Audio-Visual Scene Analysis with Self-Supervised Multisensory Features

Andrew Owens Alexei A. Efros

UC Berkeley

Abstract. The thud of a bouncing ball, the onset of speech as lips open — when visual and audio events occur together, it suggests that there might be a common, underlying event that produced both signals. In this paper, we argue that the visual and audio components of a video signal should be modeled jointly using a fused multisensory representation. We propose to learn such a representation in a self-supervised way, by training a neural network to predict whether video frames and audio are temporally aligned. We use this learned representation for three applications: (a) sound source localization, i.e. visualizing the source of sound in a video; (b) audio-visual action recognition; and (c) on/off-screen audio source separation, e.g. removing the off-screen translator's voice from a foreign official's speech. Code, models, and video results are available on our webpage: http://andrewowens.com/multisensory.

1 Introduction

As humans, we experience our world through a number of simultaneous sensory streams. When we bite into an apple, not only do we taste it, but — as Smith and Gasser [1] point out — we also hear it crunch, see its red skin, and feel the coolness of its core. The coincidence of sensations gives us strong evidence that they were generated by a common, underlying event [2], since it is unlikely that they co-occurred across multiple modalities merely by chance. These cross-modal, temporal co-occurrences therefore provide a useful learning signal: a model that is trained to detect them ought to discover multimodal structures that are useful for other tasks. In much of traditional computer vision research, however, we have been avoiding the use of other, non-visual modalities, arguably making the perception problem harder, not easier.

In this paper, we learn a temporal, multisensory representation that fuses the visual and audio components of a video signal. We propose to train this model without using any manually labeled data. That is, rather than explicitly telling the model that, e.g., it should associate moving lips with speech or a thud with a bouncing ball, we have it discover these audio-visual associations through self-supervised training [3]. Specifically, we train a neural network on a "pretext" task of detecting misalignment between audio and visual streams in synthetically-shifted videos. The network observes raw audio and video streams — some of which are aligned, and some that have been randomly shifted by a few seconds — and we task it with distinguishing between the two. This turns out to be a challenging training task that forces the network to fuse visual motion with audio information and, in the process, learn a useful audio-visual feature representation.

We demonstrate the usefulness of our multisensory representation in three audio-visual applications: (a) sound source localization, (b) audio-visual action recognition;

(a) Sound localization (b) Action recognition (c) On/off-screen audio separation

Fig. 1: Applications. We use self-supervision to learn an audio-visual representation that: (a) can be used to visualize the locations of sound sources in video; (b) is useful for visual and audio-visual action recognition; (c) can be applied to the task of separating on- and off-screen sounds. In (c), we demonstrate our source-separation model by visually masking each speaker and asking it to predict the on-screen audio. The predicted sound contains only the voice of the visible speaker. Please see our webpage for video results: http://andrewowens.com/multisensory.

and (c) on/off-screen sound source separation. Figure 1 shows examples of these applications. In Fig. 1(a), we visualize the sources of sound in a video using our network's learned attention map, i.e. the impact of an axe, the opening of a mouth, and moving hands of a musician. In Fig. 1(b), we show an application of our learned features to audio-visual action recognition, i.e. classifying a video of a chef chopping an onion. In Fig. 1(c), we demonstrate our novel on/off-screen sound source separation model's ability to separate the speakers' voices by visually masking them from the video.

The main contributions of this paper are: 1) learning a general video representation that fuses audio and visual information; 2) evaluating the usefulness of this representation qualitatively (by sound source visualization) and quantitatively (on an action recognition task); and 3) proposing a novel video-conditional source separation method that uses our representation to separate on- and off-screen sounds, and is the first method to work successfully on real-world video footage, e.g. television broadcasts. Our feature representation, as well as code and models for all applications are available online.

2 Related work

Evidence from psychophysics While we often think of vision and hearing as being distinct systems, in humans they are closely intertwined [4] through a process known as *multisensory integration*. Perhaps the most compelling demonstration of this phenomenon is the McGurk effect [5], an illusion in which visual motion of a mouth changes one's interpretation of a spoken sound[1]. Hearing can also influence vision: the timing of a sound, for instance, affects whether we perceive two moving objects to be colliding or overlapping [2]. Moreover, psychologists have suggested that humans

[1] For a particularly vivid demonstration, please see: https://www.youtube.com/watch?v=G-lN8vWm3m0 [6]

fuse audio and visual signals at a fairly early stage of processing [7,8], and that the two modalities are used jointly in perceptual grouping. For example, the McGurk effect is less effective when the viewer first watches a video where audio and visuals in a video are unrelated, as this causes the signals to become "unbound" (i.e. not grouped together) [9,10]. This multi-modal perceptual grouping process is often referred to as *audio-visual scene analysis* [11,7,12,10]. In this paper, we take inspiration from psychology and propose a self-supervised multisensory feature representation as a computational model of audio-visual scene analysis.

Self-supervised learning Self-supervised methods learn features by training a model to solve a task derived from the input data itself, without human labeling. Starting with the early work of de Sa [3], there have been many self-supervised methods that learn to find correlations between sight and sound [13,14,15,16]. These methods, however, have either learned the correspondence between static images and ambient sound [15,16], or have analyzed motion in very limited domains [14,13] (e.g. [14] only modeled drumstick impacts). Our learning task resembles Arandjelović and Zisserman [16], which predicts whether an image and an audio track are sampled from the same (or different) videos. Their task, however, is solvable from a single frame by recognizing semantics (e.g. indoor vs. outdoor scenes). Our inputs, by contrast, always come from the same video, and we predict whether they are aligned; hence our task requires motion analysis to solve. Time has also been used as supervisory signal, e.g. predicting the temporal ordering in a video [17,18,19]. In contrast, our network learns to analyze audio-visual actions, which are likely to correspond to salient physical processes.

Audio-visual alignment While we study alignment for self-supervised learning, it has also been studied as an end in itself [20,21,22] e.g. in lip-reading applications [23]. Chung and Zisserman [22], the most closely related approach, train a two-stream network with an embedding loss. Since aligning speech videos is their end goal, they use a face detector (trained with labels) and a tracking system to crop the speaker's face. This allows them to address the problem with a 2D CNN that takes 5 channel-wise concatenated frames cropped around a mouth as input (they also propose using their image features for self-supervision; while promising, these results are very preliminary).

Sound localization The goal of visually locating the source of sounds in a video has a long history. The seminal work of Hershey et al. [24] localized sound sources by measuring mutual information between visual motion and audio using a Gaussian process model. Subsequent work also considered subspace methods [25], canonical correlations [26], and keypoints [27]. Our model learns to associate motions with sounds via self-supervision, without us having to explicitly model them.

Audio-Visual Source Separation Blind source separation (BSS), i.e. separating the individual sound sources in an audio stream — also known as the *cocktail party* problem [28] — is a classic audio-understanding task [29]. Researchers have proposed many successful probabilistic approaches to this problem [30,31,32,33]. More recent deep learning approaches involve predicting an embedding that encodes the audio clustering [34,35], or optimizing a permutation invariant loss [36]. It is natural to also want to include the visual signal to solve this problem, often referred to as *Audio-Visual Source Separation*. For example, [37,25] masked frequencies based on their correlation with optical flow; [12] used graphical models; [27] used priors on harmonics; [38] used

a sparsity-based factorization method; and [39] used a clustering method. Other methods use face detection and multi-microphone beamforming [40]. These methods make strong assumptions about the relationship between sound and motion, and have mostly been applied to lab-recorded video. Researchers have proposed learning-based methods that address these limitations, e.g. [41] use mixture models to predict separation masks. Recently, [42] proposed a convolutional network that isolates on-screen speech, although this model is relatively small-scale (tested on videos from one speaker). We do on/off-screen source separation on more challenging internet and broadcast videos by combining our representation with a u-net [43] regression model.

Concurrent work Concurrently and independently from us, a number of groups have proposed closely related methods for source separation and sound localization. Gabbay et al. [44,45] use a vision-to-sound method to separate speech, and propose a convolutional separation model. Unlike our work, they assume speaker identities are known. Ephrat et al. [46] and Afouras et al. [47] separate the speech of a user-chosen speaker from videos containing multiple speakers, using face detection and tracking systems to group the different speakers. Work by Zhao et al. [48] and Gao et al. [49] separate sound for multiple visible objects (e.g. musical instruments). This task involves associating objects with the sounds they typically make based on their appearance, while ours involves the "fine-grained" motion-analysis task of separating multiple speakers. There has also been recent work on localizing sound sources using a network's attention map [50,51,52]. These methods are similar to ours, but they largely localize objects and ambient sound in static images, while ours responds to actions in videos.

3 Learning a self-supervised multisensory representation

We propose to learn a representation using self-supervision, by training a model to predict whether a video's audio and visual streams are temporally synchronized.

Aligning sight with sound During training, we feed a neural network video clips. In half of them, the vision and sound streams are synchronized; in the others, we shift the audio by a few seconds. We train a network to distinguish between these examples. More specifically, we learn a model $p(y \mid I, A)$ that predicts whether the image stream I and audio stream A are synchronized, by maximizing the log-likelihood:

$$\mathcal{L}(\theta) = \frac{1}{2}\mathbb{E}_{I,A,t}[\log(p_\theta(y = 1 \mid I, A_0)) + \log(p_\theta(y = 0 \mid I, A_t))], \tag{1}$$

where A_s is the audio track shifted by s secs., t is a random temporal shift, θ are the model parameters, and y is the event that the streams are synchronized. This learning problem is similar to noise-contrastive estimation [54], which trains a model to distinguish between real examples and noise; here, the noisy examples are misaligned videos.

Fused audio-visual network design Solving this task requires the integration of low-level information across modalities. In order to detect misalignment in a video of human speech, for instance, the model must associate the subtle motion of lips with the timing of utterances in the sound. We hypothesize that early fusion of audio and visual streams is important for modeling actions that produce a signal in both modalities. We therefore propose to solve our task using a 3D multisensory convolutional network (CNN) with an early-fusion design (Figure 2).

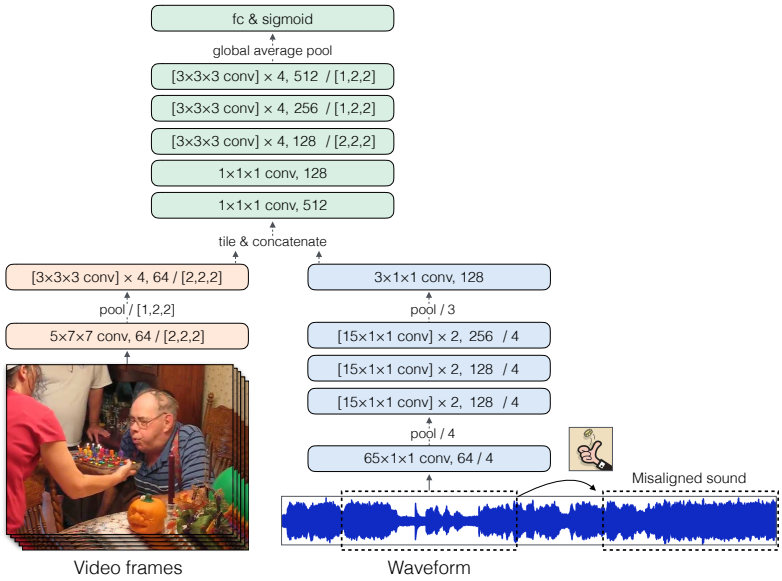

Fig. 2: Fused audio-visual network. We train an early-fusion, multisensory network to predict whether video frames and audio are temporally aligned. We include residual connections between pairs of convolutions [53]. We represent the input as a $T \times H \times W$ volume, and denote a stride by "/2". To generate misaligned samples, we synthetically shift the audio by a few seconds.

Before fusion, we apply a small number of 3D convolution and pooling operations to the video stream, reducing its temporal sampling rate by a factor of 4. We also apply a series of strided 1D convolutions to the input waveform, until its sampling rate matches that of the video network. We fuse the two subnetworks by concatenating their activations channel-wise, after spatially tiling the audio activations. The fused network then undergoes a series of 3D convolutions, followed by global average pooling [55]. We add residual connections between pairs of convolutions. We note that the network architecture resembles ResNet-18 [53] but with the extra audio subnetwork, and 3D convolutions instead of 2D ones (following work on inflated convolutions [56]).

Training We train our model with 4.2-sec. videos, randomly shifting the audio by 2.0 to 5.8 seconds. We train our model on a dataset of approximately 750,000 videos randomly sampled from AudioSet [57]. We use full frame-rate videos (29.97 Hz), resulting in 125 frames per example. We select random 224×224 crops from resized 256×256 video frames, apply random left-right flipping, and use 21 kHz stereo sound. We sample these video clips from longer (10 sec.) videos. Optimization details can be found in the supplementary material.

Task performance We found that the model obtained 59.9% accuracy on held-out videos for its alignment task (chance $= 50\%$). While at first glance this may seem low, we note that in many videos the sounds occur off-screen [15]. Moreover, we found that this task is also challenging for humans. To get a better understanding of human ability, we showed 30 participants from Amazon Mechanical Turk 60 aligned/shifted video pairs, and asked them to identify the one with out-of-sync sound. We gave them 15

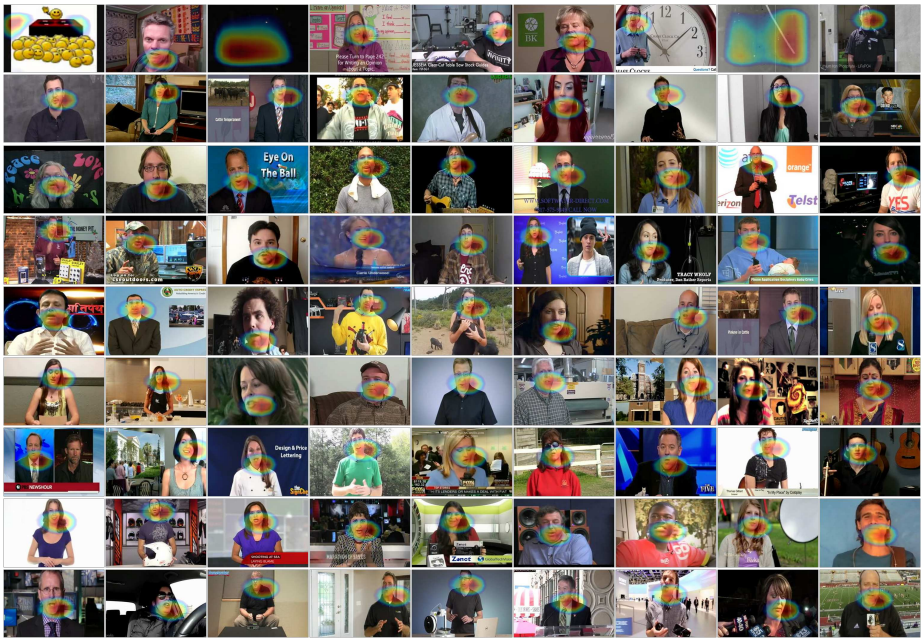

Fig. 3: Visualizing sound sources. We show the video frames in held-out AudioSet videos with the strongest class activation map (CAM) response (we scale its range per image to compensate for the wide range of values).

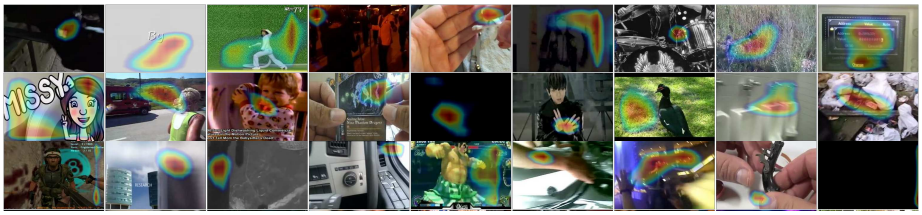

Fig. 4: Examples with the weakest class activation map response (c.f. Figure 3).

secs. of video (so they have significant temporal context) and used large, 5-sec. shifts. They solved the task with $66.6\% \pm 2.4\%$ accuracy.

To help understand what actions the model can predict synchronization for, we also evaluated its accuracy on categories from the Kinetics dataset [58] (please see the supplementary material). It was most successful for classes involving human speech: e.g., *news anchoring*, *answering questions*, and *testifying*. Of course, the most important question is whether the learned audio-visual representation is useful for downstream tasks. We therefore turn out attention to applications.

4 Visualizing the locations of sound sources

One way of evaluating our representation is to visualize the audio-visual structures that it detects. A good audio-visual representation, we hypothesize, will pay special attention to *visual sound sources* — on-screen actions that make a sound, or whose motion is highly correlated with the onset of sound. We note that there is ambiguity in the notion of a sound source for in-the-wild videos. For example, a musician's lips,

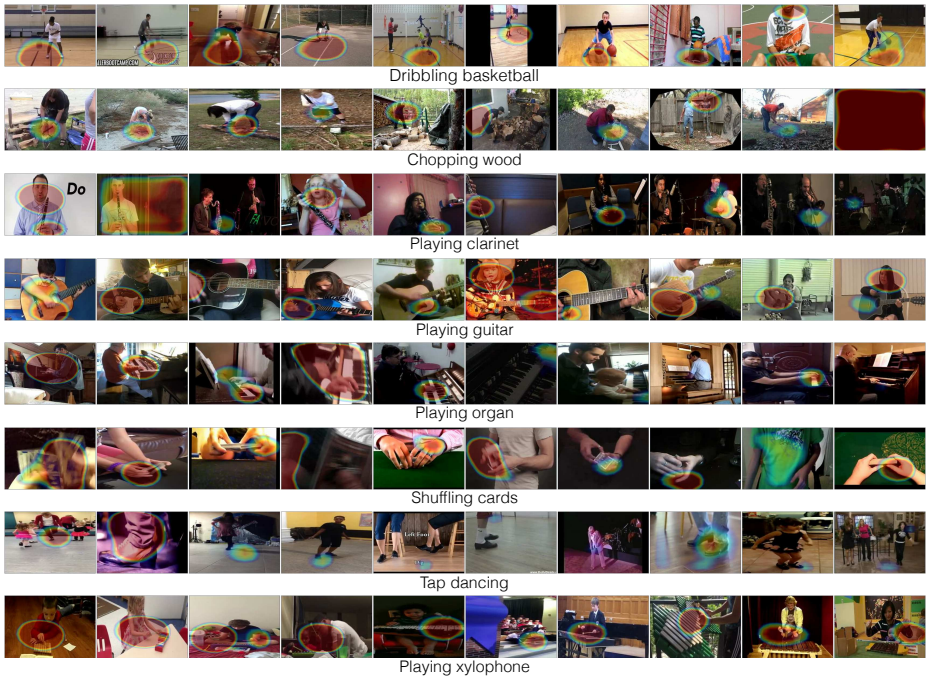

Fig. 5: Strongest CAM responses for classes in the Kinetics-Sounds dataset [16], after manually removing frames in which the activation was only to a face (which appear in almost all categories). We note that no labeled data was used for training. We do not rescale the heat maps per image (i.e. the range used in this visualization is consistent across examples).

their larynx, and their tuba could all potentially be called the source of a sound. Hence we use this term to refer to motions that are correlated with production of a sound, and study it through network visualizations.

To do this, we apply the class activation map (CAM) method of Zhou et al. [59], which has been used for localizing ambient sounds [52]. Given a space-time video patch I_x, its corresponding audio A_x, and the features assigned to them by the last convolutional layer of our model, $f(I_x, A_x)$, we can estimate the probability of alignment with:

$$p(y \mid I_x, A_x) = \sigma(w^\top f(I_x, A_x)), \tag{2}$$

where y is the binary alignment label, σ the sigmoid function, and w is the model's final affine layer. We can therefore measure the information content of a patch — and, by our hypothesis, the likelihood that it is a sound source — by the magnitude of the prediction $|w^\top f(I_x, A_x)|$.

One might ask how this self-supervised approach to localization relates to generative approaches, such as classic mutual information methods [24,25]. To help understand this, we can view our audio-visual observations as having been produced by a generative process (using an analysis similar to [60]): we sample the label y, which determines the alignment, and then conditionally sample I_x and A_x. Rather than computing mutual information between the two modalities (which requires a generative model

Model	Acc.
Multisensory (full)	82.1%
Multisensory (spectrogram)	81.1%
Multisensory (random pairing [16])	78.7%
Multisensory (vision only)	77.6%
Multisensory (scratch)	68.1%
I3D-RGB (scratch) [56]	68.1%
O3N [19]*	60.3%
Purushwalkam et al. [61]*	55.4%
C3D [62,56]*	51.6%
Shuffle [17]*	50.9%
Wang et al. [63,61]*	41.5%
I3D-RGB + ImageNet [56]	84.2%
I3D-RGB + ImageNet + Kinetics [56]	94.5%

Table 1: Action recognition on UCF-101 (split 1). We compared methods pretrained without labels (top), and with semantic labels (bottom). Our model, trained both with and without sound, significantly outperforms other self-supervised methods. Numbers annotated with "*" were obtained from their corresponding publications; we retrained/evaluated the other models.

that self-supervised approaches do not have), we find the patch/sound that provides the most information about the latent variable y, based on our learned model $p(y \mid I_x, A_x)$.

Visualizations What actions does our network respond to? First, we asked which space-time patches in our test set were most informative, according to Equation 2. We show the top-ranked patches in Figure 3, with the class activation map displayed as a heatmap and overlaid on its corresponding video frame. From this visualization, we can see that the network is selective to faces and moving mouths. The strongest responses that are not faces tend to be unusual but salient audio-visual stimuli (e.g. two top-ranking videos contain strobe lights and music). For comparison, we show the videos with the weakest response in Figure 4; these contain relatively few faces.

Next, we asked how the model responds to videos that do not contain speech, and applied our method to the Kinetics-Sounds dataset [16] — a subset of Kinetics [58] classes that tend to contain a distinctive sound. We show the examples with the highest response for a variety of categories, after removing examples in which the response was solely to a face (which appear in almost every category). We show results in Figure 5.

Finally, we asked how the model's attention varies with motion. To study this, we computed our CAM-based visualizations for videos, which we have included in the supplementary video (we also show some hand-chosen examples in Figure 1(a)). These results qualitatively suggest that the model's attention varies with on-screen motion. This is in contrast to single-frame methods models [50,52,16], which largely attend to sound-making objects rather than actions.

5 Action recognition

We have seen through visualizations that our representation conveys information about sound sources. We now ask whether it is useful for recognition tasks. To study this, we fine-tuned our model for action recognition using the UCF-101 dataset [64], initializing the weights with those learned from our alignment task. We provide the results in Table 1, and compare our model to other unsupervised learning and 3D CNN methods.

We train with 2.56-second subsequences, following [56], which we augment with random flipping and cropping, and small (up to one frame) audio shifts. At test time, we follow [65] and average the model's outputs over 25 clips from each video, and use a center 224×224 crop. See the supplementary material for optimization details.

Analysis We see, first, that our model significantly outperforms self-supervised approaches that have previously been applied to this task, including Shuffle-and-Learn [17] (82.1% vs. 50.9% accuracy) and O3N [19] (60.3%). We suspect this is in part due to the fact that these methods either process a single frame or a short sequence, and they solve tasks that do not require extensive motion analysis. We then compared our model to methods that use supervised pretraining, focusing on the state-of-the-art I3D [56] model. While there is a large gap between our self-supervised model and a version of I3D that has been pretrained on the closely-related Kinetics dataset (94.5%), the performance of our model (with both sound and vision) is close to the (visual-only) I3D pretrained with ImageNet [66] (84.2%).

Next, we trained our multisensory network with the self-supervision task of [16] rather than our own, i.e. creating negative examples by randomly pairing the audio and visual streams from different videos, rather than by introducing misalignment. We found that this model performed significantly worse than ours (78.7%), perhaps due to the fact that its task can largely be solved without analyzing motion.

Finally, we asked how components of our model contribute to its performance. To test whether the model is obtaining its predictive power from audio, we trained a variation of the model in which the audio subnetwork was ablated (activations set to zero), finding that this results in a 5% drop in performance. This suggests both that sound is important for our results, and that our visual features are useful in isolation. We also tried training a variation of the model that operated on spectrograms, rather than raw waveforms, finding that this yielded similar performance (see supplementary material for details). To measure the importance of our self-supervised pretraining, we compared our model to a randomly initialized network (i.e. trained from scratch), finding that there was a significant (14%) drop in performance — similar in magnitude to removing ImageNet pretraining from I3D. These results suggest that the model has learned a representation that is useful both for vision-only and audio-visual action recognition.

6 On/off-screen audio-visual source separation

We now apply our representation to a classic audio-visual understanding task: separating on- and off-screen sound. To do this, we propose a source separation model that uses our learned features. Our formulation of the problem resembles recent audio-visual and audio-only separation work [34,36,67,42]. We create synthetic sound mixtures by summing an input video's ("on-screen") audio track with a randomly chosen ("off-screen") track from a random video. Our model is then tasked with separating these sounds.

Task We consider models that take a spectrogram for the mixed audio as input and recover spectrogram for the two mixture components. Our simplest on/off-screen separation model learns to minimize:

$$\mathcal{L}_{\mathcal{O}}(x_M, I) = \|x_F - f_F(x_M, I)\|_1 + \|x_B - f_B(x_M, I)\|_1, \tag{3}$$

where x_M is the mixture sound, x_F and x_B are the spectrograms of the on- and off-screen sounds that comprise it (i.e. foreground and background), and f_F and f_B are our model's predictions of them conditional on the (audio-visual) video I.

We also consider models that segment the two sounds without regard for their on- or off-screen provenance, using the permutation invariant loss (PIT) of Yu et al. [36].

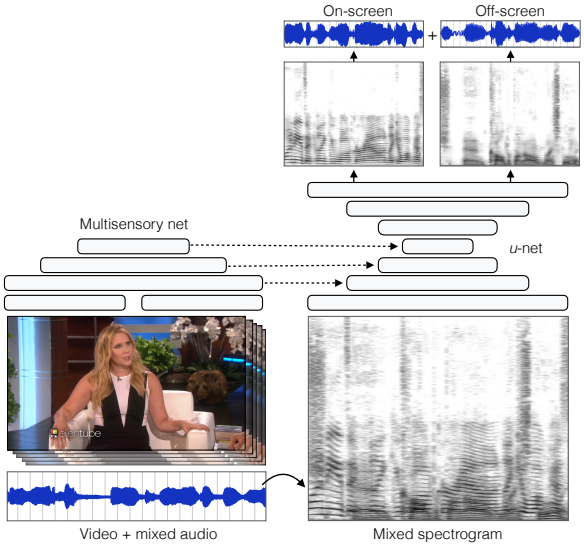

Fig. 6: Adapting our audio-visual network to a source separation task. Our model separates an input spectrogram into on- and off-screen audio streams. After each temporal downsampling layer, our multisensory features are concatenated with those of a u-net computed over spectrograms. We invert the spectrograms to obtain waveforms. The model operates on raw video, without any preprocessing (e.g. no face detection).

This loss is similar to Equation 3, but it allows for the on- and off-screen sounds to be swapped without penalty:

$$\mathcal{L}_{\mathcal{P}}(x_F, x_B, \hat{x}_1, \hat{x}_2) = \min(L(\hat{x}_1, \hat{x}_2), L(\hat{x}_2, \hat{x}_1)), \tag{4}$$

where $L(x_i, x_j) = \|x_i - x_F\|_1 + \|x_j - x_B\|_1$ and $\hat{x}_1$ and $\hat{x}_2$ are the predictions.

6.1 Source separation model

We augment our audio-visual network with a u-net encoder-decoder [43,69,70] that maps the mixture sound to its on- and off-screen components (Figure 6). To provide the u-net with video information, we include our multisensory network's features at three temporal scales: we concatenate the last layer of each temporal scale with the layer of the encoder that has the closest temporal sampling rate. Prior to concatenation, we use linear interpolation to make the video features match the audio sampling rate; we then mean-pool them spatially, and tile them over the frequency domain, thereby reshaping our 3D CNN's time/height/width shape to match the 2D encoder's time/frequency shape. We use parameters for u-net similar to [69], adding one pair of convolution layers to compensate for the large number of frequency channels in our spectrograms. We predict both the magnitude of the log-spectrogram and its phase (we scale the phase loss by 0.01 since it is less perceptually important). To obtain waveforms, we invert the predicted spectrogram. We emphasize that our model uses raw video, with no preprocessing or labels (e.g. no face detection or pretrained supervised features).

Training We evaluated our model on the task of separating speech sounds using the VoxCeleb dataset [71]. We split the training/test to have disjoint speaker identities (72%,

Method	All				Mixed sex		Same sex		GRID transfer	
	On/off	SDR	SIR	SAR	On/off	SDR	On/off	SDR	On/off	SDR
On/off + PIT	11.2	7.6	12.1	10.2	10.6	8.8	11.8	6.5	13.0	7.8
Full on/off	11.4	7.0	11.5	9.8	10.7	8.4	11.9	5.7	13.1	7.3
Mono	11.4	6.9	11.4	9.8	10.8	8.4	11.9	5.7	13.1	7.3
Single frame	14.8	5.0	7.8	10.3	13.2	7.2	16.2	3.1	17.8	5.7
No early fusion	11.6	7.0	11.0	10.1	11.0	8.4	12.1	5.7	13.5	6.9
Scratch	12.9	5.8	9.7	9.4	11.8	7.6	13.9	4.2	15.2	6.3
I3D + Kinetics	12.3	6.6	10.7	9.7	11.6	8.2	12.9	5.1	14.4	6.6
u-net PIT [36]	–	7.3	11.4	10.3	–	8.8	–	5.9	–	8.1
Deep Sep. [67]	–	1.3	3.0	8.7	–	1.9	–	0.8	–	2.2

Table 2: Source separation results on speech mixtures from the VoxCeleb (broken down by gender of speakers in mixture) and transfer to the simple GRID dataset. We evaluate the on/off-screen sound prediction error (On/off) using ℓ_1 distance to the true log-spectrograms (lower is better). We also use blind source separation metrics (higher is better) [68].

VoxCeleb short videos (200ms)				
	On-SDR	SDR	SIR	SAR
Ours (on/off)	7.6	5.3	7.8	10.8
Hou et al. [42]	4.5	–	–	–
Gabbay et al. [44]	3.5	–	–	–
PIT-CNN [36]	–	7.0	10.1	11.2
u-net PIT [36]	–	7.0	10.3	11.0
Deep Sep. [67]	–	2.7	4.2	10.3

Table 3: Comparison of audio-visual and audio-only separation methods on short (200ms) videos. We compare SDR of the on-screen audio prediction (On-SDR) with audio resampled to 2 kHz.

8%, and 20% for training, validation, and test). During training, we sampled 2.1-sec. clips from longer 5-sec. clips, and normalized each waveform's mean squared amplitude to a constant value. We used spectrograms with a 64 ms frame length and a 16 ms step size, producing 128×1025 spectrograms. In each mini-batch of the optimization, we randomly paired video clips, making one the off-screen sound for the other. We jointly optimized our multisensory network and the u-net model, initializing the weights using our self-supervised representation (see supplementary material for details).

6.2 Evaluation

We compared our model to a variety of separation methods: 1) we replaced our self-supervised video representation with other features, 2) compared to audio-only methods using blind separation methods, 3) and compared to other audio-visual models.

Ablations Since one of our main goals is to evaluate the quality of the learned features, we compared several variations of our model (Table 2). First, we replaced the multisensory features with the I3D network [56] pretrained on the Kinetics dataset — a 3D CNN-based representation that was very effective for action recognition (Section 5). This model performed significantly worse (11.4 vs. 12.3 spectrogram ℓ_1 loss for Equation 3). One possible explanation is that our pretraining task requires extensive motion analysis, whereas even single-frame action recognition can still perform well [65,72].

We then asked how much of our representation's performance comes from motion features, rather than from recognizing properties of the speaker (e.g. gender). To test this, we trained the model with only a single frame (replicated temporally to make a

Fig. 7: Qualitative results from our on/off-screen separation model. We show an input frame and spectrogram for two synthetic mixtures from our test set, and two in-the-wild internet videos containing multiple speakers. The first (a male/male mixture) contains more artifacts than the second (a female/male mixture). The third video is a real-world mixture in which a female speaker (simultaneously) translates a male Spanish speaker into English. Finally, we separate the speech of two (male) speakers on a television news show. Although there is no ground truth for these real-world examples, the source separation method qualitatively separates the two voices. Please refer to our webpage (http://andrewowens.com/multisensory) for video source separation results.

video). We found a significant drop in performance (11.4 vs. 14.8 loss). The drop was particularly large for mixtures in which two speakers had the same gender — a case where lip motion is an important cue.

One might also ask whether early audio-visual fusion is helpful — the network, after all, fuses the modalities in the spectrogram encoder-decoder as well. To test this, we ablated the audio stream of our multisensory network and retrained the separation model. This model obtained worse performance, suggesting the fused audio is helpful even when it is available elsewhere. Finally, while the encoder-decoder uses only monaural audio, our representation uses stereo. To test whether it uses binaural cues, we converted all the audio to mono and re-evaluated it. We found that this did not significantly affect performance, which is perhaps due to the difficulty of using stereo cues

in in-the-wild internet videos (e.g. 39% of the audio tracks were mono). Finally, we also transferred (without retraining) our learned models to the GRID dataset [73], a lab-recorded dataset in which people speak simple phrases in front of a plain background, finding a similar relative ordering of the methods.

Audio-only separation To get a better understanding of our model's effectiveness, we compared it to audio-only separation methods. While these methods are not applicable to on/off-screen separation, we modified our model to have it separate audio using an extra permutation invariant loss (Equation 4) and then compared the methods using blind separation metrics [68]: signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), and signal-to-artifacts ratio (SAR). For consistency across methods, we resampled predicted waveforms to 16 kHz (the minimum used by all methods), and used the mixture phase to invert our model's spectrogram, rather than the predicted phase (which none of the others predict).

We compared our model to PIT-CNN [36]. This model uses a VGG-style [74] CNN to predict two soft separation masks via a fully connected layer. These maps are multiplied by the input mixture to obtain the segmented streams. While this method worked well on short clips, we found it failed on longer inputs (e.g. obtaining 1.8 SDR in the experiment shown in Table 2). To create a stronger PIT baseline, we therefore created an audio-only version of our u-net model, optimizing the PIT loss instead of our on/off-screen loss, i.e. replacing the VGG-style network and masks with u-net. We confirmed that this model obtains similar performance on short sequences (Table 3), and found it successfully trained on longer videos. Finally, we compared with a pretrained separation model [67], which is based on recurrent networks and trained on the TSP dataset [75].

We found that our audio-visual model, when trained with a PIT loss, outperformed all of these methods, except for on the SAR metric, where the u-net PIT model was slightly better (which largely measures the presence of artifacts in the generated waveform). In particular, our model did significantly better than the audio-only methods when the genders of the two speakers in the mixture were the same (Table 2). Interestingly, we found that the audio-only methods still performed better on blind separation metrics when transferring to the lab-recorded GRID dataset, which we hypothesize is due to the significant domain shift.

Audio-visual separation We compared to the audio-visual separation model of Hou et al. [42]. This model was designed for enhancing the speech of a previously known speaker, but we apply it to our task since it is the most closely related prior method. We also evaluated the network of Gabbay et al. [45] (a concurrent approach to ours). We trained these models using the same procedure as ours ([45] used speaker identities to create hard mixtures; we instead assumed speaker identities are unknown and mix randomly). Both models take very short (5-frame) video inputs. Therefore, following [45] we evaluated 200ms videos (Table 3). For these baselines, we cropped the video around the speaker's mouth using the Viola-Jones [76] lip detector of [45] (we do not use face detection for our own model). These methods use a small number of frequency bands in their (Mel-) STFT representations, which limits their quantitative performance. To address these limitations, we evaluated only the on-screen audio, and downsampled the audio to a low, common rate (2 kHz) before computing SDR. Our model significantly outperforms these methods. Qualitatively, we observed that [45] often smooths the in-

put spectrogram, and we suspect its performance on source separation metrics may be affected by the relatively small number of frequency bands in its audio representation.

6.3 Qualitative results

Our quantitative results suggest that our model can successfully separate on- and off-screen sounds. However, these metrics are limited in their ability to convey the quality of the predicted sound (and are sensitive to factors that may not be perceptually important, such as the frequency representation). Therefore, we also provide qualitative examples.

Real mixtures In Figure 7, we show results for two synthetic mixtures from our test set, and two real-world mixtures: a simultaneous Spanish-to-English translation and a television interview with concurrent speech. We exploit the fact that our model is fully convolutional to apply it to these 8.3-sec. videos ($4\times$ longer than training videos). We include additional source separation examples in the videos on our webpage. This includes a random sample of (synthetically mixed) test videos, as well as results on in-the-wild videos that contain both on- and off-screen sound.

Multiple on-screen sound sources To demonstrate our model's ability to vary its prediction based on the speaker, we took a video in which two people are speaking on a TV debate show, visually masked one side of the screen (similar to [25]), and ran our source separation model. As shown in Figure 1, when the speaker on the left is hidden, we hear the speaker on the right, and vice versa. Please see our video for results.

Large-scale training We trained a larger variation of our model on significantly more data. For this, we combined the VoxCeleb and VoxCeleb2 [77] datasets (approx. $8\times$ as manys videos), as in [47], and modeled ambient sounds by sampling background audio tracks from AudioSet approximately 8% of the time. To provide more temporal context, we trained with 4.1-sec. videos (approx. 256 STFT time samples). We also simplified the model by decreasing the spectrogram frame length to 40 ms (513 frequency samples), predicting the spectrogram magnitude instead of its log, and increasing the weight of the phase loss to 0.2. Please see our video for results.

7 Discussion

In this paper, we presented a method for learning a temporal multisensory representation, and we showed through experiments that it was useful for three downstream tasks: (a) pretraining action recognition systems, (b) visualizing the locations of sound sources, and (c) on/off-screen source separation. We see this work as opening two potential directions for future research. The first is developing new methods for learning fused multisensory representations. We presented one method — detecting temporal misalignment — but one could also incorporate other learning signals, such as the information provided by ambient sound [15]. The other direction is to use our representation for additional audio-visual tasks. We presented several applications here, but there are other audio-understanding tasks could potentially benefit from visual information and, likewise, visual applications that could benefit from fused audio information.

Acknowledgements This work was supported, in part, by DARPA grant FA8750-16-C-0166, U.C. Berkeley Center for Long-Term Cybersecurity, and Berkeley DeepDrive. We thank Allan Jabri, David Fouhey, Andrew Liu, Morten Kolbæk, Xiaolong Wang, and Jitendra Malik for the helpful discussions.

References

1. Smith, L., Gasser, M.: The development of embodied cognition: Six lessons from babies. Artificial life **11**(1-2) (2005) 13–29 1
2. Sekuler, R.: Sound alters visual motion perception. Nature (1997) 1, 2
3. de Sa, V.R.: Learning classification with unlabeled data. Advances in neural information processing systems (1994) 1, 3
4. Shimojo, S., Shams, L.: Sensory modalities are not separate modalities: plasticity and interactions. Current opinion in neurobiology (2001) 2
5. McGurk, H., MacDonald, J.: Hearing lips and seeing voices. Nature (1976) 2
6. British Broadcasting Corporation: Is seeing believing? (2010) 2
7. Schwartz, J.L., Berthommier, F., Savariaux, C.: Audio-visual scene analysis: evidence for a" very-early" integration process in audio-visual speech perception. In: Seventh International Conference on Spoken Language Processing. (2002) 3
8. Omata, K., Mogi, K.: Fusion and combination in audio-visual integration. In: Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences. (2008) 3
9. Nahorna, O., Berthommier, F., Schwartz, J.L.: Binding and unbinding the auditory and visual streams in the mcgurk effect. The Journal of the Acoustical Society of America **132**(2) (2012) 1061–1077 3
10. Nahorna, O., Berthommier, F., Schwartz, J.L.: Audio-visual speech scene analysis: characterization of the dynamics of unbinding and rebinding the mcgurk effect. The Journal of the Acoustical Society of America **137**(1) (2015) 362–377 3
11. Barker, J.P., Berthommier, F., Schwartz, J.L.: Is primitive av coherence an aid to segment the scene? In: AVSP'98 International Conference on Auditory-Visual Speech Processing. (1998) 3
12. Hershey, J., Attias, H., Jojic, N., Kristjansson, T.: Audio-visual graphical models for speech processing. In: Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on. Volume 5., IEEE (2004) V–649 3
13. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: ICML. (2011) 3
14. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. CVPR (2016) 3
15. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Ambient sound provides supervision for visual learning. In: ECCV. (2016) 3, 5, 14
16. Arandjelović, R., Zisserman, A.: Look, listen and learn. ICCV (2017) 3, 7, 8, 9
17. Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: unsupervised learning using temporal order verification. In: European Conference on Computer Vision, Springer (2016) 527–544 3, 8, 9
18. Wei, D., Lim, J.J., Zisserman, A., Freeman, W.T.: Learning and using the arrow of time. In: CVPR. (2018) 3
19. Fernando, B., Bilen, H., Gavves, E., Gould, S.: Self-supervised video representation learning with odd-one-out networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2017) 5729–5738 3, 8, 9
20. McAllister, D.F., Rodman, R.D., Bitzer, D.L., Freeman, A.S.: Lip synchronization of speech. In: Audio-Visual Speech Processing: Computational & Cognitive Science Approaches. (1997) 3
21. Marcheret, E., Potamianos, G., Vopicka, J., Goel, V.: Detecting audio-visual synchrony using deep neural networks. In: Sixteenth Annual Conference of the International Speech Communication Association. (2015) 3

22. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Workshop on Multi-view Lip-reading, ACCV. (2016) 3
23. Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. CVPR (2017) 3
24. Hershey, J.R., Movellan, J.R.: Audio vision: Using audio-visual synchrony to locate sounds. In: NIPS. (1999) 3, 7
25. Fisher III, J.W., Darrell, T., Freeman, W.T., Viola, P.A.: Learning joint statistical models for audio-visual fusion and segregation. In: NIPS. (2000) 3, 7, 14
26. Kidron, E., Schechner, Y.Y., Elad, M.: Pixels that sound. In: CVPR. (2005) 3
27. Barzelay, Z., Schechner, Y.Y.: Harmony in motion. In: Computer Vision and Pattern Recognition, 2007. CVPR'07. Conference on. (2007) 3
28. Cherry, E.C.: Some experiments on the recognition of speech, with one and with two ears. The Journal of the acoustical society of America (1953) 3
29. Bregman, A.S.: Auditory scene analysis: The perceptual organization of sound. MIT press (1994) 3
30. Ghahramani, Z., Jordan, M.I.: Factorial hidden markov models. In: Advances in Neural Information Processing Systems. (1996) 472–478 3
31. Roweis, S.T.: One microphone source separation. In: Advances in neural information processing systems. (2001) 793–799 3
32. Cooke, M., Hershey, J.R., Rennie, S.J.: Monaural speech separation and recognition challenge. Computer Speech & Language **24**(1) (2010) 1–15 3
33. Virtanen, T.: Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE transactions on audio, speech, and language processing **15**(3) (2007) 1066–1074 3
34. Hershey, J.R., Chen, Z., Le Roux, J., Watanabe, S.: Deep clustering: Discriminative embeddings for segmentation and separation. In: Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, IEEE (2016) 31–35 3, 9
35. Chen, Z., Luo, Y., Mesgarani, N.: Deep attractor network for single-microphone speaker separation. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). (March 2017) 246–250 3
36. Yu, D., Kolbæk, M., Tan, Z.H., Jensen, J.: Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In: Acoustics, Speech and Signal Processing (ICASSP). (2017) 3, 9, 11, 13
37. Darrell, T., Fisher, J.W., Viola, P.: Audio-visual segmentation and "the cocktail party effect". In: Advances in Multimodal Interfaces (ICMI). (2000) 3
38. Pu, J., et al.: Audio-visual object localization and separation using low-rank and sparsity. In: ICASSP. (2017) 3
39. Casanovas, A.L., et al.: Blind audiovisual source separation based on sparse redundant representations. Transactions on Multimedia (2010) 4
40. Rivet, B., et al.: Audiovisual speech source separation: An overview of key methodologies. IEEE Signal Processing Magazine (2014) 4
41. Khan, F., Milner, B.: Speaker separation using visually-derived binary masks. In: Auditory-Visual Speech Processing (AVSP). (2013) 4
42. Hou, J.C., Wang, S.S., Lai, Y.H., Tsao, Y., Chang, H.W., Wang, H.M.: Audio-visual speech enhancement using multimodal deep convolutional neural networks. (2017) 4, 9, 11, 13
43. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. (2015) 4, 10
44. Gabbay, A., Ephrat, A., Halperin, T., Peleg, S.: Seeing through noise: Speaker separation and enhancement using visually-derived speech. arXiv preprint arXiv:1708.06767 (2017) 4, 11

45. Gabbay, A., Shamir, A., Peleg, S.: Visual speech enhancement using noise-invariant training. arXiv preprint arXiv:1711.08789 (2017) 4, 13
46. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. SIGGRAPH (2018) 4
47. Afouras, T., Chung, J.S., Zisserman, A.: The conversation: Deep audio-visual speech enhancement. arXiv preprint arXiv:1804.04121 (2018) 4, 14
48. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. arXiv preprint arXiv:1804.03160 (2018) 4
49. Gao, R., Feris, R., Grauman, K.: Learning to Separate Object Sounds by Watching Unlabeled Video. arXiv preprint arXiv:1804.01665 (2018) 4
50. Senocak, A., Oh, T.H., Kim, J., Yang, M.H., Kweon, I.S.: Learning to localize sound source in visual scenes. arXiv preprint arXiv:1803.03849 (2018) 4, 8
51. Arandjelović, R., Zisserman, A.: Objects that sound. arXiv preprint arXiv:1712.06651 (2017) 4
52. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Learning sight from sound: Ambient sound provides supervision for visual learning. arXiv preprint arXiv:1712.07271 (2017) 4, 7, 8
53. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016) 5
54. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. (2010) 4
55. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013) 5
56. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. arXiv preprint arXiv:1705.07750 (2017) 5, 8, 9, 11
57. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dartaset for audio events. In: ICASSP. (2017) 5
58. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017) 6, 8
59. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 2921–2929 7
60. Isola, P., Zoran, D., Krishnan, D., Adelson, E.H.: Learning visual groups from co-occurrences in space and time. arXiv preprint arXiv:1511.06811 (2015) 7
61. Purushwalkam, S., Gupta, A.: Pose from action: Unsupervised learning of pose features based on motion. arXiv preprint arXiv:1609.05420 (2016) 8
62. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV. (2015) 8
63. Wang, X., Gupta, A.: Unsupervised learning of visual representations using videos. In: ICCV. (2015) 8
64. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012) 8
65. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems. (2014) 8, 11
66. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. (2009) 9

67. Huang, P.S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P.: Joint optimization of masks and deep recurrent neural networks for monaural source separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **23**(12) (December 2015) 2136–2147 9, 11, 13
68. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. IEEE transactions on audio, speech, and language processing (2006) 11, 13
69. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arXiv preprint arXiv:1611.07004 (2016) 10
70. Michelsanti, D., Tan, Z.H.: Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification. arXiv preprint arXiv:1709.01703 (2017) 10
71. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612 (2017) 10
72. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR. (2014) 11
73. Cooke, M., Barker, J., Cunningham, S., Shao, X.: An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America (2006) 13
74. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 13
75. Kabal, P.: Tsp speech database. McGill University, Database Version (2002) 13
76. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. Volume 1., IEEE (2001) I–I 13
77. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622 (2018) 14