Learning Spatial Features from Audio-Visual Correspondence in Egocentric Videos

Sagnik Majumder[1,2] Ziad Al-Halah[3] Kristen Grauman[1,2]
[1]UT Austin [2]FAIR, Meta [3]University of Utah

Abstract

We propose a self-supervised method for learning representations based on spatial audio-visual correspondences in egocentric videos. Our method uses a masked auto-encoding framework to synthesize masked binaural audio through the synergy of audio and vision, thereby learning useful spatial relationships between the two modalities. We use our pretrained features to tackle two downstream video tasks requiring spatial understanding in social scenarios: active speaker detection and spatial audio denoising. Through extensive experiments, we show that our features are generic enough to improve over multiple state-of-the-art baselines on both tasks on two challenging egocentric video datasets that offer binaural audio, EgoCom and EasyCom. Project: http://vision.cs.utexas.edu/projects/ego_av_corr.

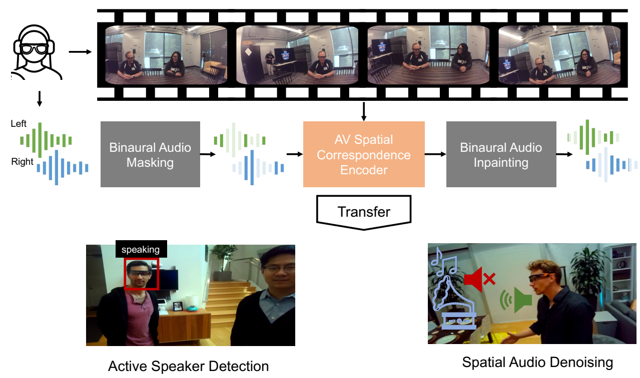


Figure 1. Given an egocentric video and binaural audio, we aim to learn spatial correspondences between vision and audio by solving the pretext task of inpainting segments of the binaural audio. The features benefit downstream social tasks where spatial localization is important: active speaker detection and audio denoising.

1. Introduction

Egocentric videos provide a first-person view of how we perceive and interact with our surroundings in daily life, and they are pushing a new frontier in multi-modal learning [9, 25, 31, 62]. A key aspect of ego-video is that it can provide a rich stream of first-person spatial audio[1] alongside visual frames when the audio is captured with multiple microphones [11, 49]. The coupling of such audio and vision provides strong spatial information about the sound sources (*e.g.* where the sound sources are, if they are in motion or not) in the context of the surrounding physical space (*e.g.* how big or small the room is, if there is a large wall nearby), as well as the camera wearer's attention in the scene revealed by how they move their head.

Such spatial cues are especially important for social settings of multiple people talking to each other, where it is valuable to be able to 1) focus on the voice(s) of interest from among various competing sounds and 2) understand where people are directing their speech activity, for better comprehension and communication. In this way, future AR applications in conversational settings could allow a hearing-impaired person to determine who is speaking in order to redirect their attention, or enhance the received audio to make it more intelligible for any listener.

We argue that this creates the need for human-centric spatially-grounded understanding of audio-visual events. Can we learn representations from video that capture audio-visual events *in the context of the persistent physical space of the environment and the human speakers in it*? Such representations would be useful for answering questions like "who is speaking right now?" and "what would the voices sound like without the audio noise (distractor sounds)?" While the former requires inferring the source location for a voice in the scene, the latter requires understanding how the perceived audio is a function of the source locations, the listener, and the surrounding environment.

Despite the significance of these problems, today's models for audio-visual learning are not human-centric and they lack spatial grounding. On the one hand, current audio-visual representation learning methods exclusively tackle exocentric (third-person) video with monaural audio [2, 23, 24, 29, 34, 50, 52]. That domain sidesteps challenges

[1]Throughout we use the term *spatial audio* to refer to *binaural* audio, as perceived by two human ears. In contrast, single-channel *monaural* audio lacks spatial information.

inherent to ego-video arising from the camera wearer's head motion and relatively limited field of view. On the other hand, the limited prior work exploring self-supervised objectives using multi-channel audio and video [18, 22, 45, 70] are also outside of the egocentric and social contexts (e.g., for sounds in empty homes [22], musical instruments [18], or single human speakers [70]), where the need for spatial understanding of multiple sound sources is limited.

We propose to learn audio-visual representations via spatial correspondence between an egocentric video and its binaural audio, for analyzing social (conversational) settings. In particular, we design a novel pretext task where the goal is to inpaint binaural (two-channel) audio using both video and audio. Given a social egocentric video clip with binaural audio, we mask segments of it and train a model based on masked autoencoding (MAE) [5, 15, 26, 30, 66] to predict the missing segments on the basis of the video and the unmasked segments in the audio. See Figure 1 (top). Additionally, we introduce a novel spatial audio masking strategy that facilitates learning strong audio-visual spatial correspondences while maintaining learning stability when vision alone is insufficient for the binauralization task. Once trained, our model's encoder provides spatial audio-visual features that can be used to address multiple downstream tasks, as we demonstrate using multiple different backbones and social egocentric video datasets.

In particular, motivated by the AR applications discussed above, we validate our feature learning method on two downstream social egocentric tasks that require strong audio-visual spatial reasoning: **1) active speaker detection:** predicting which person in the field of view of an egocentric video is speaking, and **2) spatial audio denoising:** separating audio noise (any sounds from non-speakers) from the input audio. See Figure 1 (bottom). We test the generality of our method by evaluating on two social egocentric video datasets, EgoCom [49] and EasyCom [11]—to our knowledge, the only two publicly available video datasets with binaural sound and social settings. On both, our method significantly outperforms multiple state-of-the-art task-specific and audio-visual spatial feature learning models.

2. Related Work

Audio-visual self-supervised pretraining Past work [2, 4, 34, 45, 48, 50–52] extensively studies the synergy of vision and audio for learning representations through self-supervision. They explore using both modalities to construct pretext tasks based on synthesis [51, 52], alignment [2, 4, 22, 34, 50], and masked auto-encoding (MAE) [23, 24, 29], and they focus on *semantic* downstream tasks like audio-visual event classification and retrieval. However, none of these methods are designed to extract spatial cues from video and multi-channel audio, nor do they analyze the social egocentric setting. On the contrary, we propose self-supervised learning of spatial audio-visual features from egocentric video. Further, different from existing MAE-style models [23, 24, 29], we propose a specialized masking strategy that better learns spatial audio-visual cues. Our masking idea promotes the encoding of spatial and semantic information in the learned multimodal representation, thereby improving its effectiveness for transfer learning in downstream tasks that require nuanced reasoning about both *what* and *where* aspects, such as active speaker detection and spatial audio-visual denoising. This differs from previous methods [23, 24, 29], which mainly use a learning objective that emphasizes the encoding of semantic cues and tailor to tasks like multimodal event classification or retrieval.

Audio-visual spatial correspondence learning Learning the *spatial* alignment between video and audio is important for self-supervision [45, 60, 69, 70], audio generation [7, 18, 40, 44, 46, 56, 73], audio-visual embodied learning [6, 8, 38, 39] and 3D scene mapping [41, 55]. However, these methods are either restricted to exocentric settings [7, 18, 44, 45, 56, 60, 70], or else tackle egocentric settings [8, 38, 40, 41] in simulated 3D environments that lack realism and diversity, both in terms of the audio-visual content of the videos (no people are visible, no objects are moving) and their lack of continuous camera motion from a camera-wearer's physical movements. In contrast, we learn an audio-visual representation from real-world egocentric video.

More closely related to our work are Telling Left from Right (TLR) [70], 2.5D Visual Sounds (2.5D-VS) [18], and audio-visual stereo sound ranking (SSR) [60], all of which learn spatial audio-visual features, albeit for exocentric data only. TLR predicts whether the left and right binaural channels are swapped, and SSR ranks the similarity of video to different stereo audio samples through self-supervision—both of which provide only coarse spatial information about the scene. 2.5D-VS learns to "lift" the mono input to binaural audio, which can be underconstrained from the single-channel audio and video alone. We design a novel pretext task using audio-visual inpainting of binaural audio, which is both fine-grained (requiring to capture subtleties about the arrangement of speakers in the environment) and, through our novel masking strategy, exposes better multi-modal constraints that improve learning stability and performance. Our results in Sec. 4 show our model's advantages over all three prior methods [18, 60, 70].

Active speaker detection Active speaker detection (ASD) entails predicting the active speaker(s) from among all detected faces in a video, and is a special case of generic 2D sound localization [17, 27, 32, 43, 50, 71]. While early ASD methods rely on lip motion and facial gestures [14], recent methods employ ensemble networks [3] or 3D CNNs [33, 61, 65], relation context modules [72], attention [3, 65], or graph neural networks [35, 42]. Multi-

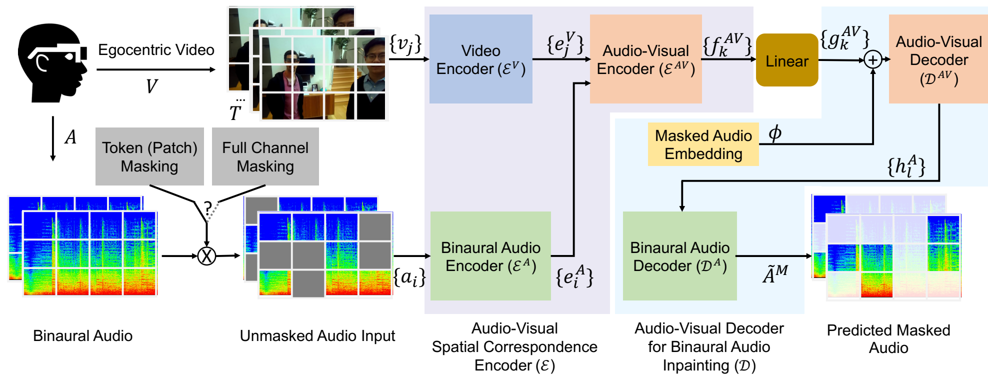


Figure 2. Our model learns the spatial correspondence between vision and binaural audio by inpainting masked tokens of the audio channels through the use of an audio-visual encoder-decoder model. We combine random token masking (which requires solving a more *local* binauralization task) with complete audio channel masking (which requires more *global* cues to synthesize unseen binaural segments). For downstream evaluation, we fuse the features from the audio-visual encoder with the backbones for downstream tasks, and finetune them.

channel audio improves ASD [32], but requires privileged information of the speaker's pose for training. Recent work explores using supervised learning to infer not only who is talking, but also to whom a camera-wearer is listening [59]. In contrast, our goal is to learn spatial audio-visual features purely from in-the-wild egocentric video through self-supervision—features generic enough to benefit multiple ASD models, as we show for both TalkNet [65] and SPELL [42].

Spatial audio denoising Audio denoising, which requires separating a target sound from noise, has traditionally been studied with single-channel (non-spatial) audio, both in the audio-only setting [28, 63, 64, 68] and audio-visual settings [1, 13, 16, 19–21, 50, 53, 54]. Using spatial audio captured with multiple microphones [12, 47, 74] naturally makes the task simpler. Different from the above, we learn task-agnostic audio-visual spatial features. That is, our contribution is the feature learning idea (which benefits both denoising and ASD), rather than a novel denoising approach.

3. Learning spatial features from egocentric audio-visual correspondence

The spatial sound perceived in an egocentric setup is shaped by the environment in which it is emitted and its source location relative to the camera-wearer. Based on this knowledge, we hypothesize that trying to solve the pretext task of audio-visual inpainting of binaural audio—that is, synthesizing missing audio segments by extracting related visual cues about the scene and the source location—can lead to learning useful audio-visual spatial correspondences. To validate our hypothesis, we propose a novel feature-learning task.

Formally, we consider an egocentric video clip $C = (V, A)$, where V and A refer to the visual and binaural audio streams, respectively. See Fig. 2 left. The visual clip V comprises T frames, such that $V = \{V_1, \dots, V_T\}$. We generate a set of visual tokens $\hat{V}$ by splitting V into P non-overlapping tubelets, such that $\hat{V} = \{\hat{V}_1, \dots, \hat{V}_P\}$, where $\hat{V}_k$ denotes the k^{th} tubelet consisting of a contiguous sequence of patches spanning all T frames. We represent the binaural audio A as two Mel-spectrograms [30], $A = \{A^L, A^R\}$, where A^L and A^R are the spectrograms for the left and right channels, respectively. We create a set of audio tokens $\hat{A}$ by splitting A into Q non-overlapping patches, such that $\hat{A} = \{\hat{A}_1, \dots, \hat{A}_Q\}$.

Next, we mask a portion of the audio tokens in $\hat{A}$ and obtain complementary subsets of masked and unmasked tokens, $\hat{A}^M$ and $\hat{A}^U$, respectively, where $\hat{A}^M = \{\ddot{A}_1, \dots, \ddot{A}_S\}$, $\hat{A}^U = \{\bar{A}_1, \dots, \bar{A}_{Q-S}\}$, and S is the number of masked tokens. Given $\{\hat{V}, \hat{A}^M, \hat{A}^U\}$, we aim to learn a self-supervised model $\mathcal{F}$ comprising an encoder $\mathcal{E}$ and decoder $\mathcal{D}$, such that $\mathcal{F} = \mathcal{D} \circ \mathcal{E}$ and $\mathcal{F}(\hat{V}, \hat{A}^U) = \tilde{A}^M$, where $\tilde{A}^M$ is an estimate of the masked audio tokens in $\hat{A}^M$. By training on this pretext task, our encoder $\mathcal{E}$ can learn rich audio-visual spatial correspondences that can be leveraged for multiple downstream tasks that require the synergy of vision and spatial audio, as we show in results.

In our method (see Fig. 2), $\mathcal{E}$ (Sec. 3.1) is an audio-visual (AV) spatial correspondence encoder that learns an implicit representation of the spatial relationships between the visual and unmasked binaural audio tokens, while $\mathcal{D}$ (Sec. 3.2) is an audio-visual decoder for binaural audio inpainting that uses this implicit representation to synthesize the masked audio tokens. We also devise a simple yet novel masking protocol (Sec. 3.3) for our inpainting task, which mixes masking random audio tokens with masking a full audio channel, and helps the model learn stronger audio-visual spatial associations that facilitate the downstream social tasks (Sec. 3.5). We train $\mathcal{F}$ to minimize the prediction error in the masked audio tokens (Sec. 3.4). Next, we describe our model design, audio masking protocol, training objective and network architecture, and downstream tasks.

3.1. Audio-visual spatial correspondence encoder

The audio-visual spatial correspondence encoder $\mathcal{E}$ (Fig. 2 center) extracts features from the visual and unmasked audio tokens $\{\hat{V}, \hat{A}^U\}$. It begins by embedding the visual and audio tokens using separate transformer encoders [15] for individually capturing the spatio-temporal features in the two modalities. Next, it uses a shared transformer encoder to jointly encode the audio and visual features, and produces a multi-modal representation suitable for binaural audio inpainting. We describe the individual encoders next.

Video and audio encoders. We first encode the visual tokens $\hat{V}$ using a linear layer to generate visual features v, such that $v = \{v_1, \dots, v_P\}$. We encode the audio tokens $\hat{A}^U$ with another linear layer to produce audio features a, such that $a = \{a_1, \dots, a_{Q-S}\}$, where S is the number of masked tokens out of a total of Q audio tokens (cf. Sec. 3). For each visual feature v_j, we add a sinusoidal positional embedding p^V_j [67] to it, where p^V_j captures cues about the 3D position of the j^{th} tubelet in the visual clip V. For an audio feature a_i, we add a sinusoidal positional embedding p^A_i and a learnable channel embedding $c \in \{c_L, c_R\}$ to it to convey information about the 2D location of the i^{th} unmasked audio token in the spectrogram and also the audio channel to which it belongs. Next, we feed the transformed visual and audio features to separate transformer encoders, $\mathcal{E}^V$ and $\mathcal{E}^A$, respectively, and obtain visual features $e^V = \{e^V_1, \dots, e^V_P\}$ and audio features $e^A = \{e^A_1, \dots, e^A_{Q-S}\}$.

Shared audio-visual encoder. Given the visual features e^V and audio features e^A, we concatenate them into e^{AV}, such that $e^{AV} = \{e^V_1, \dots, e^V_P, e^A_1, \dots, e^A_{Q-S}\}$, and re-add the sinusoidal positional embeddings p^V and p^A to the features of the respective modalities in e^{AV}. Furthermore, unlike existing audio-visual masked auto-encoders [23, 24, 29], we add the channel embeddings c to the audio features, and learnable modality embeddings $m \in \{m_A, m_V\}$ to all features in e^{AV} to help the model distinguish between the visual and audio modalities. Next, a shared audio-visual transformer $\mathcal{E}^{AV}$ encoder takes e^{AV} as input and outputs audio-visual features f^{AV}, which implicitly hold spatio-temporal information required for accurate inpainting of audio.

3.2. Audio-visual decoder for binaural audio inpainting

Our audio-visual decoder $\mathcal{D}$ (Fig. 2 right) takes f^{AV} as input and attempts to synthesize the masked binaural audio tokens by leveraging spatio-temporal cues in f^{AV}. It first projects f^{AV} to a lower-dimensional feature set g^{AV}. It then appends a learnable embedding for the masked audio tokens to g^{AV} and passes it through a shared audio-visual transformer decoder [26]. Next, it feeds the audio feature outputs of the shared decoder to another transformer decoder and uses its outputs to predict an estimate of the masked binaural audio tokens. The decoders are lightweight compared to the encoders, ensuring that the encoders are primarily responsible for driving the inpainting task and producing good audio-visual features for strong downstream performance. We next describe each component of $\mathcal{D}$ in detail.

Shared audio-visual decoder. We first create a lower-dimensional projection g^{AV} of the audio-visual encodings f^{AV} by passing it through a linear layer, and append a learnable embedding ϕ corresponding to each of the S masked audio tokens to g^{AV}. Next, we add the positional embeddings p^V and p^A, the audio channel embeddings c, and the modality embeddings m to g^{AV}, and feed it to a shallow transformer decoder $\mathcal{D}^{AV}$ that outputs an audio-visual feature set h^{AV}. We then take the audio features h^A from h^{AV} and pass them to the audio decoder for further processing.

Audio decoder. The audio decoder $\mathcal{D}^A$ re-adds the positional embeddings p^A and channel embeddings c to g^A, and feeds it to a transformer decoder, which outputs audio features d^A.

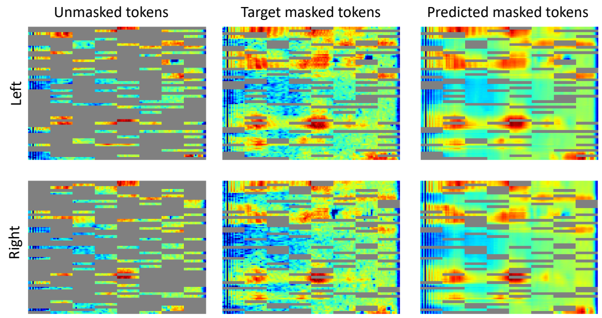
Prediction of masked audio tokens. Finally, we take the subset of all audio features d^A corresponding to the masked audio tokens $\hat{A}^M$, upsample them by passing through a linear layer, and reshape them to obtain an estimate $\tilde{A}^M$ of the masked tokens $\hat{A}^M$, such that $\tilde{A}^M = \{\tilde{A}_1, \dots, \tilde{A}_S\}$.

3.3. Audio masking

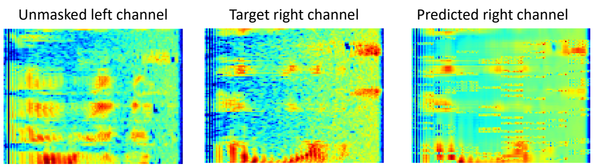
Different from other masked auto-encoding counterparts [23, 24, 29], we design an audio masking protocol that is customized to help our model better extract spatial audio-visual cues during self-supervised pretraining. In particular, we mix the strategy of randomly masking a full audio channel with that of randomly masking audio tokens using a hyperparameter r during training, where r represents the probability with which we randomly drop a full audio channel and r is sampled from a uniform distribution $U(0, 1)$:

$$\text{mask}(\hat{A}) = \begin{cases} \hat{A}^M = A^L \text{ or } \hat{A}^M = A^R & \text{if} \quad x \sim U(0,1) \leq r \\ \hat{A}^M \subseteq \{\hat{A}_1, \dots, \hat{A}_Q\} & \text{otherwise} \end{cases}$$

On the one hand, *token masking* could lead to tokens from the same location in the two audio channels being present among the unmasked tokens, providing additional spatial cues to the model and resulting in a simpler, stabler optimization objective for the inpainting task. In addition, since token masking involves masking a short span in both frequency and time domains, the model can rely more on local audio-visual cues and tolerate the global noise in both the visual and audio streams due to a camera-wearer's motion. On the other hand, *channel masking* forces the model to solve a more challenging binauralization task solely on the basis of vision,



(a) Token masking



(b) Channel masking

Figure 3. Masked targets and predictions shown alongside the unmasked inputs for (a) token masking and (b) channel masking. Our predictions accurately capture the global patterns in the target spectrograms, which depend on the scene's spatial properties.

which could help it learn even stronger spatial features. This encourages the model to reason about the camera motion at a more global scale (over the entire clip span). Towards achieving high performance on the downstream tasks, we aim to strike a fine balance between these two strategies and combine the benefits of reasoning at both temporal scales.

3.4. Training objective and network architecture

We train our model to minimize the error in predicting the masked audio tokens. In particular, we compute the mean-squared error $\mathcal{L}$ averaged over all masked audio tokens:

$$\mathcal{L} = \frac{1}{S} \sum_{i=1 \dots S} ||\ddot{A}_i - \tilde{a}_i||_2^2. \qquad (1)$$

We visualize our predicted audio tokens in Fig. 3 for the cases of token (Fig. 3a) and channel (Fig 3b) masking. Our model is able to accurately predict the masked targets and capture the global patterns in the spectrograms, which are often determined by the spatial audio-visual cues captured from of the scene (visual input not shown in Fig 3 for brevity), thereby further emphasizing our model's ability to learn useful spatial features.

Our uni-modal encoders, $\mathcal{E}^A$ and $\mathcal{E}^V$, have 8 layers, while the audio-visual encoder $\mathcal{E}^{AV}$ has 6 layers. All encoders have 12 attention heads and use 768-dimensional hidden embeddings. The audio-visual decoder $\mathcal{D}^{AV}$ and audio-only decoder $\mathcal{D}^A$ have 1 and 3 layers, respectively. Both decoders have 6 attention heads and use 384-dimensional hidden embeddings. To pretrain our model, we set the relative frequency of dropping an audio channel in our masking protocol for training to $r = 20\%$ based on disjoint validation data (see Supp.). We train our model for 200 epochs using the AdamW [37] optimizer with a weight decay of 10^{-5}, and a learning rate scheduler that reaches a peak learning rate of 2×10^{-4} over 10 warmup epochs, and then decays it through half-cycle cosine annealing [36]. For data augmentation, we perform random flipping of audio channels and video clips along the frame width. For downstream evaluation, we fuse the features from the audio-visual encoder with the backbones for downstream tasks, and finetune them. See Supp. for further details on architecture and training.

3.5. Downstream tasks

We explore two downstream tasks with our pretrained features: active speaker detection and spatial audio denoising. Active speaker detection (ASD) involves matching an audio clip with an appropriate face track from the corresponding video clip, i.e., answering "which person is speaking now?". While current SOTA methods [42, 65] rely on semantic similarities between monaural audio and vision to solve this task, we hypothesize that leveraging spatial audio can additionally reveal the sound source location in the video. In spatial audio denoising, the goal is to separate the target audio from distractors. In particular, we aim to remove the audio from sources extraneous to the conversation—out-of-view sounds from other parts of the scene. We detail the backbone models for each in the next section.

4. Experiments

We validate our learned representations on two downstream tasks and two datasets, and we compare with prior models for spatial audio-visual feature learning [18, 23, 60, 70], as well as various baselines and ablations.

Datasets. We train and evaluate our model on two challenging egocentric datasets containing video and binaural audio of people having conversations: **1)** EgoCom [49], and **2)** EasyCom [11], detailed in Supp. To our knowledge, these are the only two publicly available datasets offering binaural audio with conversations in video, whether exocentric or egocentric. In particular, Ego4D [25] and EPIC [9] do not comprise social scenarios and are not applicable. Whereas EasyCom primarily has participants sitting around a table and talking, EgoCom has videos of participants moving around a room, turning their face and body, standing up, etc. They test our method's robustness in diverse scenarios of varying difficulty.

4.1. Active speaker detection

We first evaluate on active speaker detection (ASD).

Backbone models. We consider two SOTA ASD models as the backbones for leveraging our pretrained representations: **1)** TalkNet [65], and **2)** SPELL [42]. TalkNet encodes a face track and an audio clip using attention for learning intra- and inter-modal semantic and temporal features. Next, it fuses these features and performs binary classification to predict if the face in the track is active. SPELL extracts audio-visual features for each face in a clip using ResNets [3], and learns long and short-term semantic relations among them using a graph neural network. Finally, it performs binary classification of these features for predicting active speakers.

Pretrained feature fusion. To fuse our pretrained features with the ASD backbones, we use a transformer decoder that cross-attends to the feature outputs of the shared encoder $\mathcal{E}^{AV}$ using sinsuoidal embedding as queries, with each embedding representing a clip frame index. Next, we append the decoder outputs to the cross-attention outputs in TalkNet, or the audio-visual encoder outputs in SPELL, frame by frame. In essence, while the original audio-visual encoders leverage *semantic* correlations between vision and audio, our features provide strong complementary *spatial* cues.

Baselines. For both TalkNet and SPELL, we compare against multiple baselines comprising both the unmodified backbone and improved versions of it, in addition to some naive methods:

- **All-active:** a naive model that predicts that all visible faces are always active
- **All-inactive:** a naive model that predicts that all visible faces are always inactive
- **Random:** a naive model that emits a random ASD confidence score for every visible speaker
- **Backbone w/o audio:** a vision-only version of the backbone with no audio input
- **Backbone:** the originally-proposed backbone that processes only faces and monaural audio
- **Backbone-binaural:** an improved backbone using binaural audio instead of monaural, alongside positional encodings for the faces, indicative of their relative position and depth, for better matching the face to the audio
- **Backbone-binaural w/ scene:** a further improvement over the backbone, where we also provide the scene images (uncropped video frames) to backbone-binaural
- **Backbone w/ TLR [70]:** fuses features from the SOTA Telling Left from Right (TLR) [70], which learns audio-visual spatial correspondences by predicting the spatial alignment between vision and binaural audio
- **Backbone w/ 2.5D-VS [18]:** fuses features from the SOTA audio-visual binauralization model, 2.5D Visual Sounds (2.5D-VS) [18]
- **Backbone w/ 2.5D-VS [18]++:** fuses features from 2.5D-VS with a transformer architecture
- **Backbone w/ SSR [60]++:** fuses features from the SOTA self-supervised audio-visual stereo sound ranking (SSR) [60] model with a transformer architecture

Model	TalkNet [65] *EgoCom*	TalkNet [65] *EasyCom*	SPELL [42] *EgoCom*	SPELL [42] *EasyCom*
No pretraining				
All-active	32.9	30.1	32.9	30.1
All-inactive	32.9	30.1	32.9	30.1
Random	30.8	28.0	30.8	28.0
B w/o audio	41.5	50.1	60.4	63.2
B	52.8	45.7	60.9	59.0
B-binaural	60.0	59.6	63.1	60.3
B-binaural w/ scene	60.8	66.9	61.2	61.4
Alternate pretraining methods				
B w/ TLR [70]	47.9	59.3	61.3	61.7
B w/ 2.5D-VS [18]	57.7	63.7	61.2	59.7
B w/ 2.5D-VS [18]++	63.4	68.3	65.1	64.5
B w/ SSR [60]++	61.2	70.6	61.2	67.4
B w/ AV-MAE [23]	61.1	61.3	64.4	65.2
Ours	**63.9**	**71.8**	**65.6**	**70.2**
Ours w/o pretrain	62.7	62.9	—	—
Ours w/ pretrain monaural	61.0	69.4	63.9	69.0

Table 1. Mean average precision (%) for active speaker detection with TalkNet [65] and SPELL [42] backbones on both datasets. Higher is better. 'B' refers to backbone. Note that SPELL requires storing pretrained features in the graph nodes; therefore it does not allow training from scratch.

- **Backbone w/ AV-MAE [23]:** fuses features from the SOTA modality-inpainting AV-MAE [23] model

For all alternate feature-learning methods [18, 23, 60, 70], we pretrain them on our datasets and use our feature fusion method. Thus, any advantages in performance of our approach over these SOTA representations will be attributable to our modeling ideas. Importantly, the 2.5D-VS [18]++, SSR [60]++, and AV-MAE [23] features all rely on transformers and have similar model capacity as ours; see Supp. for a detailed analysis on model capacity. We use the standard **mean average precision** (mAP) metric.

Results. Table 1 (top) reports our ASD results on both val and test splits. The three naive baselines perform poorly on both EgoCom [49] and EasyCom [11], emphasizing the difficulty of the task. For both TalkNet [65] and SPELL [42], the unchanged backbone model generally performs better than the model without audio, showing that both vision and audio are required. Upgrading from monaural to binaural audio further boosts performance, as the model can now leverage both spatial and semantic information. Additionally using scene features lets the backbone explicitly match the scene area around the inferred source location with the face, and further improves ASD, especially for EgoCom, where the background scene changes more often.

Among alternate feature learning methods, 2.5D-VS [18]++, SSR [60]++ and AV-MAE [23] consistently outperform TLR [70] and 2.5D-VS [17], and also the basic and enhanced backbones, showing that self-attention and higher model capacity enhance feature quality. Besides, 2.5D-VS

Model	SI-SDRi ↑	STFT ↓
No pretraining		
B w/o vision	1.61	7.36
B	1.46	7.27
B w/ ImageNet features	1.48	6.95
Alternate pretraining methods		
B w/ TLR [70]	1.41	7.79
B w/ 2.5D-VS [18]	1.67	7.34
B w/ 2.5D-VS [18]++	2.11	6.60
B w/ SSR [60] ++	2.04	6.70
B w/ AV-MAE [23]	2.07	6.62
Ours	**2.20**	**6.51**
B w/o pretrain	1.90	7.25
B w/ pretrain monaural audio	2.00	6.75

Table 2. Audio denoising with U-Net [70] backbone (referred to as 'B' in table) for 0 dB noise (maximum). See Supp. for varying noise levels. All STFT distance measures use base 10^{-3}.

outperforms TLR, and 2.5D-VS++ and AV-MAE generally outperform SSR++, indicating that objectives that promote reasoning directly at the spectrogram level improve results.

Our model substantially outperforms all baselines—including the SOTA AV representation learning methods—for both backbones (TalkNet and SPELL) on both datasets. This shows that our method learns stronger spatial features that are both backbone- and dataset-agnostic. In contrast, methods developed for exocentric settings with more stationary cameras (such as TLR and 2.5D-VS) rely more on the global visual context and seem to struggle in our setting, where the camera moves frequently and the sound source leaves the field of view. Finally, our improvement over 2.5D-VS++, SSR++ and AV-MAE, which use similar encoders as ours, disentangles the benefits of our masking strategy and model design from those of the model capacity.

Model analysis. Table 1 (bottom) shows ablations of our method. Upon training for ASD from scratch, we see a sharp drop in performance, showing that our advantage is not solely due to our model design, but also our self-supervised pretraining stage. Performance also declines upon pretraining with monaural audio instead of binaural, showing that our model goes beyond learning semantic features and successfully captures spatial features useful for ASD.

4.2. Spatial audio denoising

Next we evaluate spatial audio denoising on EgoCom.[2] To instantiate this task, we add the binaural audio of a target clip with the downscaled binaural audio from another randomly chosen clip, where the downscaling factor depends on the desired noise level. The goal is to extract the target sound from the mixture. We evaluate three noise levels, expressed using the signal-to-noise (SNR) ratio: **1)** 0 dB, **2)** 2.5 dB, and **3)** 5 dB. The different noise levels test our model's robustness to varying levels of task difficulty—the lower the SNR value, the higher the noise and difficulty.

[2]For EasyCom, the task setup is ill-posed for all models because mixing audio from a different EasyCom clip as noise leads to spatially overlapping sound sources, since all clips in the dataset are recorded at the same physical location (people sitting around the same table in the same room).

Backbone model. We adopt the commonly used U-Net [57] for audio-visual source separation [18, 70] as the backbone, which produces a binaural ratio mask for the target audio (see Supp. for details). We multiply the predicted ratio mask with the mixed magnitude spectrogram to get the predicted magnitude spectrogram, then convert it to a waveform using inverse short-time Fourier transform with the mixed audio phase.

Pretrained feature fusion. To use our features for denoising, we reshape the visual features and unmasked audio features produced by our audio-visual encoder $\mathcal{E}^{AV}$ to form multi-channel 2D maps, where the features align with their corresponding tokens vis-a-vis the raster order. Next, we pass the feature maps to separate convolutional layers, concatenate the outputs channel-wise, and use them to replace the visual features at the U-Net [70] bottleneck.

Baselines. We compare against the following baselines and existing methods:
- **U-Net w/o vision:** an audio-only blind denoising model
- **U-Net:** the original backbone without any alterations
- **U-Net w/ ImageNet:** pretrains the visual encoder on ImageNet [10]
- **U-Net w/ TLR [70]:** fuses features from TLR [70]
- **U-Net w/ 2.5D-VS [18]:** fuses pretrained features from 2.5D-VS [18]
- **U-Net w/ 2.5D-VS [18]++:** fuses features from the transformer-based version of 2.5D-VS
- **U-Net w/ SSR [60]++:** fuses features from the transformer-based version of SSR [60]
- **U-Net w/ AV-MAE [60]:** fuses features from the modality inpainting AV-MAE [23] model

Evaluation metric. For evaluating our denoising quality, we use standard metrics: **1) STFT** distance (the L2 error between the predicted and ground-truth spectrograms) expressed using base 10^{-3} and **2) SI-SDRi**: the improvement in SI-SDR [58], a scale-invariant estimate of the distortion in the audio, over using the mixed audio as the prediction.

Results. Table 2 (top) shows spatial audio denoising results on the challenging EgoCom dataset with 0 dB, the most difficult noise level. See Supp. for similar results with the other noise levels. The unchanged U-Net backbone lowers the STFT distance compared to the version that lacks vision,

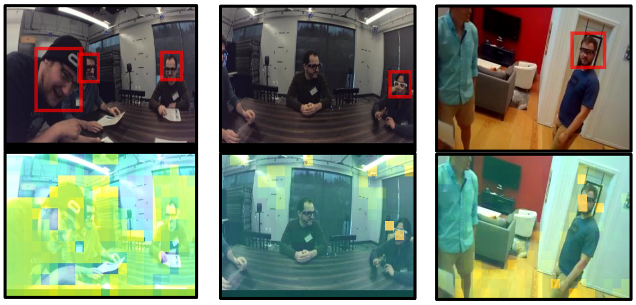


Figure 4. Heat maps showing the image areas our model's AV encoder attends to, placed alongside the images. Brighter yellow means higher attention score. Our model attends to image regions (*e.g*. faces of speakers, sound-reflecting flat regions like floor and table, etc.) that strongly determine the spatial properties of the audio, including direct sources of sound (marked in red).

showing that like ASD, vision is crucial for better denoising. Using pretrained features of 2.5D-VS [18] (++), SSR [60]++ or AV-MAE [23] further improves performance, showing that learning spatial audio-visual features aids denoising.

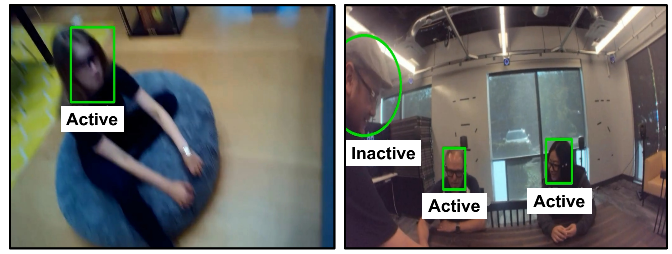
Our method outperforms all baselines ($p \leq 0.05$) across both metrics. While the improvement over the baselines that do not use self-supervised pretraining emphasizes the utility of learning spatial audio-visual relationships through self-supervision, the gain over other pretraining methods underlines the strengths of our self-supervised method design—which are consistently realized for both ASD and denoising. Further, our performance margins are larger for higher noise levels (0 and 2.5 dB), indicating that our features play a bigger role in the more difficult denoising settings.

Model analysis. In Table 2 (bottom), we ablate our pretraining method. Similar to ASD, training from scratch on the denoising task hurts performance. This disentangles the impact of our pretext task design from the model architecture and shows that our pretraining stage helps the backbone with learning better audio-visual features, leading to superior denoising quality. Furthermore, pretraining with monaural audio also degrades performance, re-emphasizing that our method is not restricted to learning semantic features—in contrast to prior work [23, 24, 29].

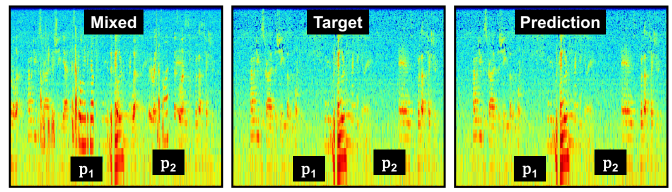
See Supp. for analysis of the effect of alternate masking choices, multi-level positional embeddings, task-specific backbones, and our finetuning strategy on performance.

4.3. Qualitative analysis

In Fig. 4, we analyze the visual attention maps of our shared encoder $\mathcal{E}^{AV}$. Note that the regions of high attention are not only limited to the direct sound sources (*e.g*., regions in and around faces of active speakers across examples), but also include large sound-reflecting objects (*e.g*., the flat surface of the table on the left; the walls on the left and in the middle; the floor on the right, etc.) that determine how sound



(a) Our model succeeds on ASD even with fast camera motion (left; note image blur), multiple active speakers (right) and partially visible faces (left and right)



(b) Our model denoises accurately—note the noise patches in mixed audio above points p_1 and p_2, which are successfully removed in our prediction.

Figure 5. Success cases for ASD (a) and denoising (b)

spatializes through early reflections, late reverberations, etc. Interestingly, our model also attends to multiple people if they are speaking at the same time (see left), thereby facilitating the detection of multiple active speakers. See Supp. for additional visualizations showing how, depending on the scene's spatial layout, our model uses one audio channel more than the other to attend to important image locations.

In Fig. 5, we qualitatively show our model's success cases. On ASD (Fig. 5a), our model can tackle drastic camera movements, multiple active speakers, and partially visible faces. On denoising (Fig. 5b), our model is able to remove interferences from distractor sources, and make predictions that closely match the ground truth in spectrogram structure.

We also observe some limitations. Our model's performance on ASD declines when there are drastic movements of the camera wearer, or there is a high overlap in speech from different conversation participants. On denoising, our model struggles when the noisy audio is semantically and acoustically similar to the target, or when it cannot extract spatial cues due to occlusions or out-of-view speakers. Refer to our Supp. video for both success and failure cases.

5. Conclusion

We introduce a novel self-supervised approach for learning audio-visual representations in social egocentric videos via spatial correspondence between the video and its binaural audio. Through extensive evaluation, we show that our learned features are strong and generic enough to improve over multiple backbone methods on multiple downstream tasks. In future work, we will explore how the learned spatial audio-visual cues may reveal the social attention of speakers.

Acknowledgements: UT Austin is supported in part by NSF CCRI and the IFML NSF AI Institute. KG is paid as a research scientist by Meta, and SM is a visiting researcher at Meta.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. *arXiv preprint arXiv:1804.04121*, 2018. 3

[2] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 208–224. Springer, 2020. 1, 2

[3] Juan Leon Alcazar, Fabian Caba Heilbron, Long Mai, Federico Perazzi, Joon-Young Lee, Pablo Arbeláez, and Bernard Ghanem. Active speakers in context. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12462–12471, 2020. 2, 6

[4] Relja Arandjelović and Andrew Zisserman. Look, listen and learn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 609–617, 2017. 2

[5] Alan Baade, Puyuan Peng, and David F. Harwath. Mae-ast: Masked autoencoding audio spectrogram transformer. In *Interspeech*, 2022. 2

[6] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 17–36. Springer, 2020. 2

[7] Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual acoustic matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18858–18868, 2022. 2

[8] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. *Advances in Neural Information Processing Systems*, 35:8896–8911, 2022. 2

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 1, 5

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 7

[11] Jacob Donley, Vladimir Tourbabin, Jung-Suk Lee, Mark Broyles, Hao Jiang, Jie Shen, Maja Pantic, Vamsi Krishna Ithapu, and Ravish Mehra. Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments. 2021. 1, 2, 5, 6

[12] Ngoc QK Duong, Emmanuel Vincent, and Rémi Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1830–1840, 2010. 3

[13] 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. *arXiv preprint arXiv:1804.03619*, 2018. 3

[14] Mark Everingham, Josef Sivic, and Andrew Zisserman. Hello! my name is... buffy" – automatic naming of characters in tv video. In *British Machine Vision Conference*, 2006. 2

[15] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022. 2, 4

[16] John W Fisher III, Trevor Darrell, William Freeman, and Paul Viola. Learning joint statistical models for audio-visual fusion and segregation. In *Advances in Neural Information Processing Systems*. MIT Press, 2000. 3

[17] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7053–7062, 2019. 2, 6

[18] Ruohan Gao and Kristen Grauman. 2.5D visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2019. 2, 5, 6, 7, 8

[19] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3879–3888, 2019. 3

[20] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. *arXiv preprint arXiv:2101.03149*, 2021.

[21] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53, 2018. 3

[22] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 658–676. Springer, 2020. 2

[23] Mariana-Iuliana Georgescu, Eduardo Fonseca, Radu Tudor Ionescu, Mario Lucic, Cordelia Schmid, and Anurag Arnab. Audiovisual masked autoencoders. *arXiv preprint arXiv:2212.05922*, 2022. 1, 2, 4, 5, 6, 7, 8

[24] Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R. Glass. Contrastive audio-visual masked autoencoder. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 4, 8

[25] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1, 5

[26] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 4

[27] John Hershey and Javier Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *Advances in Neural Information Processing Systems*. MIT Press, 1999. 2

[28] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis. Deep learning for monaural speech separation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1562–1566, 2014. 3

[29] Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Haoqi Fan, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, and Christoph Feichtenhofer. Mavil: Masked audio-video learners. *arXiv preprint arXiv:2212.08071*, 2022. 1, 2, 4, 8

[30] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. *Advances in Neural Information Processing Systems*, 35:28708–28720, 2022. 2, 3

[31] Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. Epic-sounds: A large-scale dataset of actions that sound. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1

[32] Hao Jiang, Calvin Murdock, and Vamsi Krishna Ithapu. Egocentric deep multi-channel audio-visual active speaker localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10544–10552, 2022. 2, 3

[33] Okan Köpüklü, Maja Taseska, and Gerhard Rigoll. How to design a three-stage architecture for audio-visual active speaker detection in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1193–1203, 2021. 2

[34] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems*, 31, 2018. 1, 2

[35] Juan Le'on-Alc'azar, Fabian Caba Heilbron, Ali K. Thabet, and Bernard Ghanem. Maas: Multi-modal assignation for active speaker detection. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 265–274, 2021. 2

[36] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. 5

[37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 5

[38] Sagnik Majumder and Kristen Grauman. Active audio-visual separation of dynamic sound sources. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*, pages 551–569. Springer, 2022. 2

[39] Sagnik Majumder, Ziad Al-Halah, and Kristen Grauman. Move2hear: Active audio-visual source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 275–285, 2021. 2

[40] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics. In *Advances in Neural Information Processing Systems*, 2022. 2

[41] Sagnik Majumder, Hao Jiang, Pierre Moulon, Ethan Henderson, Paul Calamia, Kristen Grauman, and Vamsi Krishna Ithapu. Chat2map: Efficient scene mapping from multi-ego conversations. *arXiv preprint arXiv:2301.02184*, 2023. 2

[42] Kyle Min, Sourya Roy, Subarna Tripathi, Tanaya Guha, and Somdeb Majumdar. Learning long-term spatial-temporal graphs for active speaker detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 371–387. Springer, 2022. 2, 3, 5, 6

[43] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 218–234. Springer, 2022. 2

[44] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360°video. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018. 2

[45] Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Self-supervised generation of spatial audio for 360° video. *Advances in Neural Information Processing Systems*, 33:4733–4744, 2020. 2

[46] Giovanni Morrone, Daniel Michelsanti, Zheng-Hua Tan, and Jesper Jensen. Audio-visual speech inpainting with deep learning. In *ICASSP*, 2021. 2

[47] Kazuhiro Nakadai, Ken-ichi Hidai, Hiroshi G Okuno, and Hiroaki Kitano. Real-time speaker localization and speech separation by audio-visual integration. In *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292)*, pages 1043–1049. IEEE, 2002. 3

[48] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and A. Ng. Multimodal deep learning. In *International Conference on Machine Learning*, 2011. 2

[49] Curtis Northcutt, Shengxin Zha, Steven Lovegrove, and Richard Newcombe. Egocom: A multi-person multi-modal egocentric communications dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 1, 2, 5, 6

[50] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 1, 2, 3

[51] 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. 2

[52] 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*, 2016. 1, 2

[53] S. Parekh, S. Essid, A. Ozerov, N. Q. K. Duong, P. Pérez, and G. Richard. Motion informed audio source separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6–10, 2017. 3

[54] Jie Pu, Yannis Panagakis, Stavros Petridis, and Maja Pantic. Audio-visual object localization and separation using low-rank and sparsity. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2901–2905. IEEE, 2017. 3

[55] Senthil Purushwalkam, Sebastia Vicenc Amengual Gari, Vamsi Krishna Ithapu, Carl Schissler, Philip Robinson, Abhinav Gupta, and Kristen Grauman. Audio-visual floorplan reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1183–1192, 2021. 2

[56] Kranthi Kumar Rachavarapu, Aakanksha, Vignesh Sundaresha, and A. N. Rajagopalan. Localize to binauralize: Audio spatialization from visual sound source localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1930–1939, 2021. 2

[57] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 7

[58] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. Sdr – half-baked or well done? *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630, 2018. 7

[59] Fiona Ryan, Hao Jiang, Abhinav Shukla, James M Rehg, and Vamsi Krishna Ithapu. Egocentric auditory attention localization in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14663–14674, 2023. 3

[60] Tomoya Sato, Yusuke Sugano, and Yoichi Sato. Self-supervised learning for audio-visual relationships of videos with stereo sounds. *IEEE Access*, 10:94273–94284, 2022. 2, 5, 6, 7, 8

[61] Rahul Sharma, Krishna Somandepalli, and Shrikanth S. Narayanan. Crossmodal learning for audio-visual speech event localization. *ArXiv*, abs/2003.04358, 2020. 2

[62] Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *CoRR*, abs/1804.09626, 2018. 1

[63] Paris Smaragdis, Bhiksha Raj, and Madhusudana Shashanka. Supervised and semi-supervised separation of sounds from single-channel mixtures. In *Independent Component Analysis and Signal Separation*, pages 414–421, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. 3

[64] Martin Spiertz and Volker Gnann. Source-filter based clustering for monaural blind source separation. In *in Proceedings of International Conference on Digital Audio Effects DAFx'09*, 2009. 3

[65] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3927–3935, 2021. 2, 3, 5, 6

[66] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022. 2

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[68] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Trans. On Audio, Speech and Lang. Processing*, 2007. 3

[69] Shanshan Wang, Archontis Politis, Annamaria Mesaros, and Tuomas Virtanen. Self-supervised learning of audio representations from audio-visual data using spatial alignment. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1467–1479, 2022. 2

[70] Karren Yang, Bryan Russell, and Justin Salamon. Telling left from right: Learning spatial correspondence of sight and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9932–9941, 2020. 2, 5, 6, 7

[71] Karren Yang, Michael Firman, Eric Brachmann, and Clément Godard. Camera pose estimation and localization with active audio sensing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 271–291. Springer, 2022. 2

[72] Yuanhang Zhang, Susan Liang, Shuang Yang, Xiao Liu, Zhongqin Wu, Shiguang Shan, and Xilin Chen. Unicon: Unified context network for robust active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3964–3972, 2021. 2

[73] Hang Zhou, Ziwei Liu, Xudong Xu, Ping Luo, and Xiaogang Wang. Vision-infused deep audio inpainting. In *ICCV*, 2019. 2

[74] Özgür Yılmaz and Scott Rickard. Blind separation of speech mixtures via time-frequency masking. In *IEEE TRANSACTIONS ON SIGNAL PROCESSING (2002) SUBMITTED*, 2004. 3