Digital Object Identifier

Increasing Importance of Joint Analysis of Audio and Video in Computer Vision: A Survey

AHMED SHAHABAZ[1], and SUDEEP SARKAR[2], (Fellow, IEEE)
[1]Computer Science and Engineering, University of South Florida (e-mail: shahabaz@usf.edu)
[2]Computer Science and Engineering, University of South Florida (e-mail: sarkar@usf.edu)

Corresponding author: Ahmed Shahabaz (e-mail: shahabaz@usf.edu).

This research was supported in part by the US National Science Foundation grant IIS 1956050

ABSTRACT The joint analysis of audio and video is a powerful tool that can be applied to various contexts, including action, speech, and sound recognition, audio-visual video parsing, emotion recognition in affective computing, and self-supervised training of deep learning models. Solving these problems often involves tackling core audio-visual tasks, such as audio-visual source localization, audio-visual correspondence, and audio-visual source separation, which can be combined in various ways to achieve the desired results. This paper provides a review of the literature in this area, discussing the advancements, history, and datasets of audio-visual learning methods for various application domains. It also presents an overview of the reported performances on standard datasets and suggests promising directions for future research.

INDEX TERMS computer vision, audio-video analysis, contrastive learning, multi-modal analysis

I. INTRODUCTION

We rely on auditory cues to perform various tasks in our daily lives, including voice recognition, object recognition, recognizing sounds of musical instruments, and identifying vehicles. While vision recognition systems are designed to visually confirm an event or an object's presence, including auditory features, can improve their accuracy. For example, it can be difficult to differentiate between a fire brigade vehicle and an ambulance from a distance based on visual information alone, but the inclusion of the sirens' sound makes it easy to distinguish between the two. In some cases, we are able to recognize events in our environment based solely on the sounds we hear without the need for visual input. For example, children may recognize the presence of an ice cream truck only by hearing its distinctive sound.

There are also interesting findings from human perception studies that show how auditory interpretation can change when video data is present. A compelling demonstration is the McGurk effect [1], a phenomenon in which the integration of auditory and visual speech information leads to the perception of a different sound than the one spoken. This occurs when a person hears a sound incompatible with the visual information presented, such as seeing a person's mouth produce a different sound than what is being heard.

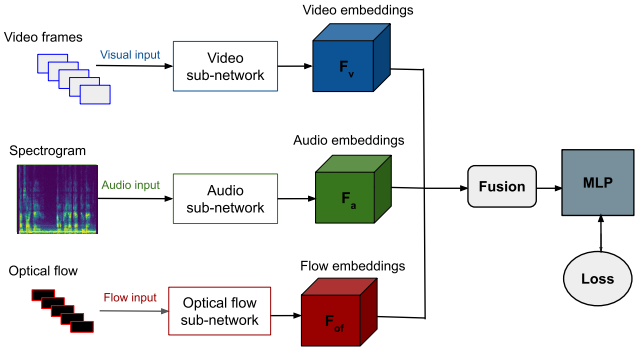

FIGURE 1: A typical multi-modal network with video (image frames), audio & optical flow modalities as input. The embeddings of all the modalities are fused together before sending the fused embedding through a classifier.

For example, if a person sees a video of someone saying "ba" but hears the sound "ga," they may perceive the sound as "da." The McGurk effect illustrates how our brains rely on auditory and visual information when interpreting speech and demonstrates visual information's strong influence on our perception of sound.

There are other cases when a video event interpretation

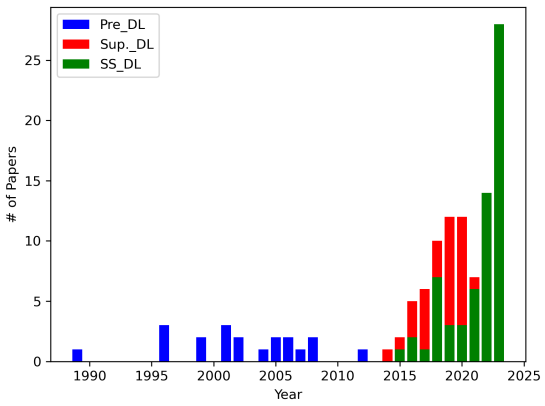

FIGURE 2: The number of audio-video papers over the years, clustered into three categories: pre-deep learning (Pre_DL, blue), supervised deep learning (Sup_DL, red), and self-supervised deep learning (SS_DL, green).

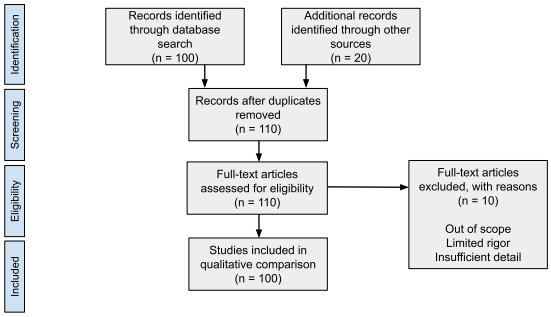

FIGURE 3: PRISMA [2] flow diagram for our review process.

is changed by audio. Imagine a video of a person walking across a room and then picking up a glass of water. If the sound of the person's footsteps is played at the same time as the video, the viewer will perceive the person as simply walking across the room. However, if the sound of the person's footsteps is played slightly before the video, the viewer may perceive the person as walking towards the camera and then stopping. This is because the sound of the footsteps arriving before the visual information creates the impression that the person is coming closer. This demonstrates how the timing of audio can affect our perception of the spatial relationships in a visual scene.

Computer vision researchers have recognized the potential for auditory features to improve accuracy and even change visual interpretation. As a result, significant research has been conducted in this area, focusing on improving performance for traditional computer vision tasks such as action or activity recognition and addressing association problems such as sound source localization and separation in images. Despite this progress, the full potential of jointly exploiting audio and video has yet to be realized. This is partly due to the challenges presented by processing auditory features, including the susceptibility of acoustic signals to noise and the difficulty of combining auditory and visual signals.

We started by selecting around 100 papers related to audio-visual learning. For this selection, we used Keith Price's Bibliography[1]. It is an already annotated database of papers with indexes for keywords, words, author, year, journal/conference. It contains resources related to the computer vision community. Some of the keywords that we used for searching through papers were *audio-visual learning, multimodal learning, sound & video*. The website first appeared in 1994. From the hundred papers, we started by sorting papers from well-known groups in recent times. Most of these recent papers use self-supervision. We gathered around 40 papers that use self-supervised learning. Later, we started including the new papers as they kept publishing. As for supervised deep learning and pre-deep learning approaches, we included everything we could find in the Keith Price Bibliography. Later, we added the newer approaches of supervised deep learning as we could find. Our survey can also be looked at in terms of PRISMA [2] (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework. In figure 3, we show the systematic approaches in accordance with PRISMA.

As shown in Figure 2, the number of audio-video papers in computer vision has significantly increased due to the advancement of deep learning. Among these deep learning approaches, self-supervised methods, which do not require extensive labeled datasets, have seen a notable rise in popularity.

In the pre-deep learning era, many methods were based on probabilistic techniques such as Gaussian mixture models (GMMs) or hidden Markov models (HMMs) (as shown in Table 2). However, to our knowledge, the total number of papers that used audio and visual data was not as much compared to the modern era of deep learning. It was probably due to the requirements of higher computing power. The most common tasks in this era were speaker localization, speaker diarization, or sound source localization.

Early deep learning architectures for audio-visual tasks often used parallel channels for audio and video data, integrating their features at higher levels and training on labeled data. However, as these tasks became more complex, the need for large amounts of labeled data became unsustainable. This led to the development of self-supervised approaches, which do not rely on labels. These approaches include contrastive learning using positive and negative pairs [3], using a teacher video network trained with labels to train an audio network without labels [4], [5], and using clustering to self-label videos [6]–[8]. Self-supervised learning approaches typically solve the audio-visual correspondence

[1]http://www.visionbib.com/bibliography/contents.html

or association task as a pretext learning task. Then, the pre-trained network is used to solve a downstream task using a limited number of manual labels to demonstrate the efficacy of the learning method.

The primary novelty of our work, delineated in Table 1, lies in its comprehensive scope within the domain of audio-visual learning, surpassing the breadth of existing review works. Contrary to most of these works, which predominantly focus on recent deep learning-based approaches or specific sub-problems within the domain, our survey extends beyond these confines. Our review embraces contemporary deep learning methodologies and the foundational pre-deep learning methods. This dual focus enables our work to provide a more holistic and inclusive overview of the field, positioning it as a more expansive and thorough review compared to the current literature. Additionally, our work delves into the various applications of audio-visual analysis methods and systems, offering insights into their diverse practical implementations. We include a qualitative analysis of current methodologies, identifying existing gaps and challenges. Based on these findings, we discuss prospective opportunities for future research in the field of audio-visual learning.

The industrial significance of this review work extends across diverse domains. It can serve as a valuable reference point for various applications:

- Self-driving car: This review can be a good reference point for the self-driving vehicle industry to search relevant literature or methodologies. Audio-visual analysis systems can be used to improve the detection of emergency vehicles in self-driving cars. Sometimes, the lanes can be too congested for emergency vehicles.
- Wildlife monitoring: Extending the reach of our work to wildlife monitoring applications to provide tools for environmental conservation. Audio-visual analysis methods we have discussed can be deployed to aid in tasks such as detecting & and tracking wildlife movements, protecting wild habitats, securing against illegal poaching, and mitigating human-wildlife conflicts.
- Meeting room diarization: In the following section, we discuss the applications and opportunities of the meeting room diarization problem in more detail. It is an area of audio-visual learning that still needs more focus or interest from different researchers.
- Classroom interaction: Another industrial usefulness of our survey is improving teacher-student interaction in the classroom, as this field heavily depends on auditory and visual cues. The qualitative comparison that we provide will be helpful for many.

This research endeavors to delve into the realm of audio-visual learning, highlighting the present methodologies, discernible gaps, and potential opportunities. It particularly accentuates the underlying challenges and assumptions inherent in various computational strategies. Our investigation primarily concentrates on the learning methodologies employed in this domain rather than engaging in direct comparative analysis of diverse deep learning architectures. While we offer a qualitative assessment of these methodologies, it is important to note that a quantitative comparison falls beyond the purview of our current study. This approach allows us to focus more on these methodologies' theoretical and conceptual aspects, providing insights into their strengths and limitations within the context of audio-visual learning.

Authors	Summary of Survey
Zhu *et al.* [9]	Focuses only on deep learning based approaches. Provides a survey on four AV tasks.
Michelsanti *et al.* [10]	Survey about deep learning based AV speech enhancement & separation. Focuses only on one task.
He *et al.* [11]	Survey of deep learning based depression recognition of humans.
Zeng *et al.* [12]	Survey of emotion recognition.
Shoumy *et al.* [13]	Survey of sentiment and emotion recognition using text, audio & video combined with physiological signals.
Katsaggelos *et al.* [14]	Results and challenges of AV fusion strategies. Discusses in terms of AV speech recognition.
Chen *et al.* [15]	Focuses on video saliency detection (VSD) task. Saliency detection refers to identifying important object/things/patterns in a video.
Akhtar & Falk [16]	Survey of multimedia quality assessment. Assessing quality of audio-visual signal at end-user.
Potamianos *et al.* [17]	Review of automatic speech recognition in audio-visual domain.
Zhang *et al.* [18]	Review of deep learning based multi-modal emotion recognition of human.
Campbell [19]	Analyzes how seeing the talker affects auditory perception of speech. Goes beyond McGurk [1] effect.

TABLE 1: Summary of other surveys in the domain of audio-visual analysis.

We summarize the relevant works in Table 2, Table 3, and Table 4 according to three categories: pre-deep learning, supervised deep learning, and self-supervised deep learning. These tables present each work by task solved, the dataset used for training and testing, input features employed, the computational method, and the method used to fuse audio and video sources. This layout of the paper follows the order of columns in these tables.

In Section II, we discuss the core audio-visual tasks solved followed by a discussion of the datasets used for these tasks in Section III. The technical approaches employed are discussed in Section IV. Section IV includes the computational approaches to solve different problems. Then, in Section V, we have summarized and discussed the performances reported by different papers along with our analysis of the quantitative performances reported. Then, we wrap up our survey by first discussing the new opportunities for future research in this field in Section VI. Then, we conclude by summarizing our analysis in Section VII.

TABLE 2: Pre-deep learning computer vision works that used audio and video. For each work, we list the task solved, the dataset used for training and testing, the input features employed, the computational method, and the method used to fuse audio and video sources.

Authors	Method	Task	Features	Fusion	Data/Dataset
Kidron *et al.* [20]	Canonical correlation analysis	Sound Localization	-	Projection to 1D subspace	Vf-A
Beal *et al.* [21], [22]	EM for params. learning & BI for tracking	Speaker Tracking	-	Gaussian approx. of delay in microphones & position in frame	Vf-S
Ben-Yacoub *et al.* [23]	Robust Correlation	Speaker Verification	A: LPCC	SVM	XM2VTS[23]
Fisher & Darrell [24]	Joint probability to calc. mutual info. between A-V data	AVC	-	Projection to lower dimension from higher	Vf-A
Naphade & Huang [25], [26]	Factor graph multinet for context EM for learning HMM& GMM	Event Understanding	V: color, histogram structure etc. A: MFCC	-	Video Clip
Kulesh *et al.* [27]	HMM & GMM	Clip Recognition	-	Vf-A	
Hung *et al.* [28]	BIC [29]	Speaker Diarization	-	-	AMI meeting [30]
Nam *et al.* [31]	V: Sapatio-temporal dynamic activity signature computation A: Gaussian Modeling	Violent scene characterisation	V: Motion sequence A: Raw audio	-	Vf-A
Rasheed & Shah [32]	V: Visual disturbance A: Energy peakiness test	Movie genre classification	V: Motion content & color A: Raw audio	-	Vf-A
Vermaak *et al.* [33]	V-Loc.: Countour Tracking S-Loc.: TDOA	Speaker Tracking	Particle filter at prediction level	-	Vf-A
Ravulapalli & Sarkar [34]	Perceptual Grouping Principle: such as Gestalt principles of similarity	A-V association	V: motion & shape A: Spectrogram	-	Vf-A
Vajaria *et al.* [35]	Speaker segmentation: BIC Speaker clustering: GSC SSL: PCA (eigen vector)	Speaker localization	V: Grey-scale difference image A: MFCC	-	Vf-A
Vajaria *et al.* [36]	A-V association & clustering [35]	Clip retrieval of a query	V: Image difference A: MFCC	-	Vf-A
Vajaria *et al.* [37]	Segmentation into ATPs: BIC A-clustering: GSC	Speaker diarization Speaker localization	V: Image difference A: MFCC	Concatenation after PCA on both	Video clip
Fisher *et al.* [38]	Maximizing mutual info. between projected low dimensional A-V data	Speaker localization	V: Pixel, motion A: Periodogram	Projection to lower dimension from higher	Vf-A
Hershey & Movellan [39]	A-V synchrony as mutual info. estimate between A-V signal	Speaker localization	V: Pixel intensity A: Avg. energy over interval	-	V: NTSC A: Raw audio

TABLE 3: Supervised deep learning computer vision works that used audio and video. For each work, we list the task solved, the dataset used for training and testing, the input features employed, the computational method, and the method used to fuse audio and video sources.

Authors	Task	Network/Method	Loss	Inputs	Fusion	Dataset
Souza *et al.* [40]	Generating structured video interpretations	Feat. Ext.: CNN [41] Classifier: SVM	Acceptor function	V: Frames, Flow A: Spectrogram BoAW	Grenander's structure [43]	CMU Kitchen [42]
Zhou *et al.* [44]	Sound generation given visual	SG: SampleRNN [45] V: VGG-19 [46]	Cross-entropy	Vf & Motion	Concatenation	VEGAS [44]
Chen *et al.* [47]	A-V navigation	Deep reinforcement learning	Proximal policy optimization	V: RGB & Depth A: Binaural spectrogram	Concatenation	Matterport3D [48] Replica [49]
Chen *et al.* [50]	A-V navigation	Reinforcement learning based on transformer	Proximal policy optimization	V: RGB & Depth A: Binaural spectrogram	-	Matterport3D [48]
Subedar *et al.* [51]	Action recognition	Bayesian DNN	ELBO Cross-entropy	Vf-A	Bayesian fusion framework	MiT [52] UCF101 [53]
Wu *et al.* [54]	A-V event localization	Dual attention matching	Cross-entropy	Vf-A	-	AVE [55]
Kazakos *et al.* [56]	Multi-modal fusion	Feat. Ext.: BN-Inception [58]	Classification loss	V: RGB, Flow A: Spectrogram	Concatenation	Epic-Kitchens [57]
Liu *et al.* [59]	Saliency prediction	V: CNN, LSTM A: 3D-CNN	KL divergence	V: RGB, Flow A: Log-MFCC	CNN	MVVA [59]
Qian *et al.* [60]	Vid. classification A-V alignment,AVC SSL, AVSS	V: ResNet-18 [61] A: CRNN [63] Grad-CAM	Binary cross-entropy	V:Frames A: Sepctrogram	-	AudioSet [62] SoundNet-Flickr [64]
Ramaswamy & Das [65]	Multi task learning for SSL via AVC	V,A Feat. Ext.: CNN Feat.->Fusion-> LSTM	A-V Triplet Gram Matrix Loss [65]	V: Frames A: Log-MFCC	MFB [65] & LSTM	AVE [55]
Tsiami *et al.* [66]	SSL via saliency estimation	STAViS [66]	Cross-entropy Correlation coefficient Norm. scanpath saliency	Vf-A	Different ways	-
Owens *et al.* [67]	Predict sound from silent video	CNN->RNN	Prediction error	V: Frames A: Sub-band envelopes	-	Greatest Hits [67]
Shi *et al.* [68]	Material Recognition	CNN		V: Geometry A: Spectrogram	MFB->Concat.	GLAudio [68]
Lee *et al.* [69]	AVSS (speech)	Encoder-decoder	Magnitude loss SI-SDR	Vf-A	Cross-modal affinity	VoxCeleb2 [70] LRS2, LRS3
Zhou *et al.* [71]	Event localization	CNN->LSTM-> Linear	Cross-entropy A-V similarity loss	Vf-A	Addition of features	AVE [55]
Tian *et al.* [55]	Event localization	CNN->Attn.->LSTM	Contrastive	Vf-A	Dual multimodal residual netw. [55]	AVE [55]
Xu*et al.* [72]	Binaural audio genearation from mono-aural audio	Using SHD & HRIR	Difference betwn. left and right channel spectrum	Vf-A	-	FAIR-Play [] MUSIC-Stereo [] YT-Music []
Gao *et al.* [73]	Efficient action recognition	Teacher-student	L_1 loss KL-divergence	V: Frames A: Spectrogram	Concatenation	Kinetics [74] Kinetics-Sound [75]
Korbar *et al.* [76]	Efficient action recognition	V: Resnet [61] A: VGGish [62]	Saliency ranking loss [76]	V: Frames A: MEL-spectrogram	-	Sports1M [77]
Rai *et al.* [78]	Hierarchical action recognition	Cooperative learning	Noise contrastive eestimation (NCE)	V: Ego & third person view A: MFCC	-	Home Action Genome []
Ji *et al.* [79]	Generating emotion controllable talking face	Encoder-decoder	Cross & self reconstruction loss	V: Vf A: MFCC	Concatenation	MEAD [80]
Panda *et al.* [81]	Efficient video recognition	LSTM	Effcy. ls.:Gumbel-softmax Clsf. ls.:Cross-entropy	V: RGB & motion A: Spectrogram	Concatenation	Kinetics-Sound [75]
Chen *et al.* [82]	Video classification	Teacher-student	Compositional contrastive learning	I: 2D-CNN V: 3D-CNN A: 1D-CNN	Concatenation & Residual block	UCF51 [53] VGG-Sound [83]
Xia & Zhao [84]	A-V event localization	Attention based background suppression	Cross-entropy	Vf-A	Cross-modal gated attention	AVE [55]
Jiang *et al.* [85]	Egocentric AVSL	V: CNN A: ResNet-18 [61]	Cross-entropy	V: Vf A: Spectrogram	Concatenation	EasyCom [86]
Li *et al.* [87]	A-V question answering	V enc.: Resnet-18 [61] A enc.: VGGish [62] Enc.->LSTM	Cross-entropy	V: Vf A: Spectrogram A: Spectrogram	-	MUSIC-AVQA [87]
Yang *et al.* [88]	Metric scale 3D pose reconstruct	3D CNN	Opt. 3D heat maps	V: Vf A: Spectrogram	Max. pool-> Concatenation	PoseKernel [88]

TABLE 4: Self-supervised deep learning computer vision works that used audio and video. For each work, we list the pretext task used, the downstream task solved, the dataset used for training and testing, the learning method explored along with the loss function, and the method used to fuse audio and video sources. (Abbreviations in ***Network*** column. **A:** Audio, **V:** Video/Visual, **F:** Face and **L:** Lip attribute extractor accordingly. For the rest of the abbreviations please refer to the Acronyms [VII])

Authors	Learning Method, Loss Function	Pretext Task	Downstream Tasks	Network	Fusion	Dataset
Arandjelović & Zisserman [75]	Contrastive	A-V correspondence	Sound & Video classification	L^3-Net [75]	Concatenation	-
Aytar *et al.* [4]	Teacher(V)-Student(A)	Sound representation learning	Sound Classification	V: [89], [90] A: CNN	-	Flickr [91]
Korbar *et al.* [92]	Curriculum learning with contrastive loss	A-V temporal synchronization	Audio classification Action recognition	V: MCx [93] A: VGG [46] like	Concatenation	Kinetics [74] SoundNet [4] AudioSet [62]
Owens & Efros [94]	Contrasting by randomly shifting audio	Predict A-V alignment	SSL , AVSS , A-V action recognition	V,A: CNN	Tile & Concatenation	AudioSet [62]
Owens *et al.* [6]	Audio cluster as label OR Binary coding of audio	Predict audio from visual input	Object & Scene recognition	V,A: CNN	-	Flickr [91]
Zhao *et al.* [95]	Multi speaker sound source separation from mixture	AVSS	SSL	V:Dilated ResNet-18 [61],A:U-Net [96]	-	MUSIC [95]
Senocak *et al.* [97]	Semi-supervised attention + Unsupervised triplet loss [98]	SSL	-	V: VGG-16 [46] A: CNN	-	Flickr [91] SoundNet [4]
Gao *et al.* [99]	Multi-modal multiple instance learning using weak labels [97]	AVSS	-	V:ResNet-152 [61] A: NMF MIML for object-audio matching	-	AudioSet [62] AV-Bench [100]–[102]
Arandjelović & Zisserman [103]	Euclidean distance between V&A embeddings	A-V correspondence	SSL	AVE-Net [103]	Euclidean distance	AudioSet [62]
Afouras *et al.* [104]	Contrastive	Audio-visual synchronization	Multi-speaker sound source separation, Speaker tracking & detecting and Correcting misaligned A-V data	V,A: VGG-M [105]	Accumulation	LRS2 [106], LRS3 [107], Columbia [108]
Asano *et al.* [7]	Sinkhorn clustering	Self-labelling using optimal transport algorithm	-	V: R(2+1)D-18 [93] A: ResNet [61]	-	VGG-Sound [83] Kinetics [74] Kinetics-Sound [75] AVE [55]
Morgado *et al.* [3]	Contrastive for CMD (AVID)+ CMA)	Action recognition	-	V:R(2+1)D-18 [93] & A:Conv2D	-	Kinetics [74], UCF-101 [53], HMDB-51 [109]
Afouras *et al.* [110]	Contrastive for SSL + Clustering for self-label	SSL & Self-labelling	Object detection	Faster R-CNN [111] for object detection	-	VGG-Sound [83], AudioSet [62]
Chen *et al.* [112]	Contrastive with cross-modal correspondence + Negative mining from background	SSL	-	V,A: CNN	-	Flickr-SoundNet [64], VGG-SS [112]
Rouditchenko *et al.* [113]	Multi speaker sound source separation from mixture + Disentanglement of initial representations	AVSS and Image segmentation	-	V:Dilated ResNet-18 [61] & A: U-Net [96]	-	AVE [55]
Nagrani *et al.* [114]	Contrastive with Curriculum mining	Person identity	-	Face:VGG-M [115] & Voice:VGG-Vox [70]	-	VoxCeleb [70]
Tian *et al.* [116]	Multi-instance multi-label learning with cross-modal attention	A-V video parsing	-	V: ResNet [61] & A: VGGish [62]	-	LLP [116]
Wu & Yang [117]	Contrastive for cross-modal Attention	A-V video parsing	-	V,A: CNN	Cross-modality attention	LLP [116]
Hu *et al.* [8]	Max-margin loss	A-V correspondence	SSL, Multi-source detection and A-V understanding	V:VGG-16 [46] & A:VGGish [62]	-	Flickr [91]
Morgado *et al.* [118]	Weighted contrastive & Instance discrimination loss to address False positive & False negative respectively	Cross modal instance discrimination	Action recognition	V:R(2+1)D [93] & A:9-layer 2D CNN	-	UCF-101 [53], HMDB-51 [109]
Morgado *et al.* [119]	Predict FOA from ZOA	Converting mono audio to spatial audio	-	V: Resnet-18 [61] Motion: FlowNet2 [120] A: 2D-CNN encoder	Concatenation	REC-STREET [119] YT-ALL [119]

Gan *et al.* [5]	Teacher(V)-Student(A) framework	SSL	Vehicle tracking	V: YOLOv2 [121] & A: CNN	-	Auditory Vehicle Tracking [5]
Gao & Grauman [122]	Consistency + co-separation loss	AVSS	-	V: Resnet-18 [61] & A: U-NET [96]	Concatenation	AudiouSet [62]
Gao & Grauman [123]	L2 loss of spectrogram prediction	Monoaural to binaural	AVSS	V: ResNet-18 [61] & A: U-Net [96]	Concatenation	FAIR-Play [123]
Harwath *et al.* [124]	Contrastive	Image-audio retrieval	Speech-prompted object localization, Clustering A-V pattern	V: VGG [46]	Similarity scoring function	Places audio caption
Hu *et al.* [125]	Energy function	Automatic speech recognition	-	Multi-modal RBM [125]	Multi-modal RBM	AVLetters [126], AVLetters2 [127]
Zhao *et al.* [128]	Curriculum learning	Exploits A-V correspondence for SSL & AVSS	-	V: Resnet-18 [61], Motion: PWC-Net [129] & Sound sep.:U-Net [96]	Attention->concatenation	MUSIC [95] & URMP [130]
Yang *et al.* [131]	Classification cross entropy	Channel flip prediction	SSL, AVSS & Audio spatialization	V:ResNet-18 [61] & A:S&E [132]	Concatenation	YouTube-ASMR [131]
Yang *et al.* [133]	Maximum correlation loss	A-V speech recognition & Activity classification	-	Correlational-RNN [133]	Correlational-RNN Encoder	AVLetters [126] & CUAVE [134]
Khosravan *et al.* [135]	Classification err. or Regression err.	AVS as binary classification & regression	-	V,A:CNN [94], (V-A concat.)->3D-CNN	Concatenation	AudioSet [62]
Gao & Grauman [136]	Cross-modal contrastive + Mask pred. + Consistency	Multi task learning for A-V speech separation using cross-modal speaker embedding	-	F: Resnet-18 [61], A: U-Net [96] & Lip mtn.:3D-CNN->TCN	Concatenation	VoxCeleb2 [137],CUAVE [134],LRS2 [106]
Tian *et al.* [138]	Cyclic co-learning of SSL & AVSS	AVSS & SSL	-	SSL: [138] & AVSS: [95]	-	MIT MUSIC [95]
Xuan *et al.* [139]	Proposal based paradigm	SSL	-	LSTM based	-	MUSIC [95], MUSIC-Synthetic [140], SSL [97]
Zhang *et al.* [141]	Transformer	Activity recognition	-	-	-	EPIC-Kitchens-55 [57]
Zhou *et al.* [142]	Video motion graph	Gesture matching	-	-	-	TED-talks [143] & Personal Story
Mercea *et al.* [144]	Triplet ls. + reconstruction ls.	A-V zero-shot learning	-	-	-	GZSL [144]
Liang *et al.* [145]	Generator [146]	Talking Head Generation	-	Encoder-Decoder	Concatenation	VoxCeleb2 [137],MEAD [80]
Zellers [147]	Contrastive span training for predicting masked snippet	Video representation learning	Visual commonsense reasoning & Activity recognition	Transformer	Joint-Encoder (Transformer)	-
Hu [148]	Cycle-consistent Random Walk [149]	SSL	-	Graph	-	MUSIC [95],VGG-Sound [83],VoxCeleb2 [137]
Afouras *et al.* [150]	Self-label [7]	Object Detection	-	Faster-RCNN	-	AudioSet-Instruments [103] VGG-Sound [83]
Song *et al.* [151]	Positive Mining	SSL	-	SimSiam [152]	-	SoundNet-Flickr [64]
Vasudevan *et al.* [153]	Multi-task learning	Multi-SSL	Multiple tasks	-	-	OmniAudio [154]
Lee *et al.* [155]	Contrastive learning	Image manipulation	Audio & Image classification	StyleGAN2 [156]	StyleGAN [157]	-
Lee *et al.* [158]	Bimodal associative memory	Associative learning	-	Memory-augmented	Concatenation	Kinetics [74], ACIVW [159]
Pan *et al.* [160]	mask ls + box ls + mutual ls	Object segmentation	-	Denoising encoder-decoder	Attention block	AVOS [160]
Yang *et al.* [161]	Re-Synthesis	A-V speech enhancement	-	Encoder-Decoder	-	Facestar [161]

II. CORE AV TASKS AND PROBLEM CONTEXTS

The joint analysis of audio and video connects the two otherwise distinct areas of research. The aim is to explore the synergies between the complementary signals for a better understanding of the dynamic scenes. A video refers to a sequence of frames. With most of the modern cameras videos are captured at 30 frames per second (fps) rate. There are high-speed cameras that can take more fps. Along with the visual information a regular video almost always contains audio. For enhancing human understanding of scenes auditory and visual information are both useful. So, researchers have started looking into the joint analysis to exploit useful information from both modalities.

Audio refers to the acoustic information in the environment. It encompasses a range of sounds and frequencies. The spectrogram refers to the visualization of the frequency over time. Convolutional neural networks (CNNs) have been very successful in extracting useful information from 2D signals such as images. Accordingly CNNs are what are normally used for spectrogram as well. Because spectrogram can be looked at as a 2D image depicting the frequency content over time. Spectrogram $S(f,t)$ can be mathematically defined using the Short-Time Fourier Transform (STFT), representing the magnitude of frequency components at various time intervals:

$$S(f,t) = |\int_{-\infty}^{\infty} x(T)w(T-t)e^{-j2\pi fT}dT| \qquad (1)$$

Here, $x(T)$ is the input audio signal, $w(T-t)$ is a window function, f denotes frequency, and t represents time.

A joint audio-visual analysis architecture 4 comprises two-stream networks. The audio sub-network takes raw audio or spectrogram as input and the video sub-network expects 2D images as input. Both of these sub-networks refer to a CNN-based neural network architecture like ResNet [61].

$$Vid = (Image, Audio)$$
$$\phi_{aud}(Audio) = f_a$$
$$\phi_{image}(Image) = f_v$$

The feature embeddings f_a and f_v are then fused together before sending the fused embeddings through the classification layer. The classification layer normally comprises an MLP (Multi-Layer Perceptron).

The joint analysis of audio-video is a powerful tool that can be applied to many problem contexts, including action/speech/sound recognition, audio-visual video parsing (AVVP), emotion recognition in affective computing, and self-supervised training of deep learning models. The solution to these problems often involves solving a set of core audio-video tasks that can be combined in various ways to achieve the desired results. Some core technical tasks involved in the audio-video analysis include the following.

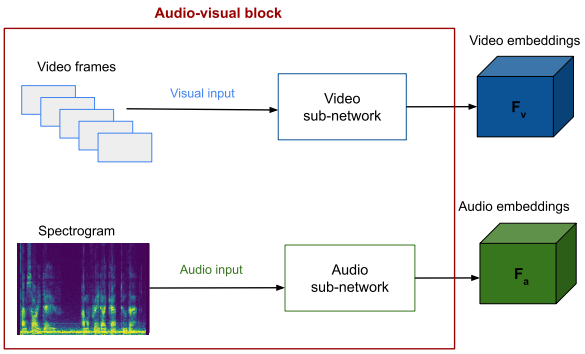

FIGURE 4: General two-stream audio-visual network showing just the feature/embedding extractor block. The embeddings can then be used for solving different audio-visual tasks.

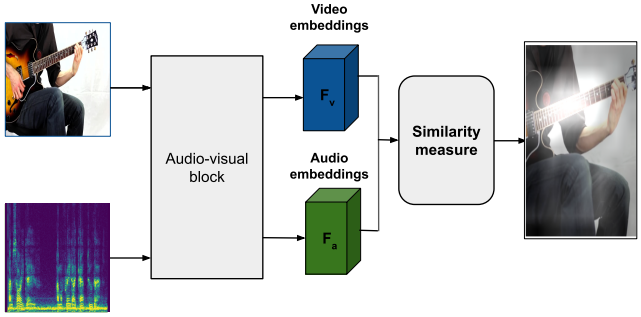

FIGURE 5: Audio-Visual Sound Source Localization (SSL) architecture. It shows the use of the extracted embeddings by A-V block from figure [4].

- **Video-based sound source localization (SSL)** [5], [8], [20], [60], [65], [66], [85], [94], [95], [97], [103], [110], [131], [138], [139], [148], [151], [153], [162]–[166] involves marking pixels' correspondence to each sound source, such as vehicles, in video frames. When the source of sound is a person, we have the audio-visual speaker localization (AVSL) [23], [35] problem, which involves identifying and locating the speaker(s) in an audio-visual scene, such as identifying and locating a person speaking in a video and tracking the speaker [21], [22], [33]. The problem's difficulty level arises if the number of pixels in a video frame for the sound source is small compared with the image size.
- **Audio-visual source separation (AVSS)** [60], [69], [95], [99], [113], [122], [123], [131], [138], [167], [168] involves separating the audio and visual components of a multimedia signal, such as separating the sound of a person speaking [37]–[39], [104], [136] from the visual image of the person. Related to this is speaker diarization [28], [37], which is the problem of identifying and labeling the different speakers in an audio signal. Humans can easily differentiate between

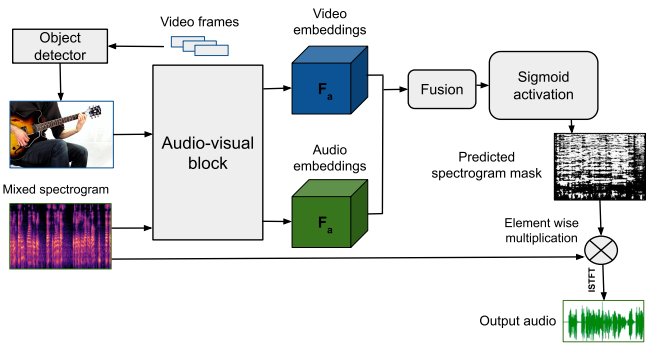

FIGURE 6: Audio-Visual Source Separation (AVSS) architecture. Conditioning on the input image it separates the corresponding audio from the mixture.

or separately identify the sounds of different objects from a mixture of multiple sounds, such as the sound of a guitar, drums, and vocals in music. In many cases, we can even differentiate between the sound of different types of guitars or two different vocals in a song that contains multiple other sounds. The ability to localize each sound source helps in the sound source separation process and vice versa.

- **Audio-Visual classification** into objects [68], [114], [150], [160], [169], actions [51], [73], [76], [78], [118], activity [55], [71], [84], [133], [141], [147], [170]–[173], speech [106], [107], [125], [174]–[176], emotions [170], [177]–[193], saliency prediction [59], [66], [76], [194] and other tasks that uses information from audio [4], [75], [79], [124], [160], [195]–[198] to enhance pure-video-based classification, such as for a moving car, dog barking, or cooking an omelet. Zhang & Li [199] have come up with a benchmark for visual-audio (audio image) denoising, which tackles the audio denoising task as an image segmentation problem in the audio image domain. Their approach also generalizes to speech denoising, audio separation, audio enhancement, and noise estimation problems.

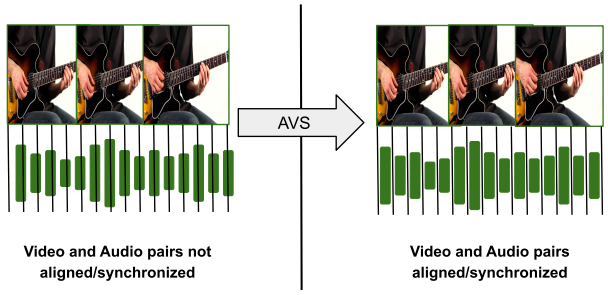

FIGURE 7: Depicting the Audio-Visual Alignment/Synchronization (AV) problem.

- **Audio-visual alignment (AVA)** involves aligning or synchronizing audio and visual signals in time, such as aligning [104], [200], [201] the sound of a person speaking with the video of their mouth moving. Similar to AVA, audio-visual Synchronization (AVS) [104], [135], [200]–[209] involves synchronizing audio and visual signals in time. This specific facet/connection within the realm of audio-visual modality has been utilized as an approach to offer simulated guidance in self-supervised [92], [94], [104], [125], [135], [163] methods.

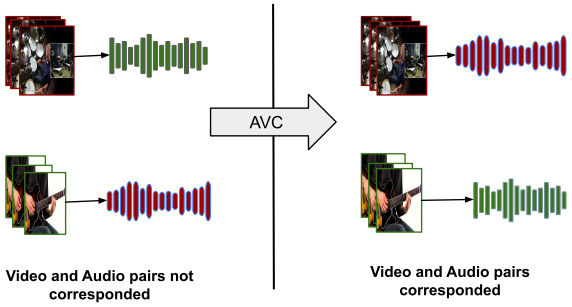

FIGURE 8: Depicting the natural Correspondence between Audio-video (AVC). Other than being a core problem by itself, AVC is also exploited by self-supervised approaches for providing pseudo labels.

  - **Audio-visual correspondence (AVC)** [24], [34], [60], [69], also known as cross-model correspondence (CMC), involves detecting and aligning audio and visual events in time and determining whether they are related. Similar to how audio-visual synchrony (AVS) has been harnessed, many have also employed audio-visual correspondence (AVC) as a method of self-supervision [8], [65], [75], [128], [155].

These core technical tasks in audio-video analysis can be combined to solve different problems. In the following subsections, we outline how solutions to these tasks can be used to solve these larger problems in computer vision.

A. VIDEO RECOGNITION OR AUDIO-VISUAL VIDEO PARSING (AVVP)

Video recognition is the process of automatically analyzing and understanding the content of a video. This typically involves extracting structured information from the video, such as identifying and labeling the different objects and people in the video, detecting and classifying events [210] or actions, and extracting semantic information, such as text or speech. A restricted version of this problem is audio-visual video parsing (AVVP) [116], [117], [211], [212] which also involves extracting structured information [40] from the video but specifically focuses on analyzing only the audio and visual components of the video. It has

many potential applications, such as clip recognition [27], [213], event recognition [25], [26], [40], [54], violent scene recognition [31], movie genre recognition [31], [32], video summarization [40], tracking [5], [21], [22], [33], video search & retrieval [36].

Audio-video analysis can be a powerful tool for these problems. It allows for the analysis of both the sound and visual components of a video and can provide important contextual information about the video. Specifically, it can be used in the following ways.

- Audio-Visual Alignment: By aligning [94] a video's audio and visual components, it is possible to understand better the relationship between the sound and the movement. This can be useful for recognizing videos that involve both sound and movement, such as talking heads [145] or music videos.
- Audio-Visual Classification: By analyzing the sound of a video, it is possible to classify [81], [82] the video based on its characteristics. For example, the sound of a person speaking can be used to classify the video as a news segment, while the sound of music can be used to classify the video as a music video.
- Sound Source Localization: By determining the location of the sound source [5], [110] in a video, it is possible to understand better the context of the video. For example, if the sound of a person speaking is coming from the left side of the screen, the person is likely facing the left.
- Audio-Visual Synchronization: By synchronizing [8], [75], [92], [104] a video's audio and visual components, it is possible to understand better the timing and duration of the video. This can be useful for recognizing videos that involve precise timing, such as sports or dance videos.

B. AUDIO-VISUAL QUESTION ANSWERING:

Audio-visual question answering (AVQA) [87] is a task that involves answering questions about an audio-visual scene. This can include answering questions about what is happening, such as identifying objects or actions or answering questions about what a person is saying in a video. It is a key area of research in computer vision and artificial intelligence, as it has numerous applications in fields such as education, entertainment, and customer service. By answering questions about an audio-visual scene, AVQA has the potential to improve our understanding of the world around us and enhance our ability to interact with it.

Here are some of the ways audio-video analysis can be used in AVQA.

- Audio-visual source separation [95] can be useful in AVQA, as it can help improve the accuracy of speech recognition algorithms and reduce the noise and distractions that can interfere with understanding a scene.
- Audio-visual alignment allows for analyzing the relationship between verbal and nonverbal cues, which can provide additional context [214], [215].
- Audio-Visual classification [216] can be useful in analyzing the relationship between sound and emotion, which can provide additional context and help improve the system's accuracy [217].
- Cross-modal analysis can be useful in the integration of information from multiple sources, which can provide a more complete understanding of a scene [87].

C. AFFECT RECOGNITION

Emotion recognition, also known as affect recognition, has been extensively studied by researchers from various fields, such as psychology, linguistics, computer vision, speech analysis, and machine learning. This field aims to recognize the emotions and moods of individuals through various communicative signals, including audio and visual cues. Traditionally, physiological data such as electrodermal activity (EDA), electrocardiography (ECG), and blood pressure have been commonly used in emotion recognition research. In computer vision, most approaches have focused on visual data, such as facial expressions or facial points, while a smaller number have used audio data. Zeng et al. [12] provides a comprehensive overview of the various approaches, techniques, and features used in emotion recognition.

Despite the longstanding interest in emotion recognition, most previous efforts have focused on using a single data modality, such as visual, audio, or physiological signals. Very few approaches have utilized multi-modal audio-visual data. One potential reason for this is the lack of sufficient datasets. Previous research has shown that multi-modal visual-physiological data can outperform approaches using a single modality of data [170], [177]–[193], [218], [219]. Therefore, it is likely that using audio data in combination with physiological and visual data could also improve the performance of emotion recognition systems.

The joint analysis of the audio and video is used to recognize affect in the following manner.

- Establishing audio-visual correspondence (AVC) [220] and Cross-modal Agreement (CMA) [221] may be useful in identifying affective cues such as body postures, facial expressions, and vocal inflections.
- Audio-visual Alignment (AVA) and Audio-visual Synchronization (AVS) may be useful in identifying and analyzing the temporal relationships between affective cues [186], [189], [192], [218], [222].
- Audio-Visual Classification may be useful in identifying affective cues such as vocal inflections or prosodic features [180], [188], [223].

D. SELF-SUPERVISED LEARNING

Self-supervised learning is a type of machine learning where the model is trained using only the data itself, without the need for explicit labels or supervision. This can be useful when labeled data is scarce or expensive to obtain. Creating a trained network using self-supervised learning consists of two parts. In the first part, the network is trained without supervision using a pretext task. Next, this pre-trained network

is refined with limited, labeled training data for a second downstream task. Audio-video analysis can be useful for the first step in self-supervised learning. The pretext tasks could include audio-visual correspondence [8], [65], [75], [103], [104], [128], [155], [195] or synchronization [92], [94], [104], [125], [135], [163], and audio-visual source separation [95], [99], [113], [122], [128], [135]. These pre-trained networks are then used for downstream tasks such as audio-visual classification and audio source localization. Table [4] gives an overview of different self-supervised learning approaches along with different pretext tasks used by them.

E. AUDIO-VISUAL LARGE LANGUAGE MODELS (AV-LLMS)

Audio-Visual Large Language Models (AV-LLMs) [224], [225] combine the power of large language models (LLMs) and audio-visual learners. The objective here is to harness the power of audio-visual learning into the LLMs. This integration enables LLMs with the capability of understanding multimedia content more comprehensively. Through this better understanding of auditory & visual contents in a video, LLMs are better able to perform tasks such as generating captions for videos, summarizing videos, and multimedia generation.

The advent of AV-LLMs represents a promising field in AI research, though the number of published papers remains limited. Models like Gemini [226] and SORA [227] leverage both audio and video data for improved comprehension of LLMs. But they don't explicitly generate audio data as output. Despite this limitation, their utilization of multi-modal data highlights a significant step towards AV-LLMs. As research progresses, we anticipate the invention of more comprehensive models through the development of AV-LLMs.

III. DATASET

Many video datasets are available, but not all include sound. This section will examine the most commonly used datasets for audio-visual learning. To our knowledge, only two audio-visual datasets existed before the rise of deep learning: XM2VTS[3], NIST Meeting Room [228].

This is because earlier vision techniques did not require large training data. However, as we know, deep learning models require a significant amount of data. As a result, several audio-visual datasets have been released or developed to meet this demand. Some of the most commonly used audio-visual datasets include VGG-Sound [83], AudioSet [62], Epic-Kitchens [57], Kinetics [74], and Flickr-SoundNet [4]. Arandjelovic and Zisserman also created a refined version of the Kinetics [74] dataset called Kinetics-Sounds [75] to ensure that audio and visual events were aligned and both visible and audible. The original Kinetics [74] dataset sometimes required alignment between audio and video tracks.

Apart from the above-mentioned popular datasets in recent a few new datasets and benchmark [229]–[232] have been released. Geng *et al.* [229] has released the first Untrimmed Audio-Visual (UnAV-100) dataset. It comprises 10K untrimmed videos with each video containing 2.8 audio-visual (30K in total) events on average. Unlike other audio-visual datasets for event localization UnAV-100 [229] contains multiple audio-visual events in each video.

Ego4D [230] by Grauman *et al.* is an egocentric audio-visual dataset of daily activities at different locations/scenarios such as home, workplace, outdoor, etc. MMG-Ego4D [231] by Gong *et al.* which is a refined and re-annotated version of the original Ego4D [230] dataset. MMG-Ego4D [231] introduces a new problem for action recognition tasks termed as Multi-modal Generalization (MMG). MMG is the study of multi-modal system behavior when one modality is missing or limited. Both Ego4D [230] and MMG-Ego4D [231] contain audio and video data.

REALIMPACT [232] by Clarke *et al.* is the first object impact sound fields that have been recorded in real life or environment. All other such datasets were modeled in simulations. REALIMPACT dataset contains 150,000 impact sound recordings, which involve 50 everyday objects. They also offer comprehensive annotations, encompassing information such as precise impact locations, microphone positions, contact force profiles, material categorizations, and RGBD images.

In the section [V-D] we report performances of different methods on these datasets. From table 11, we see that pre-training on AudioSet [62] gives the best result. Out of all the popular datasets, AudioSet [62] is the only one that contains hierarchical annotations of audio-visual events. As, AudioSet [62] was designed for audio event classification, audio and visual events are not always synchronized. So, it is not well suited for tasks that rely on both visual and auditory cues, such as Audio-Visual Sound Source Localization, Audio-Visual Sound Source Separation, etc. In the tables[12, 4] we see that VGG-Sound [83], Flick-Soundnet [64] are more commonly used for training and reporting performance of SSL, AVSS tasks. So through our analysis, we propose that self-supervised pre-training on AudioSet [62] is better for sound classification/recognition or sound representation learning tasks. But for tasks dependent on multi-modal feature representation learning it would be better to use other datasets [55], [74], [83]. These datasets are well suited when the goal is to solve tasks that require the objects producing the sounds to be visibly present in the video.

In the following sections, we will discuss some of the most commonly used datasets for audio-visual learning. Table 5 provides an overview of some of the most commonly used datasets for audio-visual learning. The table is divided into two sections. The first section lists datasets on which most models are typically trained. The second section lists

[3]http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/

Dataset	# of classes	Size	Task
Common for training			
Epic-Kitchens [57]	-	55 hrs	Action recognition
VGG-Sound [83]	309	200K	Action recognition
VGG-SS [112]	220	5k	SSL
Kinetics [74]	400	306K	Action recognition
Kinetics-Sound [75]	34	19K	Sound classification
AudioSet [62]	632	208K	Audio classification
UCF [53]	101	13K	Action recognition
Flickr-Soundnet [64]	50	-	SSL
Common for reporting results			
HMDB [109]	-	-	Action Recognition
ESC [233] DCASE [234]	-	-	Sound recognition
MUSIC [95]	-	-	AVSS
VocCeleb [70]	-	-	AVSS
Columbia [108]	-	-	SSL

TABLE 5: Overview of the common datasets

some of the datasets that are popular only for reporting results after being pre-trained on one of the datasets in the first part of Table 5.

A. EPIC-KITCHENS

Epic-Kitchens [57] is a first-person cooking video dataset created by 32 participants. The dataset contains 11.5M frames and 55 hours of videos recorded independently without scripts. It also includes 39.6K labeled action segments and 454.3K object bounding boxes, and audio of participant actions and narrations. The dataset is divided into train and test sets, with the test set featuring both seen and unseen kitchen recordings and benchmarks for object detection and action recognition. In addition to audio-visual data, the dataset also includes optical flow information.

1) EPIC-SOUNDS

EPIC-SOUNDS [235] is a version of EPIC-KITCHENS-100 [236] dataset. EPIC-SOUNDS [235] contains temporal annotations for audio events along with the actions that might have resulted into that in that event. These actions can be solely distinguished from the audio stream and then could be classified as different sound classes. Huh *et al.* [235] have provided human annotations for object materials for the actions that include colliding objects. The dataset includes 44 audio classes and a total of 75.9k audible events. This dataset is benchmarked for audio recognition tasks.

B. VGG-SOUND

The VGG-Sound dataset [83] contains over 200k YouTube video clips with a duration of 10 seconds each, representing 309 different sound classes. Each class includes 200-1000 video clips extracted from YouTube videos using various queries. A maximum of two clips were created from each downloaded video. The videos depict a range of acoustic environments with noise to simulate real-life conditions. The labels for the sound classes in this dataset are non-hierarchical, in contrast to the AudioSet dataset [62]. In most videos, a single dominant sound source is visible and audible, although other noises may be present.

1) VGG-SS

Chen *et al.* [112] have released the VGG-Sound Source (VGG SS) audio-visual localization benchmark, which includes bounding box annotations for the sound sources in the VGG-Sound dataset videos. This benchmark is well-suited for use in top computer vision journals.

C. KINETICS

The Kinetics dataset [74] is widely used for action recognition tasks and features 400 classes of human action collected from YouTube videos. The dataset contains a total of 306 videos. Where each clip is drawn from a different YouTube video and lasts for 10 seconds. Then the videos were labeled for a single dominant action, though other actions may also be present. The dataset includes 250-1000 training videos per class, 50 validation videos per class, and 100 test videos per class. Although the action labels were created based only on the visual modality, the dataset contains audio. So this dataset has been extensively used for multi-modal audio-visual action recognition tasks.

1) Kinetics-Sounds

As discussed above Kinetics-Sounds [75] is a subset of Kinetics [74] dataset. Kinetics-Sounds [75] is labeled for human actions and contains 34 human action classes. The dataset contains a total of 19k videos of 10s long. The whole dataset has been split into the train (15k), validation (1.9k), and test (1.9k) sets.

D. AUDIOSET

AudioSet [62] is a manually labeled dataset for audio event detection, featuring 632 hierarchical audio event categories organized to a depth of six. Each category contains at least 100 videos, all 10 seconds long and drawn from YouTube alongside their accompanying audio. The dataset as a whole comprises 4,971 hours of video. However, this dataset doesn't contain temporal boundary annotations thus making it unsuitable for audio-visual event localization tasks. Also, AudioSet [62] was created for audio event detection, so the audio and visual elements are not perfectly aligned. In many cases, the audio can be heard but the sounding object may not be well centered or visible at all.

1) AVE

AVE [55] a subset OF AudioSet [62] dataset. Unlike AVE [55] contains 4143 YouTube videos. The videos are 10s long and have been annotated for temporal boundaries of audio-visual events. The dataset contains a total of 28 audio-visual events. Each video consists of a minimum of one event lasting for at least 2s. The dataset contains events from different domains, such as animal activities, music performances, human activities, etc. Each event domain contains a minimum of 60 and a maximum of 188 videos.

E. SOUNDNET

Soundnet [112] is a large-scale audio-visual dataset that contains over two million videos downloaded from Flickr using popular tags [91] as queries. However, the audio and visual events in the dataset are not well aligned, and some videos do not show the sounding object in the frame. To address this issue, Senocak *et al.* [64] introduced the Flickr-Soundnet [64] benchmark, a subset of Soundnet [112] that is annotated and ensures that the object emitting sound is visible in the frame. The Flickr-Soundnet [64] dataset consists of 144K videos and is a useful resource for audio-visual localization tasks.

F. UCF

The UCF dataset [53] is a comprehensive collection of human actions, comprising 101 labels and over 13K videos spanning 27 hours. Despite the introduction of more challenging datasets such as VGG-Sound [83], and VGG-SS [112], researchers continue to report results on UCF [53] for comparison with older approaches.

G. HMDB

HMDB or HMDB51 [109], which stands for Human Motion Database. HMDB [109] is a manually labeled human action database containing 51 action classes. This dataset is a collection of 6,766 video clips from various online sources, including YouTube, digitized movies, etc. The 51 action classes can be categorized into 5 major action types. Each of those 51 classes contains at least 101 clips, and each clip contains a single action occurring for at least 1 second. These videos contain different challenging conditions, such as variations in camera viewpoint and motion, background cluttering, and changes in the position, scale, and appearances of the actors. Although the dataset does not contain audio, researchers commonly report results on this dataset for comparison with other approaches.

H. NIST MEETING ROOM

The NIST Meeting Room [228] Pilot Corpus is one of the very few datasets that became popular in the pre-deep learning era. The dataset contains audio-visual recordings and transcripts of 19 meetings totaling to 15 hours. This dataset contains 5 different types of meetings; of these 3 types were simulated or unreal. The video data was collected through 4 stationary and 1 moving/floating camera. The stationary cameras were placed at the 4 surrounding walls of the room. For experimenting with far-field recognition systems, the audio data was collected using 3 microphone arrays placed at different locations in the meeting room. Apart from the microphone arrays audio was also captured using 2 personal wireless microphones and 4 microphones fixed on the conference table. Because of the 2 wireless personal microphones, the participants were allowed to move about the meeting room, thus resulting in a more natural scenario. The total length of the multi-sensor data reaches 266 hours of audio and 77 hours of video.

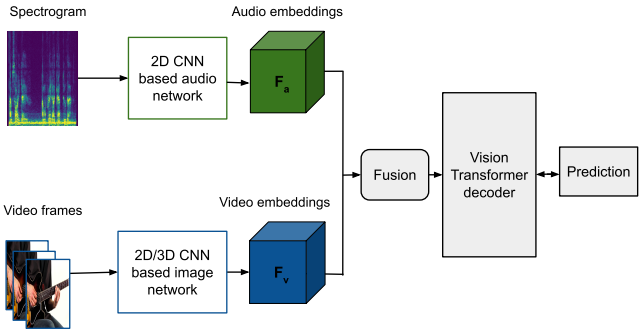

FIGURE 9: An example of a recent vision transformer-based architecture that utilizes the transformer decoder. These architectures normally use CNNs as a backbone for extracting features.

IV. COMPUTATIONAL APPROACHES

Several works have utilized audio-visual multi-modal data with labeled data for supervision in tasks such as sound generation given visual input [44], [67]], sound source localization [60], [65], [66], audio-visual source separation [60], action recognition or event localization [51], [54], [56]], and saliency estimation or prediction [59], [66] (see Table 3. Many of these approaches utilize video frames, spectrograms, or MFCC as visual and aural inputs. However, some works have also incorporated motion information [44], [56], [59] (optical flow) alongside video frames as visual input. These approaches have employed various methods for fusing multi-modal data.

Deep learning models' success heavily relied on data availability, which became abundant due to the proliferation of social media and advances in technology, particularly video recording devices. However, many of these models required supervision, leading to a need for extensive human annotation of the data. The annotation process is time-consuming, tedious, and prone to error, especially when dealing with video data, which often contains many image frames, events, actions, objects, and persons. As a result, researchers turned to self-supervised learning, which was also applied to audio-visual analysis [4], [6], [7], [75], [92], [94], [95], [97], [99], [103], [104]. A summary of self-supervised audio-visual approaches can be found in Table 4.

A. SOUND SOURCE LOCALIZATION (SSL)

In the pre-deep learning era, researchers used various methods to solve the problem of audio-visual (AV) source localization, such as probabilistic models [20], [38] and canonical correlation analysis (CCA) [20], [100]. Speaker tracking tasks [33] also require speaker or source localization to track the speaker effectively. For localization, Vermaak *et al.* [33] uses time delay of arrival. Other problems, such as speaker localization [35]–[39], also fall under the realm of sound source localization (SSL) and use probabilistic methods to

Challenges	Summary
Scarcity of annotated data	Self-supervised learning which takes advantage of the inherent synchronization of data
Absence of one modality	Self-supervised learning which takes advantage of the inherent synchronization of data
Misalignment between modalities	Mostly tackled as regression problem. Where the model tries to predict if the modalities are in sync.
Absence of one modality	Generates one modality given another. Mostly audio is generated given visual input.
Fusion Strategies	People have used different approach such as RNNs, TBN & Concatenation. Concatenation at feature level is the most common.

TABLE 6: Summary of challenges in audio-visual analysis.

solve the task of speaker localization.

Most recent approaches have used deep learning techniques, including a two-stream network with one stream for visual input and another for aural input. Some deep learning models have been trained in a self-supervised, or unsupervised manner [8], [94], [95], [97], [103], [104], [110], [112], [128], [131], [135]. A few self-supervised approaches have been trained using a contrastive loss [104], [110], [112]. Other self-supervised approaches have solved the problem of sound source localization as a downstream task while optimizing for learning audio-visual correspondence or synchronization or alignment [8], [94], [103], [104], [128], [131] as a pretext tasks. We describe some illustrative pre-deep learning and deep learning works next.

1) Pre-deep learning

Kidron *et al.* proposed a Canonical Correlation Analysis (CCA)-based algorithm for localizing pixels associated with a single sound source in [20]. This approach can also be used for speaker localization and has the advantage of having no user-defined parameters. However, CCA requires many visual features to generate reliable statistics. To address this shortcoming, the authors introduced sparsity in their approach.

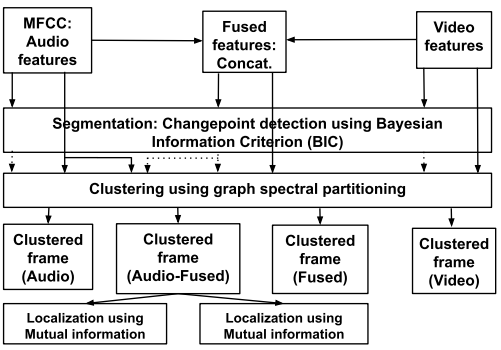

FIGURE 10: An example of SSL architecture [35] in the pre-deep learning era. Vajaria *et al.* [35] have exploited the mutual information between the multi-modal data for speaker localization.

Vajaria *et al.* proposed various approaches for speaker localization [35] and used them to solve speaker clip retrieval [36] and speaker diarization [37]. They used the Bayesian information criterion to segment feature vectors and graph spectral partitioning to cluster the segments of a speaker in a video clip. After segmenting and clustering the multi-modal feature vectors, the authors applied principal component analysis to localize the sound source or speaker. The dominant eigenvector or the vector with the largest eigenvalue identified portions of the image with motion, and the following eigenvectors separated those portions into different individuals.

Fisher *et al.* proposed a non-parametric approach for localizing speakers in a video frame in [38]. Their approach does not assume any existing density in data and instead projects audio-visual features onto a low-dimensional joint subspace, which they learned by maximizing the mutual information between the projected low-dimensional features.

Hershey and Movellan [39] proposed an approach that locates speakers by exploring audio-visual synchronization. They first measured audio-visual synchrony by calculating the mutual information between audio and visual feature vectors and then used a secondary model, such as the centroid computation model, on the mutual information estimates to localize the speaker or sound source.

Ben-Yacoub *et al.* [23] proposed a speaker verification system that uses facial data and speech. They tested their system on the XM2VTS dataset and found that the fused approach outperformed both audio-only and visual-only models. For the fused model, they used a Support Vector Machine (SVM) to fuse audio-visual features at the decision/classification level. They modeled the speaker verification problem as a binary classification problem using SVM, with the data regarded as multi-dimensional data from different modalities. To perform speaker verification, they trained a single SVM classifier on multi-dimensional multi-modal data.

Ravulapalli and Sarkar [34] demonstrated a technique for associating audio and video events using Gestalt principles of similarity from human perception. They first calculated periodicities in audio and visual domains to perform the association. They applied the Fourier transform to the audio spectrogram to calculate audio periodicity and detected significant audio events using a line detection algorithm applied to the spectrogram image. They grouped similar audio events by calculating distances between audio events and matched the resulting audio event clusters with visual event clusters. Finally, they formed audio-visual event pairs based on the maximum likelihood of the audio and visual events belonging to the same event.

2) Supervised deep learning

Several approaches have addressed the problem of SSL in a supervised manner, including [60], [65], [66]. Qian *et al.* [60] proposed a two-stage multi-task learning approach for SSL and audio-visual source separation (AVSS). In the

Application	Architecture	Backbone
Image or Audio classification	ResNet [61]	2D CNN
	VGG [46]	2D CNN
Clip or Video classification	VGG [46]	2D CNN
	L^3-Net [75]	2D CNN
Audio generation	SampleRNN [45]	RNN
Saliency prediction	STAViS [66]	V: 3D CNN A: 2D CNN
Image based sequence recognition	CRNN [63]	LSTM on top of CNN
A-V Correspondence	AVE-Net [103]	2D CNN
	L^3-Net [75]	2D CNN
Sound Source Separation	A: U-Net [96] V: Resnet-18 [61]	2D CNN
Action Recognition	V: R(2+1)D [93] A: 2D Resnet [61]	3D CNN 2D CNN

TABLE 7: Summary of different deep neural network (DNN) architecture.

first stage, their network is trained for audio and video classification tasks and a binary classification task for learning audio-visual correspondence. The second stage consists of a Grad-CAM [237] module, which uses class predictions and audio-visual features to unravel class-specific features in each modality and learn fine-grained audio-visual alignment. For visual feature extraction, the authors used ResNet-18 [61] and CRNN [63] for audio feature extraction. This model can also be used with unlabelled videos, as it utilizes a pre-trained network.

Ramaswamy and Das [65] proposed an algorithm for SSL that can be trained in supervised, weakly supervised, and self-supervised learning schemes. First, audio and visual features are extracted using [46], [238]. Then, an LSTM-based fusion module called the Audio Visual Fusion Block (AVFB) is used to fuse the extracted audio and visual features and learn spatial attention. The output of the AVFB is fed to the Segment-Wise Attention Block (SWAB) module to determine the importance of each audio-visual segment. The aggregated features from each modality are then fed to a series of fully connected layers for supervised or weakly-supervised learning, or the Audio Visual Triplet Gram Matrix Loss (AVTGML) is calculated for self-supervised learning.

Tsiami *et al.* [66] proposed a Spatio-Temporal network (STAVIS) for sound source localization through learning visual saliency. STAVIS consists of a Spatio-temporal visual network made up of a 3D-ResNet [61], and Deeply Supervised Attention Module (DSAM), and an Audio Representation Network made up of the first seven layers of SoundNet [4]. Audio and video features are projected to a common hidden dimension through affine transformations, and localization is performed using cosine similarity, weighted inner product, or bilinear transformation. The first approach provides a single localization, while the others can provide single or multiple localizations.

Liu *et al.* [59] proposed a model for estimating visual saliency using multi-modal data to predict salient faces from videos. They also introduced the MVVA database[4] for training and testing their model. The proposed model can predict salient faces in the presence of multiple faces in a single frame through three branches: one for video frames, one for audio signals, and one for cropped faces from those frames. The visual branch includes a two-stream network consisting of a network for visual frames and another for flow information, which is then concatenated and processed through Convolutional LSTMs. The audio branch consists of 3D-CNNs, and the face branch consists of LSTMs, which generate a face saliency map. These multi-modal features are then integrated with a fusion module to generate the final saliency. Cropped faces are motivated by their ability to explain most attention or fixations.

3) Self-supervised deep learning

Hu *et al.* [8] learn audio-visual correspondence by optimizing a max-margin loss to teach the network to cluster multi-modal vectors, enabling the network to capture multiple audio-visual correspondences. Other self-supervised approaches solve the problem of SSL in various ways, such as predicting audio-visual alignment from shifted audio-visual input [94], separating the sound source from the mixture of multiple audio (spectrogram) inputs [95], [122], and using an attention module to learn audio-visual synchronization [97], [135]. There have also been a few supervised deep learning approaches for SSL [60], [65], which are summarized in Table 3.

B. AUDIO-VIDEO SOURCE SEPARATION (AVSS)

In the deep learning era, the problem of audio-visual source separation (AVSS) has been approached using various machine learning methods, such as self-supervised learning [94], [99], [113], [122], [123], [128], [131], [138] and supervised learning [60], [69]. However, these approaches all have the limitation of assuming that all objects in a scene make sound, and, like in the case of sound source localization, most approaches assume that a single object is the source of the sound at any given time.

To our knowledge, the only work that handles multiple sound sources is by Tian *et al.* [138]. They recognized the interdependence of sound localization and sound separation tasks and proposed a cyclic co-learning framework. Their proposed model consists of two sub-networks: the Visual Grounding Network for sounding objects and the Audio Visual Sound Separation Network. These two sub-networks complement each other, with the Visual Grounding Network trained using a contrastive learning method and the Separation Network trained using the mix and separate method proposed by Zhao *et al.* [95]. Through the use of cyclic co-learning, the authors [138] demonstrated that their approach performs better than current approaches that only solve one of these tasks (AVSS or SSL), highlighting the reciprocal relationship between the two tasks.

[4]https://github.com/MinglangQiao/MVVA-Database

Self-supervised approaches have also been used to solve audio-visual source separation as a downstream task using networks trained on pretext tasks such as predicting audio-visual alignment, generating binaural sound from monoaural sound, and predicting if audio channels are flipped accordingly [94], [123], [131]. Other self-supervised learning approaches [99], [113], [122], [128] have employed different loss functions and learning methods to solve audio-visual source separation.

C. OBJECT OR SPEAKER TRACKING

Several works have tracked audio-visual objects or speakers using the Probabilistic Graphical, or Probabilistic Generative model [21], [22]. Beal *et al.* [21], [22] used an EM algorithm to learn the model parameters and Bayesian Inference to perform speaker tracking over a video sequence. Their setup consists of two microphones and one camera, and they fused audio and visual data by learning a linear mapping between microphone time delay and object/speaker position in the image frame. Vermaak *et al.* [33] proposed a speaker tracking system that also used generative models. Their setup, similar to that of [21], [22], consisted of a single fixed camera and a pair of microphones. The soundtrack was based on the Time Delay of Arrival (TODA) between the two microphones, which was an initialization for the visual tracking model. The visual tracking model included a generative model for motion with a likelihood-based feature search model. The authors applied a particle filter (PF) to fuse the predictions from the audio and visual model and track the speaker. In this setup, the sound and visual models complement each other to improve overall tracking capability.

D. AUDIO-VISUAL CLASSIFICATION

1) Pre-deep learning

Naphade and Huang [25], [26] proposed a probabilistic model that uses factor graphs to model context and improve event recognition for the semantic understanding of objects, sites, and events. After segmenting videos into shots, they extracted visual features such as color (histogram, moment), shape, structure, and texture. For audio features, they used Mel-frequency cepstral coefficients (MFCC), delta, and energy coefficients of different magnitudes and counts. The concepts in this work are represented using multinet [239], a collection of multijects [239] that can describe the time sequence features of an object, site, or event using a probability distribution. In this work, the authors used a mixture of Gaussian components for a site and a hidden Markov model (HMM) for object and event multijects and learned the parameters of these probabilistic models through the Expectation-Maximization (EM) algorithm. They detected events in two steps: first, detecting concepts for each video shot, and then classifying the video clip using a global constraint while taking the shot-level detection from the first step into consideration. The detection of concepts is a binary classification problem, meaning the concept is either present or absent.

Kulesh *et al.* [27] presented an approach for video clip recognition that uses an HMM and Gaussian mixture model (GMM) to model video and audio, respectively, using color histograms and MFCC as features. They estimated the transition matrix of the HMM using the Baum-Welch algorithm.

In their work, Fillipe *et al.* [240] showed how to use Grenander's pattern theory structures to build a structured semantic understanding of audio-video events by reasoning on the multiple-label decisions of deep visual and auditory models. They proposed a structured model that does not require joint training of the structural semantic dependencies and deep models but rather links them as independent components. Furthermore, they used Grenander's structures to facilitate and enhance the fusion of multimodal sensory data, particularly combining auditory and visual features. As a result, they observed significant improvements in the quality of semantic interpretations using deep models and auditory features in combination with Grenander's structures.

2) Supervised deep learning

Video clip classification:

Video clip classification involves assigning labels to an entire video clip. However, clip-level models, which aggregate clip-level classifiers for video classification, are not practical for long videos, as they are computationally complex to process. Therefore, researchers have focused on developing efficient video recognition systems that can handle long videos [73], [81], [241]–[248].

Some approaches aim to select important frames or clips by analyzing the visual modality or video frames [242]–[244], [246]–[248], while others consider both the audio and visual modalities [73], [81], [241], [245].

Gao *et al.* [73] feed the first frame of a video through their pre-trained image student model and the whole video through their video teacher network. They assume that the first frame of a video is the most important, but it could be argued that they are still using the whole video by feeding it through the video teacher network.

Panda *et al.* [81] proposed an approach that selects the modality important for the classification of a particular clip on the fly and then aggregates the results of different classifiers for the classification of the entire video. This approach does not rely on a video teacher network for guiding the self-supervised training of the video classification network.

Action classification:

Kazakos *et al.* [56] extract features from overlapping temporal windows for each modality, fusing them through temporal average pooling and concatenation. They then pass the fused features through a fully connected layer and feed the output through two different fully connected layers for learning verbs and nouns, turning it into a multi-task learning problem.

In contrast, Wang *et al.* [249] proposed the Temporal Segment Network (TSN), which performs late multi-modal

fusion. Temporal Binding Network performs mid-level fusion, generating only one prediction based on the fused features, while TSN generates a segment-level prediction for each modality and then fuses them to get the final prediction.

Subedar *et al.* [51] proposed an uncertainty-aware Bayesian deep learning approach for action recognition that fused multi-modal features based on uncertainty estimates. Their goal was to learn the predictive output distribution of action recognition through learning the posterior parameters using either Markov Chain Monte Carlo sampling, variational inference techniques, or Monte Carlo dropout approximate inference.

Yang *et al.* [133] proposed a method for activity classification from temporal data using RNNs, which was already discussed in this section.

Chen *et al.* [82] proposed network consisted of an image teacher trained on ImageNet and an audio teacher trained on the AudioSet [62] dataset, both of which taught a 3D CNN-based video student network. However, in this work, the student was not simply trained to follow the teacher. Instead, a cross-modal distillation and composition setting was used to distill and combine knowledge from the different teacher networks using different modalities of data. Another unique aspect of this approach was the fusion of audio-visual data through the use of a residual block that added a linear combination of the concatenated multi-modal features to the original features from the teacher network.

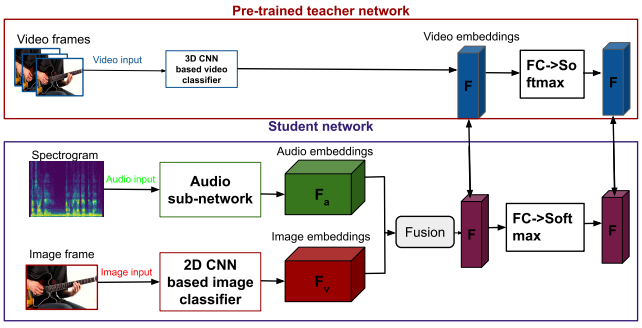

FIGURE 11: A typical A-V teacher-student network with pre-trained image/video teacher. During training the teacher sub-network is kept frozen while the student sub-network learns the embedding space of the teacher.

E. FUSION STRATEGIES FOR CLASSIFICATION

In deep learning approaches, audio and visual data are typically fed to two different sub-networks to extract features. The most common approach for fusing multi-modal data is to concatenate the extracted high-level features or to perform fusion at the prediction or classification level. In the first case, two different classification heads are used for the two modalities, and the final classification is made by fusing the results of the two classifiers. Another approach involves concatenating the extracted features and feeding the combined features to a classification head. However, these approaches do not consider the inherent correlation between the modalities.

Vanderplaetse *et al.* [250] demonstrated different methods for combining audio-visual features, such as multiplying the classification output, averaging the output logits and using a single classifier, and concatenation. These methods were applied at different points along the pipeline and the results of these different merging strategies were reported.

Authors	Approach	Fusion Method
Ngiam *et al.* [251]	RBMs	Concatenation
Slaney & Covell [252]	Optimal linear transform	Canonical correlation analysis
Sargin *et al.* [253]		Canonical correlation analysis
Fisher *et al.* [38]	Probabilistic model	Projection from higher to lower dim.
Yang *et al.* [133]	RNN	Correlational-RNN
Srivastava & Salakhutdinov [254]	Deep RBMs	Concatenation
Vanderplaetse *et al.* [250]	Neural networks	Different ways
Chen *et al.* [82]	Neural network	Residual block
Kazakos *et al.* [56]	Temporal binding network (Neural metwork)	Concat.
Tian *et al.* [55]	Dual multi-modal residual network (Neural network)	Residual block
Hossain & Muhammad [255]	CNN->ELM->SVM	Extreme Learning Machine (ELM) [256]

TABLE 8: Multi-modal Fusion

A few other complex approaches, such as Restricted Boltzmann Machines (RBMs) and autoencoders, have been used for fusing multi-modal audio-visual data. For example, Ngiam *et al.* [251] trained an RBM model to construct a modality given the other as input. Srivastava & Salakhutdinov [254] used a deep learning-based approach (Deep RBMs) for the same task as Ngiam *et al.* [251].

Yang *et al.* [133] used an RNN-based approach for fusing multi-modal audio-visual data and proposed Correlational Recurrent Neural Network (RNN), a new temporal model for this task. They showed the effectiveness of Correlational-RNN on the audio-visual speech recognition task by using maximum correlation loss and a reconstruction loss to learn audio-visual correlations.

Some traditional vision approaches, such as Sargin *et al.* [253] and Slaney & Covell [252], have used Canonical Correlation Analysis (CCA) to merge multi-modal audio-visual data while others have used the projection of high-level audio-visual data to lower dimensions for fusing the signals.

Kazakos *et al.* [56] has proposed Artificial Neural Network based approach for addressing the problem of optimal multi-modal fusion in egocentric videos. They introduced Temporal Binding Network (TBN). Their fusion strategy involves concatenating multi-modal features or embeddings within each overlapping temporal segment referred to as Temporal Binding Window (TBW).

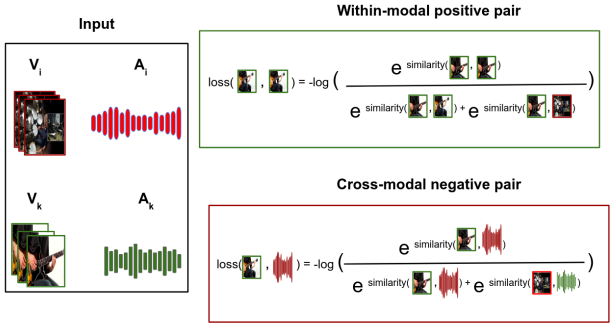

FIGURE 12: Cross and within modal contrastive loss. Contrastive loss function can contrast between similar and dissimilar data points in the high dimensional embedding space.

F. AUDIO-VIDEO ANALYSIS FOR SELF-SUPERVISED LEARNING

Audio-video analysis can be used for self-supervised learning as a pretext task for training a network. Pretext tasks are tasks the network is trained to perform without supervision and can be used to learn features from the data itself. Some examples of pretext tasks for audio-video analysis include audio-visual correspondence or synchronization and audio-visual source separation. This section reviews the details and nuances of this pretext training. Once the network is pre-trained on these tasks, it can be fine-tuned on a downstream task, such as audio-visual classification or audio source localization, using a limited amount of labeled data.

Contrastive learning is a common method for self-supervised learning in audio-visual analysis [75], [92], [104]. However, one limitation of this approach is the need for a large number of negative samples to be effective. This has led to contrastive learning being more commonly used in image classification or data analysis rather than audio-video analysis, as the cost of processing and mining negative and positive samples in video data could be high in terms of computational complexity.

One solution to this issue was the use of a memory bank, as proposed in [257], to store the features extracted from a video as it was fed through the network only once, eliminating the need for repeated feature extraction. However, this approach did have the potential drawback of missing out on backpropagation during network input. Misra & Maaten of [257] argued that this was insignificant if the number of negative samples was large enough.

Another strategy used by Korbar *et al.* [92] was to utilize curriculum learning in conjunction with contrastive learning to not only mine negative samples but also to introduce harder-to-differentiate negatives later in the training stage. These hard negatives were mined from the same video as the positive samples but with a time gap, while the easy negatives were mined from a different video. The effectiveness of curriculum learning in this context was demonstrated in [92].

Several works [6]–[8] have investigated the clustering of extracted audio-visual features, in which the assigned clusters serve as pseudo labels for data. This method is also a potential solution for contrastive learning, as clustering approaches eliminate the need for a large number of negatives.

Owens *et al.* [6] proposed two models in their work, one based on k-means clustering. They first calculate the statistical summary of sound using the approach of McDermott and Simoncelli [258] and then cluster the processed audio using k-means, using these assigned clusters as a pseudo label for visual data. However, the issue with this approach is that the value of k is a hyperparameter.

Asano *et al.* [7] addressed this problem by using a hyperparameter-free clustering method, the Sinkhorn-Knopp clustering algorithm, to solve optimal transport problems. In their approach, audio and visual features are clustered separately in a modality-agnostic way, generating a single label for the multi-modal data by synchronizing the last or output layers of the audio and video sub-networks.

Another problem with contrastive learning is sampling negative and positive samples, often done randomly, leading to potential false positives and false negatives. Morgado *et al.* [118] propose a solution to this issue by using a weighted contrastive learning loss to decrease the contribution of false positives based on the calculation of audio-visual correspondence of the same instance ($\bar{v_i}^T \bar{a_i}$). Additionally, they address the false negative problem by considering the similarity ($\bar{v_i}^T \bar{a_j}$) between instances. Sun *et al.* [164] proposed a false negative aware approach by taking the intra-modal similarity into consideration while selecting negative pairs.

Morgado *et al.* introduced an approach [3] that does not require many negative pairs, unlike other contrastive learning approaches. They achieved good performance by combining Audio-visual instance discrimination and cross-modal agreement (CMA) with only 1024 negative pairs per sample in the dataset and using 32 positive pairs to optimize CMA. Therefore, this approach can be considered an efficient contrastive learning framework. The authors evaluate performance on the downstream tasks of action recognition on the Kinetics [74] dataset and the task of audio classification on the ESC-50 [233] dataset. The authors noted that using more negative pairs did not significantly improve performance compared to the increased time complexity.

Chen et al. [112] proposed a method for automatically mining hard negatives for a contrastive learning framework from the background of video frames. They addressed the problem of audio-visual source localization (SSL) by training their model to learn the sound source localization map by calculating the similarity between extracted visual and audio features. The authors also introduced the concept of a Tri-map, an uncertain region around the sound source that is difficult to classify. While mining for hard negatives,

this uncertain region was ignored. In addition to hard negatives, which were mined from the same input video, the authors also introduced easy negatives, which were created from other videos in the dataset. They found that introducing the Tri-map and ignoring the uncertain region while generating negative samples led to better performance. They also introduced a benchmark for SSL called VGG-Sound Source and achieved state-of-the-art performance on the Flickr-SoundNet [64] dataset. For training their model, the authors only used the central frames of the raw videos as input to the visual sub-network, along with 3s of audio around these input frames. They reported that using all the frames did not significantly improve performance.

Yang *et al.* [133] proposed an RNN-based architecture (CorrRNN) using GRUs for solving audio-visual speech recognition and activity classification tasks. CorrRNN is an encoder-decoder architecture with a multi-modal encoder and multi-modal decoder. The multi-modal encoder consists of three components: a Dynamic Weighting module (DW), a GRU module, and a Correlation module. DW assigns weights to the input by calculating the coherence of the input modalities over time. The GRU module fuses the multi-modal input and creates a joint feature space representation, effectively fusing the input. The correlation module calculates the correlation between the input modalities, later used as the correlation loss. The decoder reconstructs the input from the representation created by the encoder. However, in contrast to other encoder-decoder approaches, the authors [133] used four losses: a fused-reconstruction loss, a self-reconstruction loss, and a cross-reconstruction loss. The fused reconstruction loss measures the error in reconstructing both inputs from the joint representation. The self-reconstruction loss measures the error in reconstructing the input of one modality given the encoded representation of the same modality. The cross-reconstruction loss measures the error in reconstructing the input of one modality given the encoded representation of another modality.

Other self-supervised learning techniques include using teacher-student networks [4], [5] and attempting to predict audio-visual alignment after shifting the audio [94].

G. AUDIO-VISUAL ANALYSIS IN AFFECT RECOGNITION

Table 9 presents an overview of audio-visual-based affect recognition techniques, many of which utilized pre-deep learning methods. According to a survey conducted by Zeng *et al.* [12], there are several common methods for fusing multi-modal features, including feature level, decision or classifier level, and model level fusion. Feature level fusion involves combining multi-modal features before sending them through a machine learning model or classifier, often through concatenation. However, this method can be problematic due to the various features at different time scales, temporal structures, and metric levels. Decision-level fusion involves passing data or features from each modality through different models and making a final classification decision based on the results of each model or classifier. While this addresses some of the issues of feature-level fusion, it also has its own limitations, such as treating interdependent multi-modal data as independent. Model-level fusion is a relatively understudied area that requires further research and exploration.

Authors	Feature	Technique
Busso *et al.* [190]	Markers, Prosody	SVM
Go *et al.* [187]	Eigenfaces MFCC	Linear discriminant analysis (LDA)
Hoch *et al.* [259]	Gabor feature Prosody	SVM
Pal *et al.* [185]	Vertical gray level F0-F3	k-means
Schuller *et al.* [183]	AAM,Prosody,Voice	SVM
Sebe *et al.* [182]	Motion units,Prosody	Bayessian network
Song *et al.* [181]	FAPs, Prosody	THMM
Wang & Guan [180]	Gabor wavelets Prosody, MFCC	Fisher's-LDA
Zeng *et al.* [179]	Motion units, Prosody	MFHMM
Zeng *et al.* [178]	LLP, Prosody	Adaboost + MHMM
Zeng *et al.* [260]	Motion units, Prosody	SNoW
Zeng *et al.* [221]	Motion units, Prosody	MFHMM
Zeng *et al.* [223]	Motion units, Prosody	HMM
Wöllmer*et al.* [192]	MFCC, Facial flow	LSTM
Karpouzis *et al.* [186]	FPs, Prosody	RNN
Caridakis *et al.* [189]	Facial points, Prosody	RNN
Fragopanagos & Taylor [188]	FAPs, Prosody	Neural network (NN)
Petridis & Pantic [184]	Facial points, MFCC	Adaboost, NN
Ringeval *et al.* [191]	Low level descriptor MFCC , LGBP from three orthogonal planes	-
Schoneveld *et al.* [219]	Face Image, MFCC	CNN->LSTM
Hossain & Muhammad [255]	MFCC & Image	CNN->ELM->SVM

TABLE 9: Audio-visual analysis for affect recognition is categorized based on features and techniques.

The most commonly used features in audio-visual affect recognition include facial points, MFCC, prosody, motion units, and features of faces or lips. Among the audio-visual affect recognition approaches, Busso *et al.* [190], Petridis and Pantic [184] and Schuller *et al.* [183] have employed feature-level fusion, which concatenates multi-modal features and passes them through a single affect recognition model or classifier. Decision-level fusion, on the other hand, has been used by Hoch *et al.* [259], Go *et al.* [187], Pal *et al.* [185], Wang and Guan [180], Zeng *et al.* [178], Zeng *et al.* [260], and Zeng *et al.* [223]. However, decision-level fusion ignores the inherent correlation between these multi-modal features. In an effort to benefit from both decision- and feature-level fusion, the remaining approaches in Table 9 have used model-level fusion: Caridakis *et al.* [189], Fragopanagos and Taylor [188], Karpouzis *et al.* [186], Sebe *et al.* [182], Song *et al.* [179], and Zeng *et al.* [221].

W"ollmer *et al.* [192], Caridakis *et al.* [189], Fragopanagos and Taylor [188], Karpouzis *et al.* [186], and Petridis and Pantic [184] have used neural networks (NNs) to learn and model multi-modal audio-visual data and extract

features from it. Other popular methods include SVM [183], [190], [259] and different types of HMMs [178], [179], [181], [221], [223].

As can be inferred from the above discussion, there have been relatively few approaches in the deep learning era for emotion recognition using multi-modal audio-visual data due to the lack of large-scale affect recognition datasets that include audio and visual data. Deep learning approaches, however, rely heavily on large datasets. For a more comprehensive review of emotion recognition, we refer readers to the work of Zeng *et al.* [12] and Wu *et al.* [261].

H. GENERATING ONE MODALITY FROM THE OTHER

Zhou *et al.* [44] proposed an approach for generating sound given visual input using a dataset called Visually Engaged and Grounded AudioSet (VEGAS), a subset of the AudioSet [62] dataset. The proposed architecture consists of a video encoder and a sound generator, with a three-layer SampleRNN [45] being used for the sound generator. The authors [44] experimented with three different video encoder architectures: the *Frame-to-frame method* using an Imagenet pre-trained VGG-19 [46], the *Sequence-to-sequence method* initialized with visual features from the $fc6$ layer of VGG-19, and the *Flow-based method* similar to the sequence-to-sequence method but concatenating flow features with visual features. The concatenated visual and flow features were used in all three methods to initialize the sound generator's hidden state of the coarsest RNN tier.

Other techniques for sound generation approaches include diffusion models [262], [263], Foley Analogies [264].

In alternative approaches, sound serves as a key stimulus for various applications, including the generation of visual scenes [265], the reenactment of face expressions [266]–[268], the reenactment of gestures [142], [269], the creation of emotion-controllable talking heads or faces [79], [146] and the generation of animations [270], [271].

In the domain of Generative AI, diffusion model based Computer Vision applications have seen a increase in application. As we can see from the above discussion this trend has extended to the audio-visual multi-modal domain as well.

I. VIDEO SEGMENTATION

Video segmentation requires partitioning a video into semantically meaningful clips or segments. The goal is to group together the frames or the pixels in a frame. Video segmentation is useful for better analyzing and understanding videos. The video segmentation task can be divided into two types:

1) Spatial Segmentation: [272]–[274]
   - Frame-level Segmentation: It refers to segmenting individual frames in a video while treating each frame independently. The target is to identifying the objects in a frame without considering the temporal dependencies i.e. not tracking the same object across frames.
   - Object-level Segmentation: This includes identifying and tracking specific objects across consecutive frames. The main goal is on maintaining consistency in the segmentation of objects over time.
2) Temporal Segmentation: [275]–[280]
   - Shot Boundary Detection: This refers to detecting the start and end frames of a shot. Shot is a continuous sequence of frames capturing an event without interruption. Shot boundary detection is very crucial for video summarizing and indexing.
   - Action Segmentation: Segmenting or identifying the continuous sequence of frames that captures different actions or activities performed. The is very common for applications such as surveillance, sports analysis and human-computer interaction.

J. REALTIME AUDIO-VISUAL ANALYSIS

Joint audio-visual analysis itself is a interdisciplinary field. Audio-visual analysis with the objective of real-time processing adds an additional layer of complexity for the researchers. Undertaking this task demands for distinct optimization techniques even if the target platform is robust computer devices with GPUs and powerful processing unit. But the difficulty of this task amplifies when the target is to deploy such systems on edge devices or hardware with limited power resources. In this context this inherently multidisciplinary field requires synthesis of expertise in computer vision, machine learning, edge computing and internet of things (IoT).

Real-time audio-visual systems find application in a variety of scenarios, including affective computing [281]–[283], real-time video editing [284], privacy & security [285], [286], different classification & recognition tasks [287]–[292], sound synthesis or generation [293].

V. SUMMARY OF REPORTED PERFORMANCES

The use of deep learning has consistently demonstrated superior performance in audio-visual learning tasks, similar to its impressive results in other computer vision tasks that utilize machine learning techniques. In this section, we will focus on comparing the performances of recent deep learning approaches, with a particular emphasis on self-supervised learning methods, as reported in the literature. These methods have proven highly effective in achieving state-of-the-art results in audio-visual learning tasks.

They are using multi-modal audio-visual data better [55] than relying on only one of these modalities. Especially for self-supervised techniques, using only the audio modality doesn't provide as good a performance as that of using only the visual modality. On the other hand, self-supervised learning methods are becoming increasingly popular because of the cost associated with human annotation. In the case of utilizing inherent information from within the data, multi-modality is better than uni-modality [55], [198].

However, manual supervision still proves to be superior in the case of uni-modal analysis or when multi-modal data is not available.

A. FOUR BENCHMARK TASKS

Researchers have attempted to solve various downstream tasks in order to compare and demonstrate the effectiveness of their deep-learning models. We will examine the reported performance of different methods on four popular downstream tasks: audio-visual source separation (AVSS), sound source localization (SSL), object/action/sound classification, and clustering. These tasks are widely used as benchmarks for audio-visual learning models, as they challenge a model to learn the coherent relationship between audio and visual features - the central objective of any audio-visual learner. While many self-supervised learning approaches pre-train the network to learn or solve for audio-visual correspondence (AVC) as a pretext task, it is difficult to measure or quantify this pretext task directly. Therefore, the performance of the model is often evaluated with it. The five tasks provide comprehensive ways to assess the effectiveness of a model in learning the complex interplay between audio and visual features.

B. METRICS

There is no universally accepted standard metric for evaluating the performance of the chosen tasks, and different measures are typically used for each task. Table 10] shows the common metrics used to compare or measure the performance of different tasks.

Audio-visual Tasks	Common Metrics
AV Sound Separation (AVSS)	signal distortion ratio (SDR), signal interference ratio (SIR), signal to artifacts ratio (SAR)
Sound Source Localization (SSL)	accuracy, area under the curve (AUC), consensus intersection over union (cIoU)
Object/Action Classification	accuracy, mean average precision (MAP)
Clustering	normalized mutual information (NMI)

TABLE 10: Metrics used to evaluate the performance of audio-visual tasks.

Three measures, namely signal distortion ratio (SDR), signal interference ratio (SIR), and signal-to-artifacts ratio (SAR), are commonly used in audio-visual self-supervised (AVSS) tasks, where the goal is to learn the coherent relationship between audio and visual features from unstructured and unlabeled data. AVSS methods are typically evaluated using three measures, as they provide a way to quantify the quality of the reconstructed audio signal and the separation performance of the model.

The signal distortion ratio (SDR) is a measure of the quality of a reconstructed signal compared to the original signal. It is defined as the ratio of signal power to distortion power and is typically used to evaluate the performance of audio separation algorithms.

$$\text{SDR} = 10\log_{10}\frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2} \tag{2}$$

where x_i is the original signal and $\hat{x}_i$ is the reconstructed signal.

The signal interference ratio (SIR) is a measure of the quality of a reconstructed signal compared to the interference signal, which is the unwanted part of the reconstructed signal. It is defined as the ratio of the signal power to the interference power and is often used to evaluate the performance of audio separation algorithms.

$$\text{SIR} = 10\log_{10}\frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(\hat{x}_i - z_i)^2} \tag{3}$$

where x_i is the original signal, $\hat{x}_i$ is the reconstructed signal, and z_i is the interference signal.

The signal to artifacts ratio (SAR) is a measure of the quality of a reconstructed signal compared to the artifacts, which are the errors or distortions introduced by the separation process. It is defined as the ratio of the signal power to the artifact power and is frequently used to evaluate the performance of audio separation algorithms.

$$\text{SAR} = 10\log_{10}\frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(\hat{x}_i - x_i)^2} \tag{4}$$

where x_i is the original signal and $\hat{x}_i$ is the reconstructed signal.

Among the metrics, SDR and SIR are more reliable than SAR. In our summary tables, we will report on just the signal-to-distortion ratio (SDR) metric, which is also the metric reported by most papers.

Localization in audio refers to the ability of a model to accurately identify the temporal locations of sound events in an audio signal. In the given scenario, the Consensus Intersection over Union (cIoU) and Area Under Curve (AUC) metrics are generally employed to evaluate the localization performance. The cIoU metric measures the degree of overlap between the predicted and ground-truth bounding boxes, while AUC measures the overall performance of the model in terms of true positive rate and false positive rate. To obtain the final localization map, a weighted summation is performed over valid categories using the normalized predicted probabilities as weights. This approach helps to give higher weightage to categories with higher confidence scores, thereby improving the accuracy of the final localization map. Final cIoU scores for each instrument/sound class on each frame is calculated by

$$cIoU\_class = \frac{\sum_{c=1}^{C} \theta_c cIoU_c}{\sum_{c=1}^{C} \theta_c} \tag{5}$$

The class index of instruments is represented by the variable 'c,' and the variable θ_c is set to 1 if the instrument of class 'c' makes sounds and 0 otherwise. By using this method, the evaluation score of cIoU class will only be high if the model is able to accurately associate specific classes

of instruments with the sounds they produce, thus ensuring that the evaluation metric is class-specific and accurate.

Object, action, and sound recognition can be measured using three measures: percentage accuracy, mean average precision (MAP), and area under the curve (AUC). Mean average precision (MAP) measures accuracy in a classification task. It is defined as the mean of the average precision scores for each class in the classification task.

The normalized mutual information (NMI) is a useful evaluation metric for comparing the performance of different clustering algorithms or for evaluating the performance of a clustering algorithm on a dataset. It is often used in evaluation measures for clustering tasks, particularly when the true cluster assignments are not known. NMI is a measure of the similarity between two clusterings of data. It is defined as the mutual information between the two clusterings, normalized by the average entropy of the clusterings. The normalized mutual information (NMI) is a metric that ranges from 0 to 1, with a higher value indicating a greater similarity between the two clusterings.

$$NMI = \frac{2I(C_1, C_2)}{H(C_1) + H(C_2)} \tag{6}$$

where $I(C_1, C_2)$ is the mutual information between the two clusterings, $H(C_1)$ is the entropy of the first clustering, and $H(C_2)$ is the entropy of the second clustering.

The mutual information between the two clusterings is calculated as:

$$I(C_1, C_2) = \sum_{c_1 \in C_1} \sum_{c_2 \in C_2} p(c_1, c_2) \log \frac{p(c_1, c_2)}{p(c_1)p(c_2)} \tag{7}$$

where C_1 and C_2 are the two clusterings, c_1 and c_2 are clusters in the respective clusterings, and $p(c_1)$, $p(c_2)$, and $p(c_1, c_2)$ are the probabilities of clusters c_1, c_2, and the joint probability of c_1 and c_2, respectively.

The entropy of clustering is calculated as follows:

$$H(C) = -\sum_{c \in C} p(c) \log p(c) \tag{8}$$

where C is the clustering and c is a cluster in the clustering.

C. DATASETS

We have gathered results reported on multiple datasets, including the highly regarded and challenging Audioset [62], VGG-Sound [83], and Kinetics [74]. Other datasets that we will consider include UCF-101 [53] and HMDB [109]. By comparing the performance of various approaches or models on these different datasets, we aim to provide a comprehensive understanding of the strengths and limitations of these methods in solving audio-visual tasks. It is important to consider a range of datasets in order to ensure the generalizability and robustness of these methods to diverse data.

In the domain of self-supervised learning AudioSet [62] is better for pre-training for tasks that involve sound classification/recognition. In many instances, the sounding object is not visible in AudioSet [62]. Thus making it better suited for classification or recognition task. Although the other datasets such as VGG-Sound [83], Kinetics-Sound [75], Flickr-Soundnet [64] doesn't contain so many different as Audioset [62] but the well alignment (both visible and audible) between the audio and video modality makes them better suited for other complex tasks. These tasks include Sound Source Localization, Audio-visual Source Separation, etc.

D. REPORTED PERFORMANCES

In this section, we have summarised quantitative performance on different classification, sound localization & separation, and clustering tasks. Then we have provided a conclusion by analyzing the quantitative results. For summarizing we have mainly focused on approaches that have used self-supervised pre-training. The quantitative results are taken from the reported results by the respective authors on the aforementioned different downstream tasks. We are mainly focusing on the self-supervised learning approaches because of the fact that self-supervised learning has become more popular than supervised approaches in recent times. Also, these self-supervised pre-training approaches have achieved performance that is on par with or better than the supervised approaches. Again the tasks that the supervised learning approaches have tried to solve are less popular than the ones targeted by self-supervised approaches. Also, in the most popular/challenging benchmarks/datasets, the self-supervised learning approaches have gained the best performances. Where as the supervised learning approaches show results on datasets that are not as large in terms of size and categories present. Considering all of the above facts we have put our focus on methods that were pre-trained using self-supervision.

Table 11 shows performance on a downstream object, sound, and action classification tasks. Note that the ESC-50 [233] and DCASE [234] sound datasets are relatively small, simple, and outdated compared to current sound datasets. Despite this, researchers have used these datasets to compare with approaches that came up before the introduction of the newer challenging datasets. The reasons for comparing ESC-50 [233] and DCASE [234] are twofold: 1. showing that training on newer datasets is better than training on older datasets, 2. proving the superiority of the new algorithm/approach over older approaches. The networks in table 11 were trained on different datasets such as Audioset [62], Kinetics [74], VGG-Sound [83] and SoundNet [4]. Later they are used for downstream classification tasks. By comparing the quantitative performances, we see that pre-training on Audioset [62] provides the best performance for downstream sound and action classification tasks.

Table [13] presents the performance of clustering approaches for self-labeling in the field of multi-modal analysis. However, they have yet to outperform contrastive learning approaches. However, the clustering approaches

TABLE 11: Performance comparison for object/action/sound classification task

Task	Method	Pre-training Dataset	Train-Test	mAP	%acc.
Object	Owen *et al.* [6], 2016	YFCC100M [91]	PASCAL VOC 07	47.4%	
			SUN397	21.4%	
Object	Arandjelovic & Zisserman [75], 2017	Flickr-SoundNet [64]	ImageNet		32.3%
Object	Afouras *et al.* [110], 2021	AudioSet-Instrument [103]	VGGSound [83]	52.3%	
			AudioSet [62]	44.3%	
			OpenImages (test)	39.9%	
Sound	Aytar *et al.* [4], 2016	SoundNet [4]	ESC [233]		74.2%
			DCASE [234]		88%
Sound	Arandjelovic & Zisserman [75], 2017	Flickr-SoundNet [64]	ESC [233]	79.3%	
			DCASE [234]	93%	
Sound	Korbar *et al.* [92], 2018	Audioset [62]	ESC [233]	76.7%	
			DCASE [234]	91%	
		Kinetics [74]	ESC [233]	80.6%	
			DCASE [234]	93%	
Sound	Morgado *et al.* [3], 2021	**Audioset** [62]	ESC [233]	**89.1%**	
			DCASE [234]	**96%**	
		Kinetics [74]	ESC [233]	79.1%	
			DCASE [234]	93%	
Action	Owens and Efros [94], 2018	Audioset [62]	UCF [53]	82.1%	
			HMDB [109]	-	
Action	Korbar *et al.* [92], 2018	Kinetics [74]	UCF [53]	85.8%	
			HMDB [109]	56.9%	
		Audioset [62]	UCF [53]	89.0%	
			HMDB [109]	61.6%	
Action	Morgado *et al.* [3], 2021	Kinetics [74]	UCF [53]	87.5%	
			HMDB [109]	60.8%	
		Audioset [62]	UCF [53]	**91.5%**	
			HMDB [109]	**64.7%**	
Action	Morgado *et al.* [118], 2021	Kinetics [74]	UCF [53]	85.6%	
			HMDB [109]	55.0%	
Action	Vedaldi *et al.* [112], 2021	VGG-Sound [83]	Flickr-SoundNet [64] (test)	0.590 (AUC)	

TABLE 12: Performance comparison for audio-visual sound source separation (AVSS) and sound source localization (SSL) tasks.

Task	Method	Train-Test	SDR/%acc
AVSS	Zhao *et al.* [95], 2018	MUSIC [95]	6.05 (SDR)
AVSS	Owens & Efros [94], 2018	VoxCeleb [70]	7.6 (SDR)
AVSS	Gao & Grauman [122], 2019	MUSIC [95]	7.64 (SDR)
		AudioSet [62]	4.26 (SDR)
AVSS	Aforous *et al.* [104], 2020	LRS2	10.8 (SDR)
AVSS	Chen *et al.* [167], 2023	MUSIC [95]	11.17 (SDR)
		AVE [55]	5.02 (SDR)
SSL	Aforous *et al.* [104], 2020	LRS2	99.6% (%acc)
		LRS3	99.7% (%acc)
		Columbia [108]	90.8 (F1)
SSL	Vedaldi *et al.* [112], 2021	VGG-SS [112]	0.382 (AUC)
SSL	Sun *et al.* [164], 2023	VGG-SS [112]	0.3729 (AUC)

TABLE 13: Performance comparison for just video (V), just audio (A), and audio-video (AV) clustering tasks.

Method	Train-Test	MI
Arandjelovic & Zisserman [75], 2017	Kinetics-Sound [75]	V:0.409
		A:0.330
Asano *et al.* [7], 2020	VGG-Sound [83]	V:0.528
		A:0.475
		AV:0.567
	Kinetics-Sound [75]	AV: 0.502

address some of the limitations of the contrastive learning methods, such as computation time and the need for a large number of negatives. Also, the cluster assignments can be used as week labels for solving other downstream [150] tasks.

We have summarized the performances reported by people on A-V Sound Source Separation (AVSS) and Sound Source Localization (SSL) tasks in the table 12. Most of the approaches reported here were trained on audio-visual data then, for they report these two downstream tasks on some well-known sound separation or localization benchmark datasets. In terms of numerical performance, the best SDR one can get is ∞. This is when the reconstructed when the original signal x_i and the reconstructed $\hat{x}_i$ signal are equal/same. But in terms of signal processing perspective, this can only happen when a network can truly separate the target/actual audio signal from the interference, noise, and added artifact signals. In the ideal case, it can be done by signal processing-inspired deep learning approaches. But so far we are far from achieving that.

VI. CURRENT GAPS AND FUTURE DIRECTIONS

The problems and gaps within current methodologies predominantly stem from certain oversimplified assumptions. A notable area for future exploration is the prediction of an object's material properties, an endeavor that has been minimally pursued due to data scarcity.

Current strategies for the identification of sounding objects within an image predominantly revolve around a multi-modal analysis pipeline. This process begins with the preliminary selection of potential sounding object candidates, which is typically accomplished using object detection technologies. Following this initial step, these strategies often involve calculating a score that reflects the similarity or correlation between sound and image features. This score is pivotal in pinpointing the precise-sounding objects or regions within an image. However, this reliance on object detection technologies as a first step presents a significant challenge. It heavily restricts the detection of silent objects or non-sounding regions within an image. Without the aid of such technologies, identifying these non-sounding elements becomes exceedingly difficult, if not outright unfeasible. This limitation underscores the heavy dependency on object detection technologies in current approaches, highlighting an area ripe for further research and development in the field.

Exploring temporal action segmentation in a multi-modal audio-visual context remains an area rich with research potential. One promising avenue is the utilization of audio data in conjunction with visual data. We posit that leveraging audio data can provide a form of regularization for visual network training. This approach can potentially streamline the segmentation process by using audio cues to guide and inform the visual analysis. Furthermore, incorporating foundational or intuitive knowledge into the segmentation process can significantly enhance performance. By integrating a deeper understanding of the relationships between different objects and their interactions within a scene, such knowledge can provide crucial context that aids in more accurate and efficient segmentation. This strategy not only helps tackle the inherent computational difficulties but also paves the way for more sophisticated and nuanced interpretations of audio-visual data in temporal action segmentation.

The advancement in the handling of multi-modal data, particularly the integration of audio and visual features, is a crucial area for development in the field of audio-visual learning. The current state of research in this domain has largely focused on two-stream networks, which have been the primary method for exploring audio-visual learning problems. However, this approach has its limitations, especially in terms of efficiency and adaptability to various applications.

Therefore, it is imperative to develop a standardized method for amalgamating audio and visual features. Pioneering new methods for integrating these multi-modal signals could lead to significant improvements in training efficiency. Such advancements would not only enhance the performance of audiovisual systems but also facilitate their deployment on edge devices. This is particularly relevant given the growing need for efficient, real-time processing in numerous applications.

Further, exploring innovative approaches, such as employing attention mechanisms or Vision Transformers (ViTs), may offer promising avenues for multi-modal data fusion. These technologies have the potential to revolutionize the way audio and visual data are integrated, leading to more sophisticated and effective audio-visual systems. Ultimately, such advancements could open new research frontiers and contribute substantially to the field of audio-visual learning.

Another critical gap in the audio-visual domain that still warrants attention from researchers is the challenge of continual learning & incremental learning. Although this problem has tracked some interest [294]–[296] in recent years, it is still far from solved. Continual learning refers to the ability of machine learning models to keep learning and adapting to new data without forgetting past information. In the context of audio-visual learning, continual learning is particularly important because of the influx of new data or information.

There are two ways to approach the ongoing challenges and opportunities in multi-modal research, particularly in audio-visual data. Firstly, one can focus on developing new approaches or methods for problems that have been already defined. This involves innovating novel techniques for self-supervision or crafting groundbreaking architectures. The objective here is to enhance and refine the existing methodologies for tackling specific problems which are already defined. Secondly, the focus can shift to addressing and attempting to solve new problems. This means employing audio-visual data to tackle issues that have not been previously approached in a multi-modal context.

A. NEW SOLUTIONS TO EXISTING PROBLEMS

Based on the consideration of gaps in research that we have observed in this review, we believe the following directions to be fruitful and important in the near future. They will lead to new architectures and approaches to problems that have been well studied for which datasets and benchmarks exist to measure progress.

- *New fusion architectures:* So far, people have looked at the audio-visual learning problems with two-stream networks. Finding smarter ways to fuse these multi-modal signals can improve the efficiency of the overall training procedure. Thus making the audio-visual system deployable in edge devices. This advancement could potentially unlock new avenues of opportunities for researchers. We believe that using an attention mechanism or utilizing ViT (Vision Transformer) could represent a novel and intelligent approach for the fusion of multi-modal data.
- *Improving self-supervised learning:* In the paradigm of self-supervised learning, pretext tasks AVC and AVS have traditionally been approached separately. However, recent studies by Tian *et al.* [138] have shown that SSLand AVSS can complement each other, resulting in better representation learning during pretext learning using the cyclic loss. This cyclic learning process may also benefit from curriculum learning techniques [297]. Solving AVC enables the network

to learn the correct association between audio and video. While AVS teaches the network to learn the correct temporal association, emphasizing the association of occurrence between the two modalities. These components are essential for training models that can effectively capture the intricate relationships and dependencies between audio and visual signals in multimedia data. AVC and AVS can be utilized to create harder positive and negative pairs, which is a crucial aspect of pretext learning in self-supervised learning. On the other hand, the complexity of the task dictates the training order in curriculum learning. So a technique for mining harder negatives could be beneficial in self-supervised learning approaches when used in conjunction with curriculum learning.

- *Localizing multiple sources:* Self-supervised learning has become the preferred approach for audio-visual learning. But almost all self-supervised learning approaches suffer from one of two problems: the inability to localize multiple sound sources in the same scene or to localize silent objects in the scene. People have tried using object detection or region proposal algorithms to address these issues, but the overall performance of self-supervised learning depends heavily on the region proposal algorithm. Additionally, these object detection models may need to be trained separately if they do not perform well on the training dataset, leading to many false positives. Retraining region proposals or object detection algorithms separately also increase training time.

  Most of the current methods typically rely on an object detector to select sound object candidates at the beginning of the multi-modal analysis pipeline. Similarly, silent object detection also relies heavily on object detectors, and it is difficult or almost impossible to find objects that do not make sound without using a pre-trained object detection model. Present methods for sound source localization often calculate a similarity or correlation score between sound and image features to identify sounding objects or regions in an image. Consequently, these methods can not locate objects that are not making any sound.
- *Mono to bin-aural audio:* In recent years people [131], [154] have started looking into learning spatial features from bin-aural or multi-channel audio. However, the lack of a multi-channel audio dataset has been a hindrance in the path to progress. Especially, there is a lack of multi-channel audio datasets that are collected in an unconstrained environment. Again there have been some [72], [119], [123] approaches to convert mono-aural audio to bin-aural/stereo audio. We believe this could be an interesting field of research, converting regular audio to bin-aural audio and then using the converted audio to learn spatial features from the new multi-channel audio.

B. NEW OR UNEXPLORED PROBLEM CONTEXTS

The following are some of the problem contexts that need focus from researchers in the future. Some of these are very new contexts, and some of them are old problems that are waiting to be considered again with new tools from deep learning.

- *3D action recognition:* 3D action recognition involves understanding human actions in three-dimensional space. It can be applied for solving tasks such as video surveillance and virtual reality. It is a task that prominently relies on visual cues. But audio can be used as a complimentary modality to aid in localizing where the action is happening in the 3D space.
- *Speaker diarization:* Speaker diarization is the process of segmenting the audio stream into speaker-specific segments. This involves identifying and distinguishing each speaker separately. Speaker diarization is applicable in transcription, sentiment analysis, and understanding multi-speaker conversations. This problem was more popular in the pre-deep learning era. However, in recent years, the problem still lacks the necessary attention and focus from the researchers.
- *Predicting material properties of an object*: The main focus of this problem is predicting the material property from visual and audio cues. The audio is captured when hitting the objects with another object. This problem has applications in industry for tasks like quality control, and material science.
- *Temporal action segmentation*: Temporal action segmentation is the process of segmenting video sequences into meaningful events or segments. This problem was discussed in the section IV-I. As can be seen from our discussion above temporal action segmentation problem hasn't been very popular in the audio-visual domain. Most people approach this problem using only visual modality.

  We believe that audio data can be used to regularize the training of the visual network for temporal action segmentation, and incorporating prior or common sense knowledge can also improve the segmentation task's performance by learning relationships between different objects.
- *Detecting sound source which is not present in the scene*: This task is especially a challenging one. This is also related to the sound source localization and multi-source detection problem. In most cases, people approach the problem of SSL assuming that there is only one dominant source in the video or scene and that the source is both visible and audible at the same time. There hasn't been any effort trying to locate (identifying the type of object) sounding objects that are not visible within the camera view.
- *Wildlife Biometrics*: Animal-ID has been popular in the image-only domain, where the data is collected using camera traps. To this day, there is no audio-visual

dataset for this type of task. Identifying wildlife can contribute to ecological research, conservation efforts, and monitoring wildlife populations by providing a non-intrusive method for animal identification.

- *Audio-visual analysis in camera network*: Audio-visual analysis in camera networks integrates information from both visual and auditory sources across a network of audio & visual sensors. This interdisciplinary approach has the potential to enhance surveillance and monitoring, the autonomous vehicle industry, and edge computing, thus creating a more robust and comprehensive analysis of events. This is a problem that has a lot of potential and prospective applications. However, the challenge lies in developing a network capable of handling large streams of data while remaining efficient and lightweight enough to be deployed on power-constrained devices.

The above remains an open problem for the researchers because of some assumptions that lead to the problems and gaps in the current works. Among the prospective problems in future research, predicting the material properties of an object has been attempted relatively less due to a lack of data.

In summary, future work in this field should also focus on improving current self-supervised learning methods for audio-visual learning rather than only attempting to improve the performance of some downstream tasks. The focus should be on improving the self-supervised pre-training methods so that the models learn better representations. On the other hand, as discussed above, certain persistent challenges (revisiting old problems) and untapped opportunities (exploring new problems) still await interest from researchers.

VII. CONCLUSION

We have conducted an extensive survey of the historical and current technical approaches in the domain of audio-visual learning, along with an overview of relevant datasets. Though researched for an extended period, this area has witnessed significant advancements with the rise of deep learning techniques. In most scenarios, multi-modal data has proven to be more advantageous than uni-modal data. Nonetheless, certain tasks, such as image frame classification and object detection, yield optimal results with only the visual modality, especially when human annotations are available. However, this is not the case with solely audio-based modality.

Both audio and visual modalities complement each other, particularly in unsupervised or self-supervised techniques. Typically, multi-modal analysis is employed in self-supervised learning to pre-train a network on a pseudo or pretext task. These networks, once pre-trained, can be adapted for various downstream tasks, even those focusing on a single modality. Impressively, these networks often achieve performance levels comparable to those trained through supervised methods. Yet, despite these advancements, there remains ample room for progress, especially in the realm of self-supervised learning, given the labor-intensive nature of labeling video data.

In section II, we delineated the fundamental technical tasks intrinsic to the audio-visual domain and illustrated how they address a plethora of real-world challenges. Section IV delved into the computational methodologies employed to tackle these tasks. We've categorized these methods based on their timeline and techniques into two main types: pre-deep learning and deep learning.

The paper underscores that the future trajectory of audiovisual learning research hinges on the exploration of novel applications. There is increasing interest in leveraging audiovisual learning for tasks like speaker identification, emotion recognition, and action recognition. These research avenues stand to gain immensely from multimodal data, warranting further investigation.

Our findings also reveal that the choice of dataset greatly influences the downstream objective. Thus, selecting the appropriate pretext task and pre-training dataset with the final objective in consideration is crucial. Section III sheds light on prominent datasets, their applications, and associated statistics. Furthermore, in section V, we have encapsulated the reported quantitative performances, laying special emphasis on deep learning-based self-supervised techniques, which presently dominate the challenges in audio-visual data.

In summation, the paper provides a comprehensive overview of the current state of audio-visual learning research, highlighting its successes and limitations. The findings suggest that there is still much to be explored and discovered in this field and that future research should focus on developing more efficient and effective approaches for extracting and utilizing information from multi-modal data.

ACKNOWLEDGMENT

This research was supported in part by the US National Science Foundation grant IIS 1956050.

During the preparation of this work we used ChatGPT and Grammarly to edit the text that we had drafted. After using these tools, we reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," Nature, vol. 264, no. 5588, pp. 746–748, 1976. 1, 3

[2] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, P. Group et al., "Preferred reporting items for systematic reviews and meta-analyses: the prisma statement," International journal of surgery, vol. 8, no. 5, pp. 336–341, 2010. 2

[3] P. Morgado, N. Vasconcelos, and I. Misra, "Audio-visual instance discrimination with cross-modal agreement," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12 475–12 486. 2, 6, 18, 23

[4] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," Advances in neural information processing systems, vol. 29, pp. 892–900, 2016. 2, 6, 9, 11, 13, 15, 19, 22, 23

[5] C. Gan, H. Zhao, P. Chen, D. Cox, and A. Torralba, "Self-supervised moving vehicle tracking with stereo sound," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7053–7062. 2, 7, 8, 10, 19

[6] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," in European conference on computer vision. Springer, 2016, pp. 801–816. 2, 6, 13, 18, 23

[7] Y. M. Asano, M. Patrick, C. Rupprecht, and A. Vedaldi, "Labelling unlabelled videos from scratch with multi-modal self-supervision," in NeurIPS, 2020. 2, 6, 7, 13, 18, 23

[8] D. Hu, F. Nie, and X. Li, "Deep multimodal clustering for unsupervised audiovisual learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9248–9257. 2, 6, 8, 9, 10, 11, 14, 15, 18

[9] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, and R. He, "Deep audio-visual learning: A survey," International Journal of Automation and Computing, vol. 18, pp. 351–376, 2021. 3

[10] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1368–1396, 2021. 3

[11] L. He, M. Niu, P. Tiwari, P. Marttinen, R. Su, J. Jiang, C. Guo, H. Wang, S. Ding, Z. Wang et al., "Deep learning for depression recognition with audiovisual cues: A review," Information Fusion, vol. 80, pp. 56–86, 2022. 3

[12] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE transactions on pattern analysis and machine intelligence, vol. 31, no. 1, pp. 39–58, 2008. 3, 10, 19, 20

[13] N. J. Shoumy, L.-M. Ang, K. P. Seng, D. M. Rahaman, and T. Zia, "Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals," Journal of Network and Computer Applications, vol. 149, p. 102447, 2020. 3

[14] A. K. Katsaggelos, S. Bahaadini, and R. Molina, "Audiovisual fusion: Challenges and new approaches," Proceedings of the IEEE, vol. 103, no. 9, pp. 1635–1653, 2015. 3

[15] C. Chen, M. Song, W. Song, L. Guo, and M. Jian, "A comprehensive survey on video saliency detection with auditory information: The audio-visual consistency perceptual is the key!" IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 2, pp. 457–477, 2023. 3

[16] Z. Akhtar and T. H. Falk, "Audio-visual multimedia quality assessment: A comprehensive survey," IEEE Access, vol. 5, pp. 21 090–21 117, 2017. 3

[17] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," Issues in visual and audio-visual speech processing, vol. 22, p. 23, 2004. 3

[18] S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, and X. Zhao, "Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects," Expert Systems with Applications, p. 121692, 2023. 3

[19] R. Campbell, "The processing of audio-visual speech: empirical and neural bases," Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 363, no. 1493, pp. 1001–1010, 2008. 3

[20] E. Kidron, Y. Schechner, and M. Elad, "Pixels that sound," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, 2005, pp. 88–95 vol. 1. 4, 8, 13, 14

[21] M. Beal, N. Jojic, and H. Attias, "A graphical model for audiovisual object tracking," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 7, pp. 828–836, 2003. 4, 8, 10, 16

[22] M. J. Beal, H. Attias, and N. Jojic, "Audio-video sensor fusion with probabilistic graphical models," in Computer Vision — ECCV 2002, A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 736–750. 4, 8, 10, 16

[23] S. Ben-Yacoub, J. Luttin, K. Jonsson, J. Matas, and J. Kittler, "Audio-visual person verification," in Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), vol. 1, 1999, pp. 580–585 Vol. 1. 4, 8, 14

[24] J. W. Fisher and T. Darrell, "Probabalistic models and informative subspaces for audiovisual correspondence," in Computer Vision — ECCV 2002, A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 592–603. 4, 9

[25] M. Naphade and T. Huang, "Recognizing high-level audio-visual concepts using context," in Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205), vol. 3, 2001, pp. 46–49 vol.3. 4, 10, 16

[26] ——, "Detecting semantic concepts using context and audiovisual features," in Proceedings IEEE Workshop on Detection and Recognition of Events in Video, 2001, pp. 92–98. 4, 10, 16

[27] V. Kulesh, V. Petrushin, and I. Sethi, "Video clip recognition using joint audio-visual processing model," in Object recognition supported by user interaction for service robots, vol. 1, 2002, pp. 500–503 vol.1. 4, 10, 16

[28] H. Hung, Y. Huang, C. Yeo, and D. Gatica-Perez, "Associating audio-visual activity cues in a dominance estimation framework," in 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008, pp. 1–6. 4, 8

[29] Y. Huang, O. Vinyals, G. Friedland, C. Muller, N. Mirghafori, and C. Wooters, "A fast-match approach for robust, faster than real-time speaker diarization," in 2007 IEEE Workshop on Automatic Speech Recognition Understanding (ASRU), 2007, pp. 693–698. 4

[30] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos et al., "The ami meeting corpus," in Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research, vol. 88. Citeseer, 2005, p. 100. 4

[31] J. Nam, M. Alghoniemy, and A. Tewfik, "Audio-visual content-based violent scene characterization," in Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269), vol. 1, 1998, pp. 353–357 vol.1. 4, 10

[32] Z. Rasheed and M. Shah, "Movie genre classification by exploiting audio-visual features of previews," in Object recognition supported by user interaction for service robots, vol. 2, 2002, pp. 1086–1089 vol.2. 4, 10

[33] J. Vermaak, M. Gangnet, A. Blake, and P. Perez, "Sequential monte carlo fusion of sound and vision for speaker tracking," in Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, vol. 1, 2001, pp. 741–746 vol.1. 4, 8, 10, 13, 16

[34] S. Ravulapalli and S. Sarkar, "Association of sound to motion in video using perceptual organization," in 18th International Conference on Pattern Recognition (ICPR'06), vol. 1, 2006, pp. 1216–1219. 4, 9, 14

[35] H. Vajaria, T. Islam, S. Sarkar, R. Sankar, and R. Kasturi, "Audio segmentation and speaker localization in meeting videos," in 18th International Conference on Pattern Recognition (ICPR'06), vol. 2, 2006, pp. 1150–1153. 4, 8, 13, 14

[36] H. Vajaria, S. Sarkar, and R. Kasturi, "Clip retrieval using multi-modal biometrics in meeting archives," in 2008 19th International Conference on Pattern Recognition, 2008, pp. 1–4. 4, 10, 13, 14

[37] ——, "Exploring co-occurence between speech and body movement for audio-guided video localization," vol. 18, no. 11, 2008, pp. 1608–1617. 4, 8, 13, 14

[38] J. W. Fisher III, T. Darrell, W. T. Freeman, and P. A. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in Advances in neural information processing systems, 2001, pp. 772–778. 4, 8, 13, 14, 17

[39] J. Hershey and J. Movellan, "Audio vision: Using audio-visual synchrony to locate sounds," in Advances in Neural Information Processing Systems, S. Solla, T. Leen, and K. Müller, Eds., vol. 12. MIT Press, 2000. [Online]. Available: https://proceedings.neurips.cc/paper/1999/file/b618c3210e934362ac261db280128c22-Paper.pdf 4, 8, 13, 14

[40] F. D. M. de Souza, S. Sarkar, and G. Cámara-Chávez, "Building semantic understanding beyond deep learning from sound and vision," in 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 2097–2102. 5, 9, 10

[41] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in European conference on computer vision. Springer, 2004, pp. 25–36. 5

[42] F. De la Torre, J. Hodgins, A. Bargteil, X. Martin, J. Macey, A. Collado, and P. Beltran, "Guide to the carnegie mellon university multimodal activity (cmu-mmac) database," 2009. 5

[43] U. Grenander, Elements of pattern theory. JHU Press, 1996. 5

[44] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg, "Visual to sound: Generating natural sound for videos in the wild," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018. 5, 13, 20

[45] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "Samplernn: An unconditional end-to-end

neural audio generation model," arXiv preprint arXiv:1612.07837, 2016. 5, 15, 20

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015. 5, 6, 7, 15, 20

[47] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman, "Soundspaces: Audio-visual navigation in 3d environments," in ECCV, 2020. 5

[48] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," arXiv preprint arXiv:1709.06158, 2017. 5

[49] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. Engel, R. Mur-Artal, C. Ren, S. Verma, A. Clarkson, M. Yan, B. Budge, Y. Yan, X. Pan, J. Yon, Y. Zou, K. Leon, N. Carter, and R. Newcombe, "The replica dataset: A digital replica of indoor spaces," 06 2019. 5

[50] C. Chen, Z. Al-Halah, and K. Grauman, "Semantic audio-visual navigation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15 516–15 525. 5

[51] M. Subedar, R. Krishnan, P. L. Meyer, O. Tickoo, and J. Huang, "Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6301–6310. 5, 9, 13, 17

[52] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfreund, C. Vondrick et al., "Moments in time dataset: one million videos for event understanding," IEEE transactions on pattern analysis and machine intelligence, vol. 42, no. 2, pp. 502–508, 2019. 5

[53] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," 2012. 5, 6, 12, 13, 22, 23

[54] Y. Wu, L. Zhu, Y. Yan, and Y. Yang, "Dual attention matching for audio-visual event localization," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6292–6300. 5, 10, 13

[55] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 247–263. 5, 6, 9, 11, 12, 17, 20, 23

[56] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, "Epic-fusion: Audio-visual temporal binding for egocentric action recognition," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5492–5501. 5, 13, 16, 17

[57] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price et al., "Scaling egocentric vision: The epic-kitchens dataset," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 720–736. 5, 7, 11, 12

[58] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in International conference on machine learning. PMLR, 2015, pp. 448–456. 5

[59] Y. Liu, M. Qiao, M. Xu, B. Li, W. Hu, and A. Borji, "Learning to predict salient faces: A novel visual-audio saliency model," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. Springer, 2020, pp. 413–429. 5, 9, 13, 15

[60] R. Qian, D. Hu, H. Dinkel, M. Wu, N. Xu, and W. Lin, "Multiple sound sources localization from coarse to fine," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. Springer, 2020, pp. 292–308. 5, 8, 9, 13, 14, 15

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778. 5, 6, 7, 8, 15

[62] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 776–780. 5, 6, 7, 11, 12, 17, 20, 22, 23

[63] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 11, pp. 2298–2304, 2016. 5, 15

[64] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. S. Kweon, "Learning to localize sound source in visual scenes," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4358–4366. 5, 6, 7, 11, 12, 13, 19, 22, 23

[65] J. Ramaswamy and S. Das, "See the sound, hear the pixels," in 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2959–2968. 5, 8, 9, 11, 13, 14, 15

[66] A. Tsiami, P. Koutras, and P. Maragos, "Stavis: Spatio-temporal audiovisual saliency network," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4766–4776. 5, 8, 9, 13, 14, 15

[67] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2405–2413. 5, 13

[68] F. Shi, J. Guo, H. Zhang, S. Yang, X. Wang, and Y. Guo, "Glavnet: Global-local audio-visual cues for fine-grained material recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14 433–14 442. 5, 9

[69] J. Lee, S.-W. Chung, S. Kim, H.-G. Kang, and K. Sohn, "Looking into your speech: Learning cross-modal affinity for audio-visual speech separation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1336–1345. 5, 8, 9, 15

[70] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," arXiv preprint arXiv:1706.08612, 2017. 5, 6, 12, 23

[71] J. Zhou, L. Zheng, Y. Zhong, S. Hao, and M. Wang, "Positive sample propagation along the audio-visual event line," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8436–8444. 5, 9

[72] X. Xu, H. Zhou, Z. Liu, B. Dai, X. Wang, and D. Lin, "Visually informed binaural audio generation without binaural audios," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15 485–15 494. 5, 25

[73] R. Gao, T.-H. Oh, K. Grauman, and L. Torresani, "Listen to look: Action recognition by previewing audio," 2020. 5, 9, 16

[74] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308. 5, 6, 7, 11, 12, 18, 22, 23

[75] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 609–617. 5, 6, 9, 10, 11, 12, 13, 15, 18, 22, 23

[76] B. Korbar, D. Tran, and L. Torresani, "Scsampler: Sampling salient clips from video for efficient action recognition," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6232–6242. 5, 9

[77] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in CVPR, 2014. 5

[78] N. Rai, H. Chen, J. Ji, R. Desai, K. Kozuka, S. Ishizaka, E. Adeli, and J. C. Niebles, "Home action genome: Cooperative compositional action understanding," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11 184–11 193. 5, 9

[79] X. Ji, H. Zhou, K. Wang, W. Wu, C. C. Loy, X. Cao, and F. Xu, "Audio-driven emotional video portraits," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14 080–14 089. 5, 9, 20

[80] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "Mead: A large-scale audio-visual dataset for emotional talking-face generation," in European Conference on Computer Vision. Springer, 2020, pp. 700–717. 5, 7

[81] R. Panda, C.-F. Chen, Q. Fan, X. Sun, K. Saenko, A. Oliva, and R. Feris, "Adamml: Adaptive multi-modal learning for efficient video recognition," 2021. 5, 10, 16

[82] Y. Chen, Y. Xian, A. Koepke, Y. Shan, and Z. Akata, "Distilling audio-visual knowledge by compositional contrastive learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7016–7025. 5, 10, 17

[83] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," 2020. 5, 6, 7, 11, 12, 13, 22, 23

[84] Y. Xia and Z. Zhao, "Cross-modal background suppression for audio-visual event localization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19 989–19 998. 5, 9

[85] H. Jiang, C. Murdock, and V. K. Ithapu, "Egocentric deep multi-channel audio-visual active speaker localization," in Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10 544–10 552. 5, 8

[86] J. Donley, V. Tourbabin, J.-S. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, and R. Mehra, "Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments," arXiv preprint arXiv:2107.04174, 2021. 5

[87] G. Li, Y. Wei, Y. Tian, C. Xu, J.-R. Wen, and D. Hu, "Learning to answer questions in dynamic audio-visual scenarios," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19 108–19 118. 5, 10

[88] Z. Yang, X. Fan, V. Isler, and H. S. Park, "Posekernellifter: Metric lifting of 3d human pose using sound," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13 179–13 189. 5

[89] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014. 6

[90] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, vol. 25, pp. 1097–1105, 2012. 6

[91] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m," Communications of the ACM, vol. 59, no. 2, p. 64–73, Jan 2016. [Online]. Available: http://dx.doi.org/10.1145/2812802 6, 13, 23

[92] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," in Advances in Neural Information Processing Systems, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/c4616f5a24a66668f11ca4fa80525dc4-Paper.pdf 6, 9, 10, 11, 13, 18, 23

[93] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," 2018. 6, 15

[94] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 631–648. 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 19, 23

[95] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 570–586. 6, 7, 8, 10, 11, 12, 13, 14, 15, 23

[96] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015. 6, 7, 15

[97] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. S. Kweon, "Learning to localize sound source in visual scenes," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4358–4366. 6, 7, 8, 13, 14, 15

[98] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," 2018. 6

[99] R. Gao, R. Feris, and K. Grauman, "Learning to separate object sounds by watching unlabeled video," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 35–53. 6, 8, 11, 13, 15, 16

[100] H. Izadinia, I. Saleemi, and M. Shah, "Multimodal analysis for identification and segmentation of moving-sounding objects," IEEE Transactions on Multimedia, vol. 15, no. 2, pp. 378–390, 2012. 6, 13

[101] K. Li, J. Ye, and K. A. Hua, "What's making that sound?" New York, NY, USA: Association for Computing Machinery, 2014. [Online]. Available: https://doi.org/10.1145/2647868.2654936 6

[102] J. Pu, Y. Panagakis, S. Petridis, and M. Pantic, "Audio-visual object localization and separation using low-rank and sparsity," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 2901–2905. 6

[103] R. Arandjelovic and A. Zisserman, "Objects that sound," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 435–451. 6, 7, 8, 11, 13, 14, 15, 23

[104] T. Afouras, A. Owens, J. S. Chung, and A. 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. Springer, 2020, pp. 208–224. 6, 8, 9, 10, 11, 13, 14, 18, 23

[105] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," arXiv preprint arXiv:1405.3531, 2014. 6

[106] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," IEEE transactions on pattern analysis and machine intelligence, vol. 44, no. 12, pp. 8717–8727, 2018. 6, 7, 9

[107] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," arXiv preprint arXiv:1809.00496, 2018. 6, 9

[108] P. Chakravarty and T. Tuytelaars, "Cross-modal supervision for learning active speaker detection in video," in European Conference on Computer Vision. Springer, 2016, pp. 285–301. 6, 12, 23

[109] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in 2011 International conference on computer vision. IEEE, 2011, pp. 2556–2563. 6, 12, 13, 22, 23

[110] T. Afouras, Y. M. Asano, F. Fagan, A. Vedaldi, and F. Metze, "Self-supervised object detection from audio-visual correspondence," arXiv preprint arXiv:2104.06401, 2021. 6, 8, 10, 14, 23

[111] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," arXiv preprint arXiv:1506.01497, 2015. 6

[112] A. Vedaldi, H. Chen, W. Xie, T. Afouras, A. Nagrani, and A. Zisserman, "Localizing visual sounds the hard way," in Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). Institute of Electrical and Electronics Engineers, 2021. 6, 12, 13, 14, 18, 23

[113] A. Rouditchenko, H. Zhao, C. Gan, J. McDermott, and A. Torralba, "Self-supervised audio-visual co-segmentation," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 2357–2361. 6, 8, 11, 15, 16

[114] A. Nagrani, S. Albanie, and A. Zisserman, "Learnable pins: Cross-modal embeddings for person identity," 2018. 6, 9

[115] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE, 2018, pp. 67–74. 6

[116] Y. Tian, D. Li, and C. Xu, "Unified multisensory perception: weakly-supervised audio-visual video parsing," arXiv preprint arXiv:2007.10558, 2020. 6, 9

[117] Y. Wu and Y. Yang, "Exploring heterogeneous clues for weakly-supervised audio-visual video parsing," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 6, 9

[118] P. Morgado, I. Misra, and N. Vasconcelos, "Robust audio-visual instance discrimination," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12 934–12 945. 6, 9, 18, 23

[119] P. Morgado, N. Nvasconcelos, T. Langlois, and O. Wang, "Self-supervised generation of spatial audio for 360°video," in Advances in Neural Information Processing Systems, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/01161aaa0b6d1345dd8fe4e481144d84-Paper.pdf 6, 25

[120] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2462–2470. 6

[121] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7263–7271. 7

[122] R. Gao and K. Grauman, "Co-separating sounds of visual objects," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019. 7, 8, 11, 15, 16, 23

[123] ——, "2.5d visual sound," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019. 7, 8, 15, 16, 25

[124] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 649–665. 7, 9

[125] D. Hu, X. Li et al., "Temporal multimodal learning in audiovisual speech recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3574–3582. 7, 9, 11

[126] I. Matthews, T. Cootes, J. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 2, pp. 198–213, 2002. 7

[127] H. Bear, R. Harvey, B.-J. Theobald, and Y. Lan, "Which phoneme-to-viseme maps best improve visual-only computer lip-reading?" 12 2014. 7

[128] H. Zhao, C. Gan, W.-C. Ma, and A. Torralba, "The sound of motions," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1735–1744. 7, 9, 11, 14, 15, 16

[129] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8934–8943. 7

[130] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications," IEEE Transactions on Multimedia, vol. 21, no. 2, pp. 522–535, 2019. 7

[131] K. Yang, B. Russell, and J. 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, 2020, pp. 9932–9941. 7, 8, 14, 15, 16, 25

[132] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141. 7

[133] X. Yang, P. Ramesh, R. Chitta, S. Madhvanath, E. A. Bernal, and J. Luo, "Deep multimodal representation learning from temporal data," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5447–5455. 7, 9, 17, 19

[134] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "Cuave: A new audio-visual database for multimodal human-computer interface research," in 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, 2002, pp. II–2017–II–2020. 7

[135] N. Khosravan, S. Ardeshir, and R. Puri, "On attention modules for audio-visual synchronization." in CVPR Workshops, 2019, pp. 25–28. 7, 9, 11, 14, 15

[136] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," arXiv preprint arXiv:2101.03149, 2021. 7, 8

[137] J. Son Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," arXiv e-prints, p. arXiv:1806.05622, Jun. 2018. 7

[138] Y. Tian, D. Hu, and C. Xu, "Cyclic co-learning of sounding object visual grounding and sound separation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2745–2754. 7, 8, 15, 24

[139] H. Xuan, Z. Wu, J. Yang, Y. Yan, and X. Alameda-Pineda, "A proposal-based paradigm for self-supervised sound source localization in videos," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1029–1038. 7, 8

[140] D. Hu, R. Qian, M. Jiang, X. Tan, S. Wen, E. Ding, W. Lin, and D. Dou, "Discriminative sounding objects localization via self-supervised audio-visual matching," Advances in Neural Information Processing Systems, vol. 33, pp. 10 077–10 087, 2020. 7

[141] Y. Zhang, H. Doughty, L. Shao, and C. G. Snoek, "Audio-adaptive activity recognition across video domains," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13 791–13 800. 7, 9

[142] Y. Zhou, J. Yang, D. Li, J. Saito, D. Aneja, and E. Kalogerakis, "Audio-driven neural gesture reenactment with video motion graphs," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3418–3428. 7, 20

[143] A. Siarohin, O. J. Woodford, J. Ren, M. Chai, and S. Tulyakov, "Motion representations for articulated animation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13 653–13 662. 7

[144] O.-B. Mercea, L. Riesch, A. Koepke, and Z. Akata, "Audio-visual generalised zero-shot learning with cross-modal attention and language," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10 553–10 563. 7

[145] B. Liang, Y. Pan, Z. Guo, H. Zhou, Z. Hong, X. Han, J. Han, J. Liu, E. Ding, and J. Wang, "Expressive talking head generation with granular audio-visual control," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3387–3396. 7, 10

[146] E. Burkov, I. Pasechnik, A. Grigorev, and V. Lempitsky, "Neural head reenactment with latent pose descriptors," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 13 786–13 795. 7, 20

[147] R. Zellers, J. Lu, X. Lu, Y. Yu, Y. Zhao, M. Salehi, A. Kusupati, J. Hessel, A. Farhadi, and Y. Choi, "Merlot reserve: Neural script knowledge through vision and language and sound," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16 375–16 387. 7, 9

[148] X. Hu, Z. Chen, and A. Owens, "Mix and localize: Localizing sound sources in mixtures," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10 483–10 492. 7, 8

[149] A. Jabri, A. Owens, and A. Efros, "Space-time correspondence as a contrastive random walk," in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 19 545–19 560. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/e2ef524fbf3d9fe611d5a8e90fefdc9c-Paper.pdf 7

[150] T. Afouras, Y. M. Asano, F. Fagan, A. Vedaldi, and F. Metze, "Self-supervised object detection from audio-visual correspondence," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10 575–10 586. 7, 9, 23

[151] Z. Song, Y. Wang, J. Fan, T. Tan, and Z. Zhang, "Self-supervised predictive learning: A negative-free method for sound source localization in visual scenes," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3222–3231. 7, 8

[152] X. Chen and K. He, "Exploring simple siamese representation learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15 750–15 758. 7

[153] A. B. Vasudevan, D. Dai, and L. Van Gool, "Sound and visual representation learning with multiple pretraining tasks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14 616–14 626. 7, 8

[154] A. B. Vasudevan, D. Dai, and L. V. Gool, "Semantic object prediction and spatial sound super-resolution with binaural sounds," in European conference on computer vision. Springer, 2020, pp. 638–655. 7, 25

[155] S. H. Lee, W. Roh, W. Byeon, S. H. Yoon, C. Kim, J. Kim, and S. Kim, "Sound-guided semantic image manipulation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3377–3386. 7, 9, 11

[156] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 8110–8119. 7

[157] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401–4410. 7

[158] S. Lee, H.-I. Kim, and Y. M. Ro, "Weakly paired associative learning for sound and image representations via bimodal associative memory," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10 534–10 543. 7

[159] V. Sanguineti, P. Morerio, N. Pozzetti, D. Greco, M. Cristani, and V. Murino, "Leveraging acoustic images for effective self-supervised audio representation learning," in European Conference on Computer Vision. Springer, 2020, pp. 119–135. 7

[160] W. Pan, H. Shi, Z. Zhao, J. Zhu, X. He, Z. Pan, L. Gao, J. Yu, F. Wu, and Q. Tian, "Wnet: Audio-guided video object segmentation via wavelet-based cross-modal denoising networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1320–1331. 7, 9

[161] K. Yang, D. Marković, S. Krenn, V. Agrawal, and A. Richard, "Audio-visual speech codecs: Rethinking audio-visual speech enhancement by re-synthesis," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8227–8237. 7

[162] S. Mo and Y. Tian, "Audio-visual grouping network for sound localization from mixtures," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10 565–10 574. 8

[163] C. Huang, Y. Tian, A. Kumar, and C. Xu, "Egocentric audio-visual object localization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 22 910–22 921. 8, 9, 11

[164] W. Sun, J. Zhang, J. Wang, Z. Liu, Y. Zhong, T. Feng, Y. Guo, Y. Zhang, and N. Barnes, "Learning audio-visual source localization via false negative aware contrastive learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 6420–6429. 8, 18, 23

[165] X. Zhou, D. Zhou, D. Hu, H. Zhou, and W. Ouyang, "Exploiting visual context semantics for sound source localization," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), January 2023, pp. 5199–5208. 8

[166] D. Fedorishin, D. D. Mohan, B. Jawade, S. Setlur, and V. Govindaraju, "Hear the flow: Optical flow-based self-supervised visual sound source localization," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), January 2023, pp. 2278–2287. 8

[167] J. Chen, R. Zhang, D. Lian, J. Yang, Z. Zeng, and J. Shi, "iquery: Instruments as queries for audio-visual sound separation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 14 675–14 686. 8, 23

[168] R. Tan, A. Ray, A. Burns, B. A. Plummer, J. Salamon, O. Nieto, B. Russell, and K. Saenko, "Language-guided audio-visual source separation via trimodal consistency," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 10 575–10 584. 8

[169] C. Gungor and A. Kovashka, "Complementary cues from audio help combat noise in weakly-supervised object detection," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), January 2023, pp. 2185–2194. 9

[170] M. Chen, L. Xing, Y. Wang, and Y. Zhang, "Enhanced multimodal representation learning with cross-modal kd," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 11 766–11 775. 9, 10

[171] T. Mahmud and D. Marculescu, "Ave-clip: Audioclip-based multi-window temporal transformer for audio visual event localization," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), January 2023, pp. 5158–5167. 9

[172] M. Cristani, M. Bicego, and V. Murino, "Audio-visual event recognition in surveillance video sequences," IEEE Transactions on Multimedia, vol. 9, no. 2, pp. 257–267, 2007. 9

[173] Y.-L. Kang, J.-H. Lim, M. Kankanhalli, C.-S. Xu, and Q. Tian, "Goal detection in soccer video using audio/visual keywords," in 2004 International Conference on Image Processing, 2004. ICIP '04., vol. 3, 2004, pp. 1629–1632 Vol. 3. 9

[174] J. Hong, M. Kim, J. Choi, and Y. M. Ro, "Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 18 783–18 794. 9

[175] M. Burchi and R. Timofte, "Audio-visual efficient conformer for robust speech recognition," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 2258–2267. 9

[176] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," arXiv preprint arXiv:1904.08779, 2019. 9

[177] Y. Li, Y. Wang, and Z. Cui, "Decoupled multimodal distilling for emotion recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 6631–6640. 9, 10

[178] Z. Zeng, Y. Hu, G. Roisman, Z. Wen, and Y. Fu, "Audio-visual spontaneous emotion recognition," 01 2007, pp. 72–90. 9, 10, 19, 20

[179] Z. Zeng, Y. Hu, M. Liu, Y. Fu, and T. S. Huang, "Training combination strategy of multi-stream fused hidden markov model for audio-visual affect recognition," in Proceedings of the 14th ACM International Conference on Multimedia, ser. MM '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 65–68. [Online]. Available: https://doi.org/10.1145/1180639.1180661 9, 10, 19, 20

[180] Y. Wang and L. Guan, "Recognizing human emotion from audiovisual information," in Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., vol. 2, 2005, pp. ii/1125–ii/1128 Vol. 2. 9, 10, 19

[181] M. Song, J. Bu, C. Chen, and N. Li, "Audio-visual based emotion recognition - a new approach," in Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., vol. 2, 2004, pp. II–II. 9, 10, 19, 20

[182] N. Sebe, I. Cohen, T. Gevers, and T. Huang, "Emotion recognition based on joint visual and audio cues," in 18th International Conference on Pattern Recognition (ICPR'06), vol. 1, 2006, pp. 1136–1139. 9, 10, 19

[183] B. Schuller, R. Müeller, B. Höernler, A. Höethker, H. Konosu, and G. Rigoll, "Audiovisual recognition of spontaneous interest within conversations," in Proceedings of the 9th International Conference on Multimodal Interfaces, ser. ICMI '07. New York, NY, USA: Association for Computing Machinery, 2007, p. 30–37. [Online]. Available: https://doi.org/10.1145/1322192.1322201 9, 10, 19, 20

[184] S. Petridis and M. Pantic, "Audiovisual discrimination between laughter and speech," in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, pp. 5117–5120. 9, 10, 19

[185] P. Pal, A. Iyer, and R. Yantorno, "Emotion detection from infant facial expressions and cries," in 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, vol. 2, 2006, pp. II–II. 9, 10, 19

[186] K. Karpouzis, G. Caridakis, L. Kessous, N. Amir, A. Raouzaiou, L. Malatesta, and S. Kollias, "Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition," in Artifical Intelligence for Human Computing, T. S. Huang, A. Nijholt, M. Pantic, and A. Pentland, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 91–112. 9, 10, 19

[187] H.-J. Go, K.-C. Kwak, D.-J. Lee, and M.-G. Chun, "Emotion recognition from the facial image and speech signal," in SICE 2003 Annual Conference (IEEE Cat. No.03TH8734), vol. 3, 2003, pp. 2890–2895 Vol.3. 9, 10, 19

[188] N. Fragopanagos and J. Taylor, "Emotion recognition in human–computer interaction," Neural Networks, vol. 18, no. 4, pp. 389–405, 2005, emotion and Brain. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608005000390 9, 10, 19

[189] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Raouzaiou, and K. Karpouzis, "Modeling naturalistic affective states via facial and vocal expressions recognition," in Proceedings of the 8th International Conference on Multimodal Interfaces, ser. ICMI '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 146–154. [Online]. Available: https://doi.org/10.1145/1180995.1181029 9, 10, 19

[190] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in Proceedings of the 6th International Conference on Multimodal Interfaces, ser. ICMI '04. New York, NY, USA: Association for Computing Machinery, 2004, p. 205–211. [Online]. Available: https://doi.org/10.1145/1027933.1027968 9, 10, 19, 20

[191] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in Proceedings of the 5th international workshop on audio/visual emotion challenge, 2015, pp. 3–8. 9, 10, 19

[192] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "Lstm-modeling of continuous emotions in an audiovisual affect recognition framework," Image and Vision Computing, vol. 31, no. 2, pp. 153–163, 2013. 9, 10, 19

[193] P. Chakravarthula, J. A. D'Souza, E. Tseng, J. Bartusek, and F. Heide, "Seeing with sound: Long-range acoustic beamforming for multimodal scene understanding," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 982–991. 9, 10

[194] J. Xiong, G. Wang, P. Zhang, W. Huang, Y. Zha, and G. Zhai, "Casp-net: Rethinking video saliency prediction from an audio-visual consistency perceptual perspective," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 6441–6450. 9

[195] K. Heidler, L. Mou, D. Hu, P. Jin, G. Li, C. Gan, J.-R. Wen, and X. X. Zhu, "Self-supervised audiovisual representation learning for remote sensing data," International Journal of Applied Earth Observation and Geoinformation, vol. 116, p. 103130, 2023. 9, 11

[196] C. Feng, Z. Chen, and A. Owens, "Self-supervised video forensics by audio-visual anomaly detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 10 491–10 503. 9

[197] Y.-B. Lin, Y.-L. Sung, J. Lei, M. Bansal, and G. Bertasius, "Vision transformers are parameter-efficient audio-visual learners," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 2299–2309. 9

[198] Z. Lin, S. Yu, Z. Kuang, D. Pathak, and D. Ramanan, "Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 19 325–19 337. 9, 20

[199] Y. Zhang and J. Li, "Birdsoundsdenoising: Deep visual audio denoising for bird sounds," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), January 2023, pp. 2248–2257. 9

[200] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in Workshop on Multi-view Lip-reading, ACCV, 2016. 9

[201] S.-W. Chung, J. S. Chung, and H.-G. Kang, "Perfect match: Improved cross-modal embeddings for audio-visual synchronisation," in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 3965–3969. 9

[202] T. D. C. Little, C. Y. R. Chen, C. Chang, and P. B. Berra, "Multimedia synchronization," IEEE Data Eng. Bull., vol. 14, no. 3, pp. 26–35, 1991. 9

[203] Y. Liu and Y. Sato, "Recovery of audio-to-video synchronization through analysis of cross-modality correlation," Pattern Recognition Letters, vol. 31, no. 8, pp. 696–701, 2010, award winning papers from the 19th International Conference on Pattern Recognition (ICPR). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865509001925 9

[204] P. Shrestha, M. Barbieri, H. Weda, and D. Sekulovski, "Synchronization of multiple camera videos using audio-visual features," IEEE Transactions on Multimedia, vol. 12, no. 1, pp. 79–92, 2010. 9

[205] A. Llagostera Casanovas and A. Cavallaro, "Audio-visual events for multi-camera synchronization," Multimedia Tools and Applications, vol. 74, no. 4, pp. 1317–1340, 2015. 9

[206] E. Marcheret, G. Potamianos, J. Vopicka, and V. Goel, "Detecting audio-visual synchrony using deep neural networks," in Sixteenth Annual Conference of the International Speech Communication Association, 2015. 9

[207] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: Learning lip sync from audio," ACM Trans. Graph., vol. 36, no. 4, jul 2017. [Online]. Available: https://doi.org/10.1145/3072959.3073640 9

[208] T. Kikuchi and Y. Ozasa, "Watch, listen once, and sync: Audio-visual synchronization with multi-modal regression cnn," in 2018 IEEE international conference on acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 3036–3040. 9

[209] J. Wang, Z. Fang, and H. Zhao, "Alignnet: A unifying approach to audio-visual alignment," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), March 2020. 9

[210] A. Senocak, J. Kim, T.-H. Oh, D. Li, and I. S. Kweon, "Event-specific audio-visual fusion layers: A simple and new perspective on video understanding," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), January 2023, pp. 2237–2247. 9

[211] J. Gao, M. Chen, and C. Xu, "Collecting cross-modal presence-absence evidence for weakly-supervised audio-visual event perception," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 18 827–18 836. 9

[212] S. Mo and Y. Tian, "Multi-modal grouping network for weakly-supervised audio-visual video parsing," Advances in Neural Information Processing Systems, vol. 35, pp. 34 722–34 733, 2022. 9

[213] D. S. S., A. Gupta, C. V. Jawahar, and M. Tapaswi, "Unsupervised audio-visual lecture segmentation," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), January 2023, pp. 5232–5241. 10

[214] H. Yun, Y. Yu, W. Yang, K. Lee, and G. Kim, "Pano-avqa: Grounded audio-visual question answering on 360deg videos," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2031–2041. 10

[215] H. M. Fayek and J. Johnson, "Temporal reasoning via audio question answering," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2283–2294, 2020. 10

[216] C. Hori, H. Alamri, J. Wang, G. Wichern, T. Hori, A. Cherian, T. K. Marks, V. Cartillier, R. G. Lopes, A. Das, I. Essa, D. Batra, and D. Parikh, "End-to-end audio visual scene-aware dialog using multimodal attention-based video features," in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2352–2356. 10

[217] M. Lao, N. Pu, Y. Liu, K. He, E. M. Bakker, and M. S. Lew, "Coca: Collaborative causal regularization for audio-visual question answering," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 11, 2023, pp. 12 995–13 003. 10

[218] Z. Zeng, J. Tu, M. Liu, T. S. Huang, B. Pianfetti, D. Roth, and S. Levinson, "Audio-visual affect recognition," IEEE Transactions on Multimedia, vol. 9, no. 2, pp. 424–428, 2007. 10

[219] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging recent advances in deep learning for audio-visual emotion recognition," Pattern Recognition Letters, vol. 146, pp. 1–7, 2021. 10, 19

[220] J. Han, Z. Zhang, N. Cummins, F. Ringeval, and B. Schuller, "Strength modelling for real-worldautomatic continuous affect recognition from audiovisual signals," Image and Vision Computing, vol. 65, pp. 76–86, 2017. 10

[221] Z. Zeng, J. Tu, B. Pianfetti, M. Liu, T. Zhang, Z. Zhang, T. Huang, and S. Levinson, "Audio-visual affect recognition through multi-stream fused hmm for hci," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 2, 2005, pp. 967–972 vol. 2. 10, 19, 20

[222] H. Chen, Y. Deng, S. Cheng, Y. Wang, D. Jiang, and H. Sahli, "Efficient spatial temporal convolutional features for audiovisual continuous affect recognition," in Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, 2019, pp. 19–26. 10

[223] Z. Zeng, Z. Zhang, B. Pianfetti, J. Tu, and T. Huang, "Audio-visual affect recognition in activation-evaluation space," in 2005 IEEE International Conference on Multimedia and Expo, 2005, pp. 4 pp.–. 10, 19, 20

[224] S. Chen, H. Li, Q. Wang, Z. Zhao, M. Sun, X. Zhu, and J. Liu, "Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset," Advances in Neural Information Processing Systems, vol. 36, 2024. 11

[225] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending clip to image, text and audio," in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 976–980. 11

[226] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth et al., "Gemini: a family of highly capable multimodal models," arXiv preprint arXiv:2312.11805, 2023. 11

[227] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh, "Video generation models as world simulators," 2024. [Online]. Available: https://openai.com/research/video-generation-models-as-world-simulators 11

[228] J. S. Garofolo, M. Michel, V. M. Stanford, and E. Tabassi, "The nist meeting room pilot corpus." 2004. [Online]. Available: https://doi.org/10.35111/800p-fv08 11, 13

[229] T. Geng, T. Wang, J. Duan, R. Cong, and F. Zheng, "Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 22 942–22 951. 11

[230] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, A. Gebreselasie, C. González, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolář, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. Ruiz, M. Ramazanova, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Z. Zhao, Y. Zhu, P. Arbeláez, D. Crandall, D. Damen, G. M. Farinella, C. Fuegen, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. Newcombe, A. Oliva, H. S. Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, L. Torresani, M. Yan, and J. Malik, "Ego4d: Around the world in 3,000 hours of egocentric video," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022, pp. 18 995–19 012. 11

[231] X. Gong, S. Mohan, N. Dhingra, J.-C. Bazin, Y. Li, Z. Wang, and R. Ranjan, "Mmg-ego4d: Multimodal generalization in egocentric action recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 6481–6491. 11

[232] S. Clarke, R. Gao, M. Wang, M. Rau, J. Xu, J.-H. Wang, D. L. James, and J. Wu, "Realimpact: A dataset of impact sound fields for real objects," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 1516–1525. 11

[233] K. J. Piczak, "Esc: Dataset for environmental sound classification," in Proceedings of the 23rd ACM international conference on Multimedia, 2015, pp. 1015–1018. 12, 18, 22, 23

[234] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," IEEE Transactions on Multimedia, vol. 17, no. 10, pp. 1733–1746, 2015. 12, 22, 23

[235] J. Huh, J. Chalk, E. Kazakos, D. Damen, and A. Zisserman, "EPIC-SOUNDS: A Large-Scale Dataset of Actions that Sound," in IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP), 2023. 12

[236] D. Damen, H. Doughty, G. M. Farinella, , A. Furnari, J. Ma, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100," International Journal of Computer Vision (IJCV), vol. 130, p. 33–55, 2022. [Online]. Available: https://doi.org/10.1007/s11263-021-01531-2 12

[237] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626. 15

[238] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold et al., "Cnn architectures for large-scale audio classification," in 2017 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, 2017, pp. 131–135. 15

[239] M. Naphade, T. Kristjansson, B. Frey, and T. Huang, "Probabilistic multimedia objects (multijects): a novel approach to video indexing and retrieval in multimedia systems," in Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269), 1998, pp. 536–540 vol.3. 16

[240] F. D. M. de Souza, S. Sarkar, and G. Cámara-Chávez, "Building semantic understanding beyond deep learning from sound and vision," in 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 2097–2102. 16

[241] D. Kiela, E. Grave, A. Joulin, and T. Mikolov, "Efficient large-scale multi-modal classification," CoRR, vol. abs/1802.02892, 2018. [Online]. Available: http://arxiv.org/abs/1802.02892 16

[242] Q. Fan, C. Chen, H. Kuehne, M. Pistoia, and D. D. Cox, "More is less: Learning efficient video representations by big-little network and depthwise temporal aggregation," CoRR, vol. abs/1912.00869, 2019. [Online]. Available: http://arxiv.org/abs/1912.00869 16

[243] Z. Wu, C. Xiong, C.-Y. Ma, R. Socher, and L. S. Davis, "Adaframe: Adaptive frame selection for fast video recognition," 2019. 16

[244] N. Hussein, M. Jain, and B. E. Bejnordi, "Timegate: Conditional gating of segments in long-range activities," 2020. 16

[245] B. Korbar, D. Tran, and L. Torresani, "Scsampler: Sampling salient clips from video for efficient action recognition," 2019. 16

[246] Y. Meng, C.-C. Lin, R. Panda, P. Sattigeri, L. Karlinsky, A. Oliva, K. Saenko, and R. Feris, "Ar-net: Adaptive frame resolution for efficient action recognition," 2020. 16

[247] W. Wu, D. He, X. Tan, S. Chen, and S. Wen, "Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition," 2019. 16

[248] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei, "End-to-end learning of action detection from frame glimpses in videos," 2017. 16

[249] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in European conference on computer vision. Springer, 2016, pp. 20–36. 16

[250] B. Vanderplaetse and S. Dupont, "Improved soccer action spotting using both audio and video streams," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, pp. 3921–3931. 17

[251] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in ICML, 2011. 17

[252] M. Slaney and M. Covell, "Facesync: A linear operator for measuring synchronization of video facial images and audio tracks," in Advances in Neural Information Processing Systems, T. Leen, T. Dietterich, and V. Tresp, Eds., vol. 13. MIT Press, 2001. [Online]. Available: https://proceedings.neurips.cc/paper/2000/file/9f6992966d4c363ea0162a056cb45fe5-Paper.pdf 17

[253] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," IEEE transactions on Multimedia, vol. 9, no. 7, pp. 1396–1403, 2007. 17

[254] N. Srivastava, R. Salakhutdinov et al., "Multimodal learning with deep boltzmann machines." in NIPS, vol. 1. Citeseer, 2012, p. 2. 17

[255] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," Information Fusion, vol. 49, pp. 69–78, 2019. 17, 19

[256] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: a new learning scheme of feedforward neural networks," in 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541), vol. 2, 2004, pp. 985–990 vol.2. 17

[257] I. Misra and L. van der Maaten, "Self-supervised learning of pretext-invariant representations," 2019. 18

[258] J. H. McDermott and E. P. Simoncelli, "Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis," Neuron, vol. 71, no. 5, pp. 926–940, 2011. 18

[259] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll, "Bimodal fusion of emotional data in an automotive environment," in Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., vol. 2, 2005, pp. ii/1085–ii/1088 Vol. 2. 19, 20

[260] Z. Zeng, J. Tu, M. Liu, T. Zhang, N. Rizzolo, Z. Zhang, D. Roth, and S. Levinson, "Bimodal hci-related affect recognition," 01 2004, pp. 137–143. 19

[261] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Survey on audiovisual emotion recognition: databases, features, and data fusion strategies," APSIPA transactions on signal and information processing, vol. 3, 2014. 20

[262] L. Ruan, Y. Ma, H. Yang, H. He, B. Liu, J. Fu, N. J. Yuan, Q. Jin, and B. Guo, "Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 10 219–10 228. 20

[263] K. Su, K. Qian, E. Shlizerman, A. Torralba, and C. Gan, "Physics-driven diffusion models for impact sound synthesis from videos," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 9749–9759. 20

[264] Y. Du, Z. Chen, J. Salamon, B. Russell, and A. Owens, "Conditional generation of audio from video via foley analogies," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 2426–2436. 20

[265] K. Sung-Bin, A. Senocak, H. Ha, A. Owens, and T.-H. Oh, "Sound to visual scene generation by audio-to-visual latent alignment," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 6430–6440. 20

[266] R. Huang, P. Lai, Y. Qin, and G. Li, "Parametric implicit face representation for audio-driven facial reenactment," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 12 759–12 768. 20

[267] A. Chatziagapi and D. Samaras, "Avface: Towards detailed audio-visual 4d face reconstruction," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 16 878–16 889. 20

[268] M. Agarwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "Audio-visual face reenactment," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 5178–5187. 20

[269] L. Zhu, X. Liu, X. Liu, R. Qian, Z. Liu, and L. Yu, "Taming diffusion models for audio-driven co-speech gesture generation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 10 544–10 553. 20

[270] S. Shen, W. Zhao, Z. Meng, W. Li, Z. Zhu, J. Zhou, and J. Lu, "Difftalk: Crafting diffusion models for generalized audio-driven portraits animation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 1982–1991. 20

[271] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang, "Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 8652–8661. 20

[272] A. Rouditchenko, H. Zhao, C. Gan, J. McDermott, and A. Torralba, "Self-supervised audio-visual co-segmentation," in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2357–2361. 20

[273] K. Li, Z. Yang, L. Chen, Y. Yang, and J. Xiao, "Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation," in Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 1485–1494. 20

[274] Y. Mao, J. Zhang, M. Xiang, Y. Zhong, and Y. Dai, "Multimodal variational auto-encoder based audio-visual segmentation," in Proceedings

of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2023, pp. 954–965. 20
[275] O. Gillet, S. Essid, and G. Richard, "On the correlation of automatic audio and visual segmentations of music videos," IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, no. 3, pp. 347–355, 2007. 20
[276] G. Sargent, P. Hanna, and H. Nicolas, "Segmentation of music video streams in music pieces through audio-visual analysis," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 724–728. 20
[277] J. Nam and A. Tewfik, "Combined audio and visual streams analysis for video sequence segmentation," in 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 4, 1997, pp. 2665–2668 vol.4. 20
[278] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso, "Temporal video segmentation to scenes using high-level audiovisual features," IEEE Transactions on Circuits and Systems for Video Technology, vol. 21, no. 8, pp. 1163–1177, 2011. 20
[279] ——, "On the use of audio events for improving video scene segmentation," Analysis, Retrieval and Delivery of Multimedia Content, pp. 3–19, 2013. 20
[280] H. Sundaram and S.-F. Chang, "Video scene segmentation using video and audio features," in 2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No.00TH8532), vol. 2, 2000, pp. 1145–1148 vol.2. 20
[281] Y. Tahir, D. Chakraborty, T. Maszczyk, S. Dauwels, J. Dauwels, N. Thalmann, and D. Thalmann, "Real-time sociometrics from audio-visual features for two-person dialogs," in 2015 IEEE International Conference on Digital Signal Processing (DSP), 2015, pp. 823–827. 20
[282] A. Adeel, J. Ahmad, H. Larijani, and A. Hussain, "A novel real-time, lightweight chaotic-encryption scheme for next-generation audio-visual hearing aids," Cognitive Computation, vol. 12, pp. 589–601, 2020. 20
[283] M. S. Hossain and G. Muhammad, "An audio-visual emotion recognition system using deep learning fusion for a cognitive wireless framework," IEEE Wireless Communications, vol. 26, no. 3, pp. 62–68, 2019. 20
[284] P. Motlicek, S. Duffner, D. Korchagin, H. Bourlard, C. Scheffler, J.-M. Odobez, G. Del Galdo, M. Kallinger, and O. Thiergart, "Real-time audio-visual analysis for multiperson videoconferencing," Advances in Multimedia, vol. 2013, 01 2013. 20
[285] H. Huang, L. Hu, F. Xiao, A. Du, N. Ye, and F. He, "An eeg-based identity authentication system with audiovisual paradigm in iot," Sensors, vol. 19, no. 7, p. 1664, 2019. 20
[286] H. Xu, Z. Cai, D. Takabi, and W. Li, "Audio-visual autoencoding for privacy-preserving video streaming," IEEE Internet of Things Journal, vol. 9, no. 3, pp. 1749–1761, 2021. 20
[287] Y. Zhao, P. Barnaghi, and H. Haddadi, "Multimodal federated learning on iot data," in 2022 IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation (IoTDI), 2022, pp. 43–54. 20
[288] A. Saxena, K. Shinghal, R. Misra, and A. Agarwal, "Automated enhanced learning system using iot," in 2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU), 2019, pp. 1–5. 20
[289] K. Mikolajczyk, O. Bown, and S. Ferguson, "A study of creative development with an iot-based audiovisual system: Creative strategies and impacts for system design," in Proceedings of the 15th Conference on Creativity and Cognition, 2023, pp. 139–149. 20
[290] Y. He, K. P. Seng, and L. M. Ang, "Generative adversarial networks (gans) for audio-visual speech recognition in artificial intelligence iot," Information, vol. 14, no. 10, p. 575, 2023. 20
[291] Y. Wang, W. Gao, S. Yang, Q. Chen, C. Ye, H. Wang, Q. Zhang, J. Ren, Z. Ning, X. Chen et al., "Humanoid intelligent display platform for audiovisual interaction and sound identification," Nano-Micro Letters, vol. 15, no. 1, p. 221, 2023. 20
[292] G. Peruzzi, A. Pozzebon, and M. Van Der Meer, "Fight fire with fire: Detecting forest fires with embedded machine learning models dealing with audio and images on low power iot devices," Sensors, vol. 23, no. 2, p. 783, 2023. 20
[293] S. Ghose and J. J. Prevost, "Enabling an iot system of systems through auto sound synthesis in silent video with dnn," in 2020 IEEE 15th International Conference of System of Systems Engineering (SoSE), 2020, pp. 563–568. 20
[294] K. Doshi and Y. Yilmaz, "Rethinking video anomaly detection - a continual learning approach," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), January 2022, pp. 3961–3970. 24
[295] S. Mo, W. Pian, and Y. Tian, "Class-incremental grouping network for continual audio-visual learning," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2023, pp. 7788–7798. 24
[296] W. Pian, S. Mo, Y. Guo, and Y. Tian, "Audio-visual class-incremental learning," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2023, pp. 7799–7811. 24
[297] D. Hu, Z. Wang, H. Xiong, D. Wang, F. Nie, and D. Dou, "Curriculum audiovisual learning," arXiv preprint arXiv:2001.09414, 2020. 24

Acronyms

AVC Audio-Visual Correspondence. 4, 5, 24, 25
AVID Audio-Visual Instance Discrimination. 6
AVS Audio-visual Synchronization. 7, 24, 25
AVSL Audio-Visual Speaker Localization. 5
AVSS Audio-Visual Source Separation. 5–7, 11, 12, 24

BI Bayesian Inference. 4
BIC Bayesian Information Criterion. 4

CAM Class Activation Mapping. 5
CMA Cross-modal Agreement. 6, 18
CMD Cross-modal Discrimination. 6

EM Expectation-Maximization Algorithm. 4, 16

FOA First Order Ambisonics. 6

GMM Gaussian Mixutre Model. 4
GSC Graph Spectral Cluster. 4

HMM Hidden Markov Model. 4
HRIR Head-Related Impulse Response. 5

LGBP Local Gabor Binary Patterns from. 19
LPCC Linear Prediction Cepstral Coefficients. 4

MFB Multi-modal Factorized Bilinear pooling. 5
MFCC Mel-Frequency Cepstral Coefficients. 4, 5, 13, 19
MIML Multi-instance multi-label learning. 6

NMF Non-negative Matrix Factorization. 6

RBM Restricted Boltzmann Machine. 7

SAR Signal to Artifacts Ratio. 21
SDR Signal to Distortion Ratio. 21, 23
SG Sound Generator. 5
SHD Spherical Harmonic Decomposition. 5
SI-SDR scale-invariant source-to-distortion ratio. 5
SIR Signal to Interference Ratio. 21
SSL Sound Source Localization. 4–7, 11, 12, 18, 19, 24, 25
SVM Support Vector Machine. 4

TCN Temporal Convolutional Network. 7
TDOA Time Delay of Arrival. 4

ZOA Zeroth Order Ambisonics. 6

AHMED SHAHBAZ is a PhD student in the department of Computer Science at University of South Florida. He works under the supervision od Dr. Sudeep Sarkar in the USF Institute for Artificial Intelligence + X. His research area interest is Computer Vision and Machine Learning.

SUDEEP SARKAR is a Distinguished University Professor, Chair of Computer Science and Engineering at the University of South Florida, Tampa, and Co-Director of the USF Institute for Artificial Intelligence + X. He received his M.S. and Ph.D. in electrical engineering on a University Presidential Fellowship from The Ohio State University, Columbus, and his B. Tech degree from the Indian Institute of Technology, Kanpur. He has 35 years of experience conducting and directing fundamental research in computer vision, predictive learning, biometrics, and artificial intelligence. His use-inspired contributions are in systems to recognize persons from how they walk (gait biometrics), automated recognition of actions, activities, and events in a video, economic activity from satellite images, and extracting precise medically relevant information from medical images. He has directed 22 Doctoral and 25 Master's students on these topics. He is a co-Editor-in-Chief of Pattern Recognition Letters and was the President of the IEEE Biometrics Council. He is a Fellow of the National Academy of Inventors (NAI), American Association for the Advancement of Science (AAAS), American Institute of Medical and Biological Engineers (AIMBE), Institute of Electrical and Electronics Engineers (IEEE), and International Association for Pattern Recognition (IAPR) and member of the Academy of Science, Engineering, and Medicine of Florida. He was the recipient of the National Science Foundation CAREER award in 1994, the USF Teaching Incentive Program Award for Undergraduate Teaching Excellence in 1997, the Outstanding Undergraduate Teaching Award in 1998, the Theodore and Venette Askounes-Ashford Distinguished Scholar Award in 2004, and William R. Jones Outstanding Mentor Award, 2017.