principal component of the learned audio embedding and assigning the bins to different horizontal locations in a video frame. As an example, notice in the first row that the yellow bar follows the sound of the moving subject from left to right. These results suggest that

Model	AP50	IOU	X-error	Y-Error
Mono audio	10.2	30.6	11.9	10.4
Stereo audio	15.1	34.8	10.0	10.7
Visual context	27.9	41.0	12.9	**5.5**
Ours	**43.4**	**47.1**	**9.1**	**5.5**
Mono audio	13.9	23.8	17.3	10.0
Stereo audio	26.0	39.0	**9.7**	9.6
Visual context	21.6	27.0	20.7	6.5
Ours	**44.1**	**45.0**	10.8	**5.9**
- Pretext	23.9	35.1	12.4	8.9
+ Pretext	**34.9**	**39.5**	**12.3**	**6.5**

Table 3. Results of audio-visual localization on the YouTube-ASMR test data (Rows 1-4, 9-10) and filtered test data (Rows 5-8). See text for details. Our model outperforms the baselines, and using the pretrained weights from the flipping pretext task outperforms no pretraining.

the audio sub-network learns spatial cues in order to solve the proposed audio-visual correspondence task.

Next, we evaluate whether our model has learned to match the spatial cues in the audio embedding to the positions of sound sources in the visual stream. Specifically, we determine to which regions the visual sub-network attends as described in the Supplemental. We find qualitatively that visual attention is directed to the sound sources in the visual frame, and that there is notable correlation between the region of visual attention and the approximate location of a sound source based on the log-energy difference between the two audio channels ($R = 0.286$, Supplemental). In Figure 4(b), we have visualized examples of this correspondence. Notice in the top row that the visual sub-network attends to the speaker's face (right image), in correspondence with the audio sub-network, which identifies the sound as coming from the right (left image). Overall, the analysis suggests that our spatial alignment model learns a representation that maps spatial audio cues to the positions of sound sources in the visual stream.

Tracking Sounding Faces using Stereo Sound. To leverage our pretrained features for sound localization, we devise a new audio-visual face tracking task on the YouTube-ASMR dataset. The goal is to generate bounding boxes for sounding faces using stereo audio and a visual reference frame taken from a different part of the same video (Figure 3(b)). In practice, such a system could be used to augment vision-based tracking, *e.g.*, spatial sound can allow a system to reason through visual occlusion of a sounding object. Although the sounds in YouTube-ASMR are produced by a variety of objects, we focus on tracking faces for three reasons: (1) almost all videos feature front-facing individuals, (2) many sounds are mouth-based sounds (*e.g.*, whispering, eating) that contain useful signal for localizing faces, and (3) pretrained vision networks such as Reti-

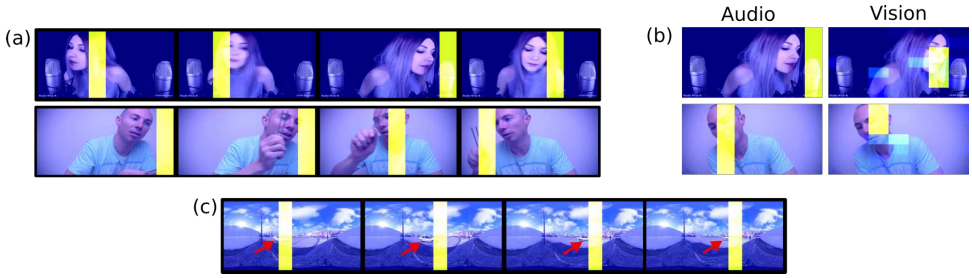

Figure 4. (a) Sound localization results on YouTube-ASMR using audio only. (b) Comparison of localization results using the spatial cues captured by the audio sub-network (left) and the regions of importance of the visual sub-network (right). The visual attention of the model is localized to sound sources and corresponds to the spatial cues learned by the audio sub-network. (c) Sound localization results on YouTube-360 using audio only.

naFace [9] provide a cheap yet reliable source of pseudo ground-truth labels. Our task is motivated by Gan *et al*. [12], which proposes tracking vehicles from stereo sound and camera metadata without visual input.

Model. Our model (shown in Figure 3(b)) is built around the pretrained audio and visual sub-networks from our pretext task. For input, the model takes a one-second audio clip centered around the frame containing the localization target, as well as a reference visual frame from a different part of the same video. The inputs are passed through the sub-networks and concatenated in a one-dimensional vector. Similar to Gan *et al*. [12], the features are then passed through several deconvolution blocks to predict the coordinates of bounding boxes relative to anchor (prior) boxes. We use the object detection loss of YOLOv2 [31] for training. We initialize the audio and visual sub-networks with pretrained weights from the pretext task and fine tune through the entire network.

Results. For evaluation, we consider three metrics: (1) average precision with intersection-over-union (IOU) threshold set at 50 ("AP50"), (2) average IOU of the highest-confidence box per frame ("IOU"), and (3) average error in the x and y coordinates of the highest-confidence box per frame as a percentage of the frame dimension ("X-error", "Y-error"). Rows 1-4 of Table 3 show the performance of our model against several baselines: models that only use mono or stereo audio ("mono audio" and "stereo audio"), and a model that only uses visual context ("visual context"). We find that our model outperforms all of the baselines. However, it is surprising that the performance boost from using stereo audio is not larger. We hypothesize that many videos may feature stationary faces (enabling the visual context model to do well), or contain sounds that do not originate from the face. Therefore, we also evaluate our models on a subset of test videos that are likely to contain moving faces and mouth sounds, *i.e*., we filtered for videos in which there are large left-right shifts in the face's horizontal location that correlate with changes in log-energy difference between the two audio channels. Rows 5-8 of Table 3 show the models evaluated on this subset of test clips. We now observe a significant improvement when using stereo audio. Finally, to determine the extent to which our pretrained features from the pretext task are helpful for this task, we train our model on the full data with fixed sub-network weights (no fine-tuning), with and without pretraining. Using the pretrained weights ("+ pretext task") improves performance across all metrics compared to no pretraining ("- pretext task") (Table 3, Rows 9-10).

5.2. Audio Spatialization (Upmixing)

The goal of upmixing is to convert mono audio to multi-channeled spatial audio, providing the listener with the sensation that sounds are localized in space. Recent work has used a concurrent visual stream to provide spatial information [14, 24, 25]. Specifically, the model is tasked with upmixing the audio stream by matching sounds to the perceived locations of their sources in the video frame. Since our spatial alignment model matches spatial audio cues to visual sound sources, we hypothesize that the pretrained features could be useful for the upmixing task.

Model. We adopt the Mono2Binaural framework of Gao *et al*. [14]. The model takes mono audio and visual frames as input and produces binaural (two-channel) spatial audio as output by producing a complex mask for the difference between the two channels, illustrated in Figure 3(c). We use a U-Net to upmix the audio input, and concatenate our pretrained visual features to the innermost layer of this network. Our model implementation is almost identical to that of Gao *et al*. [14], except we use Tanh activation for producing the complex mask instead of Sigmoid activation because we noticed that the asymmetry of the Sigmoid layer for producing the difference mask biases the upmixing in one direction. These effects were only noticeable because of the strong binaural cues in our dataset. Switching to a Tanh activation layer resolved the bias and quantitatively improved the results, so we maintained this change.

Visual sub-network	YouTube-ASMR	FAIR-Play
Supervised	0.0858	0.403
No visual features	0.0924	0.418
No pretraining	0.0891	0.413
Mismatch task [3]	0.0877	0.412
Shift task [22, 27]	0.0861	0.409
Flip task (ours)	**0.0853**	**0.401**

Table 4. Test set error of upmixing on the YouTube-ASMR and FAIR-Play datasets. Our pretrained features outperform other features, including ResNet-18 trained on ImageNet classification.

Baselines and Evaluation Criteria. We compare our pretrained visual sub-network features ("flip task") against several baselines: (i) No visual features; (ii) ResNet-18 without any training ("no pretraining"); (iii) ResNet-18 trained on the audio-visual correspondence tasks of detecting mismatching semantic information ("mismatch task") [3] or (iv) shifted temporal alignment ("shift task") [22, 27], which use non-spatial audio cues to learn visual features; (v) ResNet-18 trained on ImageNet ("supervised"), which is the visual network used in Gao *et al*. [14]. For training and evaluation criteria, we used the L1 distance between the output and target complex spectrograms averaged over time-frequency bins.

Results. The test set errors are shown in Table 4. Our pretrained features from the spatial alignment detection task improve audio spatialization on both the YouTube-ASMR and the FAIR-Play datasets, outperforming all of the baselines. The difference between our model and the rest is significant based on Wilcoxon signed-rank tests on YouTube-ASMR ($p < 0.1$ for shift task, $p < 0.05$ for rest)[2]. This result suggests that our pretext task, which uses spatial audio cues to guide the video sub-network, successfully teaches the visual sub-network to extract features corresponding to sound sources. Importantly, our features outperform the pretrained features from other audio-visual self-supervised correspondence tasks that use non-spatial cues trained on the same dataset. These results suggest that in the YouTube-ASMR and FAIR-Play datasets, the spatial information in the audio may be a richer source of signal than the semantic information. Overall, we show that spatial audio cues are a powerful alternative to semantic audio cues for learning visual features in a self-supervised manner.

5.3. Audio-Visual Source Separation

Next, we evaluate our pretrained audio-visual features on sound source separation, where the objective is to estimate individual sound sources based on their mixture using visual information. Since we are working with video datasets with spatial audio, the binaural cues could also be valuable in aiding separation; human listeners can distinguish sound sources not only based on differences in timbre, but also due to their spatial position. Therefore we hypothesize that a joint audio-visual representation that captures spatial information could improve performance on this task.

[2]See project website for spatialization examples

Visual sub-network	YouTube-ASMR	FAIR-Play
No visual features	0.0946	0.423
No pretraining	0.0953	0.422
Mismatch task [3]	0.0923	0.423
Shift task [22, 27]	0.0918	0.422
Flip task (ours)	0.0898	0.410
Supervised	0.0885	0.362
Ours + supervised	**0.0863**	**0.350**

Table 5. Test set error of source separation performed on YouTube-ASMR and FAIR-Play datasets. Our pretrained features outperform other self-supervised features and boost the performance of supervised ResNet-18 features.

Model. To test this hypothesis, we used the mix-and-separate framework [37] adapted for stereo audio [14]. Our model takes mixed stereo audio tracks and visual frames as input and produces separated audio corresponding to the given visual stream (Figure 3(d)), whereas Gao *et al*. [14] uses two visual streams for separation. We use a U-Net to generate a mask for the mixed audio input and produce the separated output. Pretrained audio-visual features are concatenated to the innermost layer of the network.

Baselines and Evaluation Criteria. We compare our pretrained audio-visual network with several baselines on the source separation task: (i) No audio-visual features; (ii) Our network without any pretraining ("no pretraining"); Networks trained on the correspondence tasks of detecting (iii) mismatching semantic information ("mismatch task") [3] or (iv) shifted temporal alignment ("shift task") [22, 27], which use non-spatial audio cues to learn visual features; (v) ResNet-18 trained on ImageNet classification ("supervised"). For the training and evaluation criteria, we used the L1 distance between the output and target magnitude spectrograms averaged over time-frequency bins.

Results. The evaluation results on the test set are shown in Table 5. The pretrained features from the audio-visual spatial correspondence task are useful for source separation on both the YouTube-ASMR and the FAIR-Play datasets, outperforming all of the baselines except the strongly supervised ResNet-18 model trained on ImageNet. This observation is consistent with the fact that the source separation task depends on visually discriminating between different sound sources, for which ImageNet classification is well-suited. On the other hand, our features capture the spatial position of sound sources from both the audio and the visual frames, which may be helpful to the source separation task in complementary ways, *e.g*., if the mixed sounds are

coming from two different locations. To determine if this is the case, we trained a new model by concatenating our audio-visual features with the pretrained ImageNet features ("Ours + supervised"). The addition of our features yielded a significant boost over using only the strongly supervised ResNet-18 features. The difference between this model and the rest is significant based on Wilcoxon signed-rank tests on YouTube-ASMR ($p < 0.05$ for all tasks). These results indicate that capturing audio-visual spatial correspondence is useful for sound source separation.

6. Spatial Alignment in 360-Degree Video

Our pretext task for learning spatial correspondence between sight and sound can be extended to 360-degree videos with full-sphere first-order ambisonics (FOA) audio. A key motivation for generalizing the task to this domain is learning a spatial representation of surround sound that can be applied to direction-of-arrival (DOA) estimation [2]. However, recordings with DOA annotations are extremely difficult to obtain, and currently training models relies on synthetic datasets for strongly supervised labels [1, 2, 19, 23]. To address this constraint, we introduce a generalization of our audio-visual correspondence task that learns strong spatial audio cues from 360-degree videos with real spatial audio in a self-supervised manner.

Problem Formulation. First-order ambisonics (FOA) extends stereo audio to the 3D setting, with extra channels to capture sound depth and height at time t: $a(t) = (a_w(t), a_y(t), a_z(t), a_x(t))$, where $a_w(t)$ represents omni-directional sound pressure, and $(a_y(t), a_z(t), a_x(t))$ are front-back, up-down, and left-right sound pressure gradients respectively. FOA is often provided with 360-degree video to give viewers a full-sphere surround image and sound experience. Analogous to the case of stereo audio, we can train a model to detect whether the visual and audio streams are spatially aligned in 360-degree videos. To generate misaligned examples, we propose the transformation, $\tilde{a}(t) = (a_w(t), a_x(t)\sin\theta + a_y(t)\cos\theta, a_z(t), a_x(t)\cos\theta - a_y(t)\sin\theta)$ which rotates the audio about the z-axis by θ.

Implementation Details, Training and Results. We use the same model architecture as for field-of-view video and stereo audio (depicted in Figure 3). The main difference is that the input to the audio sub-network has the four FOA channels instead of two stereo channels. To train our model, we use the YouTube-360 dataset, which contains over a thousand 360-degree videos with FOA audio [25]. Our model achieves about 60% classification accuracy on the YouTube-360 test set; see Supplemental for details.

Sound Localization. Does the audio sub-network trained on the audio-visual spatial correspondence task learn a strong representation of 360-degree surround sound? We first investigate whether the audio embeddings extracted from held-out data contain 360-degree spatial audio cues. We observe that the audio features are strongly correlated with the directional energy of the sound based on an unsupervised projection using principal component analysis. Similar to Section 5.1, we bin the embeddings based on their first two principal components and project the bin over the horizontal range of the video to track sound sources in the video using our self-supervised audio features. We show qualitative results in Figure 4(c): notice that a moving vehicle is tracked from left to right.

Method	Mean Error
Trained from scratch	20.5
Pretrained weights	**17.5**
One-shot SVM on random embeddings	78.6
One-shot SVM on our embeddings	**29.5**

Table 6. DOA estimation performance on the Tau Spatial Sound dataset. Error shown is mean prediction error in degrees (lower is better, maximum is 180, random is 90). Standard deviation of the top models are 1.17 and 0.62 based on 4-fold cross validation.

To determine whether the learned spatial audio cues provide a quantitative boost on spatial audio localization tasks, we evaluated our audio sub-network on the Tau Spatial Sound dataset [2]. This dataset consists of 400 minute-long recordings of multiple sound events synthetically up-mixed to 36 different directions along the X-Y plane (every 10 degrees). Our objective is to predict the direction-of-arrival (DOA) of sound events given FOA audio input. We first compared a baseline model trained from scratch with one that was initialized using the weights from our pretrained audio sub-network. We found that our pretrained weights provided a notable boost over the baseline, using the error in the azimuth angle as the evaluation criteria (Table 6). Next, we extracted the embeddings of the Tau Spatial Sound audio clips using our pretrained audio sub-network and performed classification using a linear SVM, providing only one example of an event from each azimuth angle (one-shot learning). Even in this case, we were able to predict the direction of arrival with a mean error of only about 30 degrees. These results suggest that the spatial information extracted using our self-supervised task are useful for acoustic localization.

7. Conclusion

We have demonstrated a simple, yet effective, approach for self-supervised representation learning from video with spatial audio and its application to three downstream audio-visual tasks. Critical to our approach was the ability to train on a large corpus of video with spatial audio. Our work opens up the possibility of exploring network architectures and cross-modal self-supervised training losses [34] that jointly leverage the spatial and semantic cues present in the visual and spatial audio channels in an effective manner.

References

[1] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):34–48, 2018.

[2] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. A multi-room reverberant dataset for sound event localization and detection. *arXiv preprint arXiv:1905.08546*, 2019.

[3] Relja Arandjelovi and Andrew Zisserman. Look, listen and learn. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017.

[4] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451, 2018.

[5] Zohar Barzelay and Yoav Y Schechner. Harmony in motion. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[6] Taras Butko, Fran González Pla, Carlos Segura, Climent Nadeu, and Javier Hernando. Two-source acoustic event detection and localization: Online implementation in a smart-room. In *2011 19th European Signal Processing Conference*, pages 1317–1321. IEEE, 2011.

[7] Yin Cao, Turab Iqbal, Qiuqiang Kong, Miguel Galindo, Wenwu Wang, and Mark Plumbley. Two-stage sound event localization and detection using intensity vector and generalized cross-correlation. Technical report, DCASE2019 Challenge, June 2019.

[8] Virginia R de Sa. Learning classification with unlabeled data. In *Advances in neural information processing systems*, pages 112–119, 1994.

[9] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019.

[10] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Trans. Graph*, 37:11, 2018.

[11] John W Fisher III, Trevor Darrell, William T Freeman, and Paul A Viola. Learning joint statistical models for audio-visual fusion and segregation. In *Advances in neural information processing systems*, pages 772–778, 2001.

[12] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2019.

[13] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[14] Ruohan Gao and Kristen Grauman. 2.5D visual sound. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[15] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2019.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] John R Hershey and Javier R Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *Advances in neural information processing systems*, pages 813–819, 2000.

[18] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[19] Toni Hirvonen. Classification of spatial audio location and content using convolutional neural networks. In *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.

[20] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[21] Einat Kidron, Yoav Y Schechner, and Michael Elad. Pixels that sound. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 88–95. IEEE, 2005.

[22] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[23] Kuba Lopatka, Jozef Kotus, and Andrzej Czyzewski. Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations. *Multimedia Tools and Applications*, 75(17):10407–10439, 2016.

[24] Yu-Ding Lu, Hsin-Ying Lee, Hung-Yu Tseng, and Ming-Hsuan Yang. Self-supervised audio spatialization with correspondence classifier. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3347–3351. IEEE, 2019.

[25] Pedro Morgado, Nuno Vasconcelos, U C San, Diego Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360° video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[26] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[27] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2018.

[28] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2405–2413, 2016.

[29] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European conference on computer vision*, pages 801–816. Springer, 2016.

[30] Lord Rayleigh. On our perception of the direction of a source of sound. *Proceedings of the Musical Association*, 2:75–84, 1875.

[31] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[32] Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh Mcdermott, and Antonio Torralba. Self-supervised audio-visual co-segmentation. In *ICASSP*, 2019.

[33] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018.

[34] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *ArXiv*, abs/1906.05849, 2019.

[35] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[36] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2019.

[37] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh Mcdermott, and Antonio Torralba. The sound of pixels. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2018.