International Journal of Automation and Computing
www.ijac.net

18(3), June 2021, 351-376
DOI: 10.1007/s11633-021-1293-0

Deep Audio-visual Learning: A Survey

Hao Zhu[1,2] Man-Di Luo[2,3] Rui Wang[1,2] Ai-Hua Zheng[1] Ran He[2,3,4]

[1] Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei 230601, China

[2] Center for Research on Intelligent Perception and Computing (CRIPAC) and National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

[3] School of Artificial Intelligence, University of the Chinese Academy of Sciences, Beijing 100049, China

[4] Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai 200031, China

Abstract: Audio-visual learning, aimed at exploiting the relationship between audio and visual modalities, has drawn considerable attention since deep learning started to be used successfully. Researchers tend to leverage these two modalities to improve the performance of previously considered single-modality tasks or address new challenging problems. In this paper, we provide a comprehensive survey of recent audio-visual learning development. We divide the current audio-visual learning tasks into four different subfields: audio-visual separation and localization, audio-visual correspondence learning, audio-visual generation, and audio-visual representation learning. State-of-the-art methods, as well as the remaining challenges of each subfield, are further discussed. Finally, we summarize the commonly used datasets and challenges.

Keywords: Deep audio-visual learning, audio-visual separation and localization, correspondence learning, generative models, representation learning.

Citation: H. Zhu, M. D. Luo, R. Wang, A. H. Zheng, R. He. Deep audio-visual learning: A survey. *International Journal of Automation and Computing*, vol.18, no.3, pp.351–376, 2021. http://doi.org/10.1007/s11633-021-1293-0

1 Introduction

Human perception is multidimensional and includes vision, hearing, touch, taste, and smell. In recent years, along with the vigorous development of artificial intelligence technology, the trend from single-modality learning to multimodality learning has become crucial to better machine perception. Analyses of audio and visual information, representing the two most important perceptual modalities in our daily life, have been widely developed in both academia and industry in the past decades. Prominent achievements include speech recognition[1, 2], facial recognition[3–5], fine-grained visual classification[6–8], etc. Audio-visual learning (AVL) using both modalities has been introduced to overcome the limitation of perception tasks in each modality. In addition, exploring the relationship between audio and visual information leads to more interesting and important research topics and ultimately better perspectives on machine learning.

The purpose of this article is to provide an overview of the key methodologies in audio-visual learning, which aims to discover the relationship between audio and visual data for many challenging tasks. In this paper, we mainly divide these efforts into four categories: 1) audio-visual separation and localization, 2) audio-visual corresponding learning, 3) audio and visual generation, and 4) audio-visual representation.

Audio-visual separation and localization aim to separate specific sounds emanating from the corresponding objects and localize each sound in the visual context, as illustrated in Fig. 1(a). Audio separation has been investigated extensively in the signal processing community during the past two decades. With the addition of the visual modality, audio separation can be transformed into audio-visual separation, which has proven to be more effective in noisy scenes[9–11]. Furthermore, introducing the visual modality allows for audio localization, i.e., the localization of a sound in the visual modality according to the audio input. The tasks of audio-visual separation and localization themselves not only lead to valuable applications but also provide the foundation for other audio-visual tasks, e.g., generating spatial audio for 360° video[12]. Most studies in this area focus on unsupervised learning due to the lack of training labels.

Audio-visual correspondence learning focuses on

Review
Manuscript received December 4, 2020; accepted March 15, 2021; published online April 15, 2021
Recommended by Associate Editor Nazim Mir-Nasiri
Colored figures are available in the online version at https://link.springer.com/journal/11633

discovering the global semantic relation between audio and visual modalities, as shown in Fig. 1(b). It consists of audio-visual retrieval and audio-visual speech recognition tasks. The former uses audio or an image to search for its counterpart in another modality, while the latter derives from the conventional speech recognition task that leverages visual information to provide a more semantic prior to improve recognition performance. Although both of these two tasks have been extensively studied, they still entail major challenges, especially for fine-grained cross-modality retrieval and homonyms in speech recognition.

Audio-visual generation tries to synthesize the other modality based on one of them, which is different from the above two tasks leveraging both audio and visual modalities as inputs. Trying to make a machine that is creative is always challenging, and many generative models have been proposed[13, 14]. Audio-visual cross-modality generation has recently drawn considerable attention. It aims to generate audio from visual signals, or vice versa. Although it is easy for a human to perceive the natural correlation between sounds and appearance, this task is challenging for machines due to heterogeneity across modalities. As shown in Fig. 1(c), vision to audio generation mainly focuses on recovering speech from lip sequences or predicting the sounds that may occur in the given scenes. In contrast, audio to vision generation can be classified into three categories: audio-driven image generation, body motion generation, and talking face generation.

Audio-visual representation learning aims to automatically discover the representation from raw data. A human can easily recognize audio or video based on long-term brain cognition. However, machine learning algorithms such as deep learning models are heavily dependent on data representation. Therefore, learning suitable data representations for machine learning algorithms may improve performance.

Unfortunately, real-world data such as images, videos, and audio do not possess specific algorithmically defined features[15]. Therefore, an effective representation of data determines the success of machine learning algorithms. Recent studies seeking better representation have designed various tasks, such as audio-visual correspondence (AVC)[16] and audio-visual temporal synchronization (AVTS)[17]. By leveraging such a learned representation, one can more easily solve audio-visual tasks mentioned in the very beginning.

In this paper, we present a comprehensive survey of the above four directions of audio-visual learning. The rest of this paper is organized as follows. We introduce the four directions in Sections 2–5. Section 6 summarizes the commonly used public audio-visual datasets. Finally, Section 8 concludes the paper.

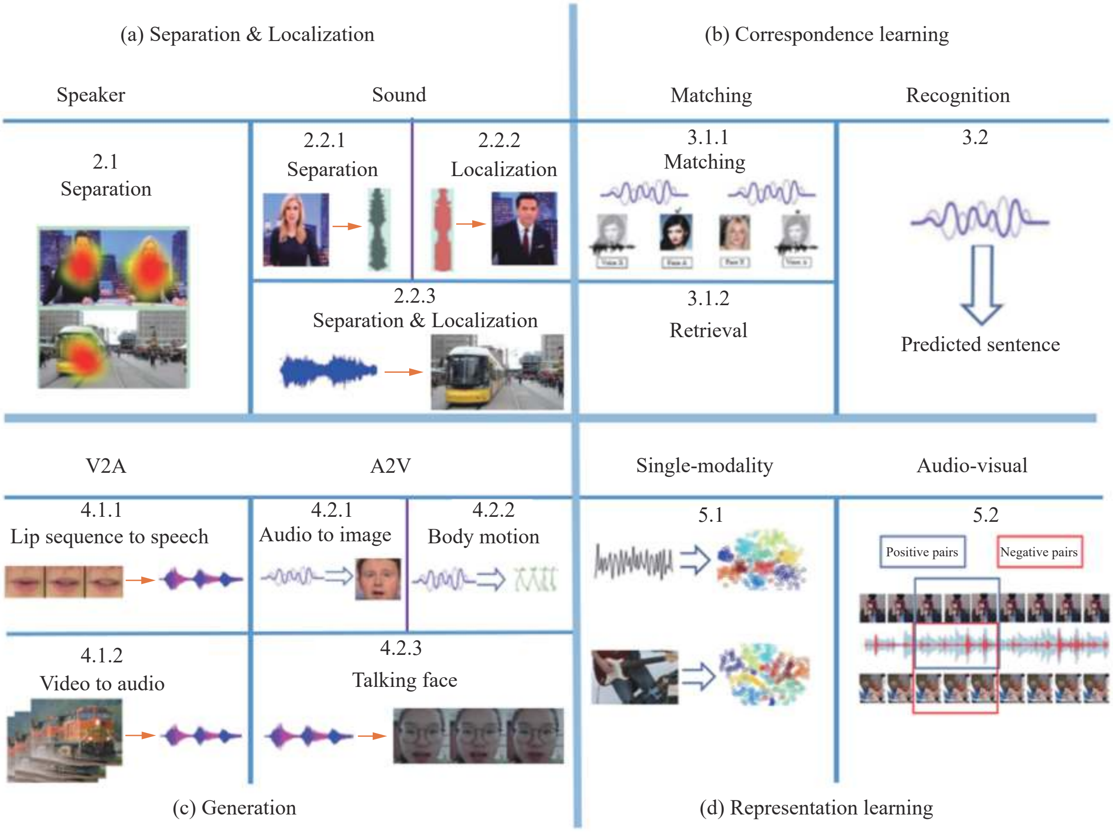


Fig. 1 Illustration of four categories of tasks in audio-visual learning

2 Audio-visual separation and localization

The objective of audio-visual separation is to separate different sounds from the corresponding objects, while audio-visual localization mainly focuses on localizing a sound in a visual context. As shown in Fig. 2, we classify types of this task by different identities: speakers (Fig. 2(a)) and objects (Fig. 2(b)). The former concentrates on a person's speech that can be used for television programs to enhance the target speakers' voice, while the latter is a more general and challenging task that separates arbitrary objects rather than speakers only. In this section, we provide an overview of these two tasks, examining the motivations, network architectures, advantages, and disadvantages as shown in Tables 1 and 2.

2.1 Speaker separation

The speaker separation task is a challenging task and is also known as the cocktail party problem. It aims to isolate a single speech signal in a noisy scene. Some studies tried to solve the problem of audio separation with only the audio modality and achieved exciting results[18, 19]. Advanced approaches[9, 11] tried to utilize visual information to aid the speaker separation task and significantly surpassed single modality-based methods. The early attempts leveraged mutual information to learn the joint distribution between the audio and the video[20, 21]. Subsequently, several methods focused on analyzing videos containing salient motion signals and the corresponding audio events (e.g., a mouth starting to move or a hand on piano suddenly accelerating)[22, 23].

Gabbay et al.[9] proposed isolating the voice of a specific speaker and eliminating other sounds in an audio-visual manner. Instead of directly extracting the target speaker's voice from the noisy sound, which may bias the training model, the researchers first fed the video frames into a video-to-speech model and then predicted the speaker's voice by the facial movements captured in the video. Afterwards, the predicted voice was used to filter the mixtures of sounds, as shown in Fig. 3. Although Gabbay et al.[9] improved the quality of separated voice by adding the visual modality, their approach was only applicable in controlled environments.

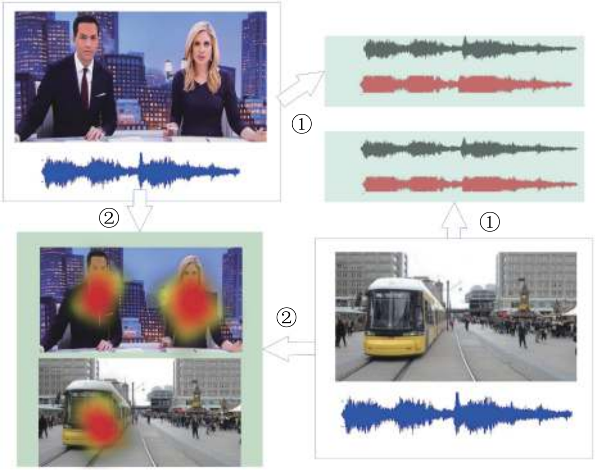


Fig. 2 Illustration of audio-visual separation and localization task. Paths 1 and 2 denote separation and localization tasks, respectively.

In contrast to previous approaches that require training a separate model for each speaker of interest (speaker-dependent models), recent researches focus on obtaining intelligible speech in an unconstrained environment. Afouras et al.[10] proposed a deep audio-visual speech enhancement network to separate the speaker's voice of the given lip region by predicting both the magnitude and phase of the target signal. They treated the spectrograms as temporal signals rather than images for a network. Additionally, instead of directly predicting clean signal magnitudes, they also tried to generate a more effective soft mask for filtering. Ephrat et al.[11] proposed a speaker-independent model that was only trained once and was then applicable to any speaker. This approach even outperformed the state-of-the-art speaker-dependent audio-visual speech separation methods. The relevant model consists of multiple visual streams and one audio stream, concatenating the features from different streams into a joint audio-visual representation. This feature is further processed by a bidirectional long short-term memory (LSTM)[24] and three fully connected layers. Finally, an elaborate spectrogram mask is learned for each speaker to be multiplied by the noisy input. Finally, the researchers converted it back to waveforms to obtain an isolated speech signal for each speaker. Lu et al.[25] designed a network similar to that of [11]. The difference is that Lu et al.[25] enforced an audio-visual matching network to distinguish the correspondence between speech and human lip movements. Therefore, they could obtain clear speech.

Instead of directly utilizing video as a condition, Morrone et al.[26] further introduced landmarks as a fine-grained feature to generate time-frequency masks to filter mixed-speech spectrogram.

2.2 Separating and localizing objects' sounds

Instead of matching a specific lip movement from a noisy environment, as in the speaker separation task, humans focus more on objects while dealing with sound separation and localization. It is difficult to find a clear correspondence between audio and visual modalities due to the challenge of exploring the prior sounds from different objects.

2.2.1 Separation

The early attempt to solve this localization problem can be traced back to 2000[27] and a study that synchronized low-level features of sounds and videos. Fisher et al.[21] later proposed using a nonparametric approach to

Table 1 Summary of recent audio-visual separation and localization approaches

Category	Method	Ideas & strengths	Weaknesses
Speaker separation	Gabbay et al.[9]	Predict speaker's voice based on faces in video used as a filter	Can only be used in controlled environments
	Afouras et al.[10]	Generate a soft mask for filtering in the wild	–
	Lu et al.[25]	Distinguish the correspondence between speech and human speech lip movements	Two speakers only; Hardly applied for background noise
	Ephrat et al.[11]	Predict a complex spectrogram mask for each speaker; Trained once, applicable to any speaker	The model is too complicated and lacks explanation
	Gu et al.[31]	All information of speakers; Robustness	Complex network; Plenty of preparation
	Zhu and Rahtu[32]	Strong capacity of sub-network; Single image	Small scope of application
	Morrone et al.[26]	Use landmarks to generate time-frequency masks	Additional landmark detection required
Separate and localize objects' sounds	Gao et al.[30]	Disentangle audio frequencies related to visual objects	Separated audio only
	Senocak et al.[34]	Focus on the primary area by using attention	Localized sound source only
	Tian et al.[37]	Joint modeling of auditory and visual modalities	Localized sound source only
	Pu et al.[23]	Use low rank to extract the sparsely correlated components	Not for the in-the-wild environment
	Zhao et al.[39]	Mix and separate a given audio; Without traditional supervision	Motion information is not considered
	Zhao et al.[40]	Introduce motion trajectory and curriculum learning	Only suitable for synchronized video and audio input
	Sharma et al.[38]	State-of-the-art for detection unconstrained videos entertainment media	Additional audio visual detection localize sound source only
	Sun et al.[43]	3D space; Low computational complexity	–
	Rouditchenko et al.[41]	Separation and localization use only one modality input	Does not fully utilize temporal information
	Parekh et al.[42]	Weakly supervised learning via multiple-instance learning	Only a bounding box proposed on the image

Table 2 A quantitative study on audio-visual separation and localization

Category	Method	Dataset	Result
Speaker separation	Gabbay et al.[9]	GRID[82] and TCD TIMIT[145]	SAR: 9.49 (on GRID)
	Afouras et al.[10]	LRS2[84] and VoxCeleb2[156]	–
	Lu et al.[25]	WSJ0 and GRID[82]	SAR: 10.11 (on GRID)
	Morrone et al.[26]	GRID[82] and TCD TIMIT[145]	PESQ: 2.45 (on TCD TIMIT)
Separate and localize objects' sounds	Gao et al.[30]	AudioSet[165]	SDR: 2.53
	Senocak et al.[34]	Base on filckr-SoundNet[138]	–
	Tian et al.[37]	Subset of AudioSet[165]	Prediction accuracy: 0.727
	Sharma et al.[38]	Movies	Recall: 0.512 9

learn a joint distribution of visual and audio signals and then project both of them to a learned subspace. Furthermore, several acoustic-based methods[28, 29] were described that required specific devices for surveillance and instrument engineering, such as microphone arrays used to capture the differences in the arrival of sounds.

To learn audio source separation from large-scale in-the-wild videos containing multiple audio sources per video, Gao et al.[30] suggested learning an audio-visual localization model from unlabeled videos and then exploiting the visual context for audio source separation. Researchers' approach relied on a multi-instance multilabel learning framework to disentangle the audio frequencies related to individual visual objects even without observing or hearing them in isolation. The multilabel learning framework was fed by a bag of audio basis vectors for

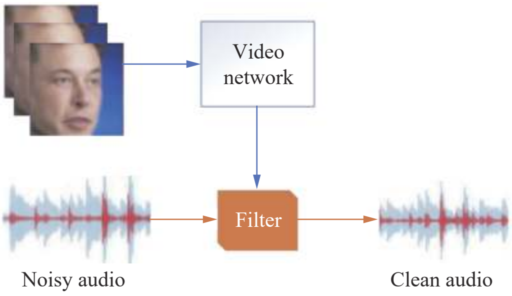


Fig. 3 Basic pipeline of a noisy audio filter

each video, and then, the bag-level prediction of the objects presented in the audio was obtained.

The information of the target speaker is helpful for sound separation tasks such as lip movement, tone, and spatial location etc. Therefore, Gu et al.[31] exploited all this information by obtaining semantic information of each modal via a fusion method based on factorized attention. Similarly, Zhu and Rahtu[32] attempted to add extra information (appearance) by introducing an appearance attention module to separate the different semantic representations.

2.2.2 Localization

Instead of only separating audio, can machines localize the sound source merely by observing sound and visual scene pairs as a human can? There is evidence both in physiology and psychology that sound localization of acoustic signals is strongly influenced by the synchronicity of their visual signals[27]. The past efforts in this domain were limited to requiring specific devices or additional features. Izadinia et al.[33] proposed utilizing the velocity and acceleration of moving objects as visual features to assign sounds to them. Zunino et al.[29] presented a new hybrid device for sound and optical imaging that was primarily suitable for automatic monitoring.

As the number of unlabeled videos on the Internet has been increasing dramatically, recent methods mainly focus on unsupervised learning. Additionally, modeling audio and visual modalities simultaneously tend to outperform independent modeling. Senocak et al.[34] learned to localize sound sources by merely watching and listening to videos. The relevant model mainly consisted of three networks, namely, sound and visual networks and an attention network trained via the distance ratio[35] unsupervised loss.

Attention mechanisms cause the model to focus on the primary area[36]. They provide prior knowledge in a semi-supervised setting. As a result, the network can be converted into a unified one that can learn better from data without additional annotations. To enable cross-modality localization, Tian et al.[37] proposed capturing the semantics of sound-emitting objects via the learned attention and leveraging temporal alignment to discover the correlations between the two modalities.

To make full use of media content in multiple modalities, Sharma et al.[38] proposed a novel network consisting of 3D convolution neural networks (3D CNNs) and bidirectional long short term memory (BiLSTMs) to fuse the complementary information of two modalities in a weaker-than-full supervision fashion.

2.2.3 Simultaneous separation and localization

Sound source separation and localization can be strongly associated with each other by assigning one modality′s information to another. Therefore, several researchers attempted to perform localization and separation simultaneously. Pu et al.[23] used a low-rank and sparse framework to model the background. The researchers extracted components with sparse correlations between the audio and visual modalities. However, the scenario of this method had a major limitation: It could only be applied to videos with a few sound-generating objects. Therefore, Zhao et al.[39] introduced a system called PixelPlayer that used a two-stream network and presented a mix-and-separate framework to train the entire network. In this framework, audio signals from two different videos were added to produce a mixed signal as input. The input was then fed into the network that was trained to separate the audio source signals based on the corresponding video frames. The two separated sound signals were treated as outputs. The system thus learned to separate individual sources without traditional supervision.

Instead of merely relying on image semantics while ignoring the temporal motion information in the video, Zhao et al.[40] subsequently proposed an end-to-end network called deep dense trajectory to learn the motion information for audio-visual sound separation. Furthermore, due to the lack of training samples, directly separating sound for a single class of instruments tend to lead to overfitting. Therefore, Zhao et al.[40] further proposed a curriculum strategy, starting by separating sounds from different instruments and proceeding to sounds from the same instrument. This gradual approach provided a good start for the network to converge better on the separation and localization tasks.

The methods of previous studies[23, 39, 40] could only be applied to videos with synchronized audio. Hence, Rouditchenko et al.[41] tried to perform localization and separation tasks using only video frames or sound by disentangling concepts learned by neural networks. The researchers proposed an approach to produce sparse activations that could correspond to semantic categories in the input using the sigmoid activation function during the training stage and softmax activation during the fine-tuning stage. Afterwards, the researchers assigned these semantic categories to intermediate network feature channels using labels available in the training dataset. In other words, given a video frame or a sound, the approach used the category-to-feature-channel correspondence to select a specific type of source or object for separation or localization. Aiming to introduce weak labels to improve performance, Parekh et al.[42] designed an approach based on multiple-instance learning, a well-known strategy for weakly supervised learning.

Inspired by the human auditory system, which accepts information selectively, Sun et al.[43] proposed a metamaterial-based single-microphone listening system (MSLS) to localize and separate the fixed sound signal in 3D space. The core part of the system is metamaterial enclosure which consisted of multiple second-order acoustic filters to decide the frequency response of different directions.

2.3 Discussions

Speaker voice separation has achieved great progress in various specific fields in the past decades, especially in audio-only modality. Introducing visual modality has increased both the performance and applications scenarios. Due to the explicit pattern between the voice and video, for example, the lip movement of the target speaker is highly related to the voice, recent efforts tend to leverage this pattern in sound separation tasks[31]. However, it hard to capture an explicit pattern between audio and visual in the more general tasks, such as object′s sound separation and localization. Therefore, researchers introduced effective strategies (such as sparse correlation, temporal motion information, multiple-instance learning, etc.) and more powerful networks for this task.

3 Audio-visual correspondence learning

In this section, we introduce several studies that explored the global semantic relation between audio and visual modalities. We name this branch of research audio-visual correspondence learning; it consists of 1) the audio-visual matching task and 2) the audio-visual speech recognition task. We summarize the advantages and disadvantages in Tables 3 and 4.

3.1 Audio-visual matching

Biometric authentication, ranging from facial recognition to fingerprint and iris authentication is a popular topic that has been researched over many years, while evidence shows that this system can be attacked maliciously. To detect such attacks, recent studies particularly focus on speech antispoofing measures.

Sriskandaraja et al.[44] proposed a network based on a Siamese architecture to evaluate the similarities between pairs of speech samples. Białobrzeski et al.[45] presented a two-stream network, where the first network was a Bayesian neural network assumed to be overfitting, and the second network was a CNN used to improve generalization. Gomez-Alanis et al.[46] further incorporated LightCNN[47] and a gated recurrent unit (GRU)[48] as a robust feature extractor to represent speech signals in utterance-level analysis to improve performance.

We note that cross-modality matching is a special form of such authentication that has recently been extensively studied. It attempts to learn the similarity between pairs. We divide this matching task into fine-grained voice-face matching and coarse-grained audio-image retrieval.

3.1.1 voice-face matching

Given facial images of different identities and the corresponding audio sequences, voice-face matching aims to identify the face that the audio belongs to (the V2F task) or vice versa (the F2V task), as shown in Fig. 4. The key point is finding the embedding between audio and visual modalities. Nagrani et al.[49] proposed using three networks to address the audio-visual matching problem: a static network, a dynamic network, and an N-way network. The static network and the dynamic network could only handle the problem with a specific number of images and audio tracks. The difference was that the dynamic network added to each image temporal information such as the optical flow or a 3D convolution[50, 51]. Based on the static network, Nagrani et al.[49] increased the number of samples to form an N-way network that was able to solve the $N:1$ identification problem.

However, the correlation between the two modalities was not fully utilized in the above method. Therefore, Wen et al.[52] proposed a disjoint mapping network (DIMNets) to fully use the covariates (e.g., gender and nationality)[53, 54] to bridge the relation between voice and face information. The intuitive assumption was that for a given voice and face pair, the more covariates were shared between the two modalities, the higher the probability of being a match. The main drawback of this framework was that a large number of covariates led to high data costs. Therefore, Hoover et al.[55] suggested a low-cost but robust approach of detection and clustering on audio clips and facial images. For the audio stream, the researchers applied a neural network model to detect speech for clustering and subsequently assigned a frame cluster to the given audio cluster according to the majority principle. Doing so required a small amount of data for pretraining.

To further enhance the robustness of the network, Chung et al.[56] proposed an improved two-stream training method that increased the number of negative samples to improve the error-tolerance rate of the network. The cross-modality matching task, which is essentially a classification task, allows for wide-ranging applications of the triplet loss. However, it is fragile in the case of multiple samples. To overcome this defect, Wang et al.[57] proposed a novel loss function to expand the triplet loss for multiple samples and a new elastic network (called EmNet) based on a two-stream architecture that can tolerate a variable number of inputs to increase the flexibility of the network. Most recently, Zheng et al.[58] proposed a novel adversarial-metric learning model that generates a modality-independent representation for each individual in each modality by adversarial learning while

Table 3 Summary of audio-visual correspondence learning

Category	Method	Ideas & strengths	Weaknesses
Voice-face matching	Nagrani et al.[49]	The method is novel and incorporates dynamic information	As the sample size increases, the accuracy decreases excessively
	Wen et al.[52]	The correlation between modes is utilized	Dataset acquisition is difficult
	Wang et al.[57]	Can deal with multiple samples; Can change the size of input	Static image only; Low robustness
	Hoover et al.[55]	Easy to implement; Robust and Efficient	Cannot handle large-scale data
	Zheng et al.[58]	Adversarial learning and metric learning is leveraged to explore the better feature representation	No high level semantic information is taken into account
Audio-visual retrieval	Hong et al.[64]	Preserve modality- specific characteristics; Soft intra-modality structure loss	Complex network
	Sanguineti et al.[67]	Acoustic images contain more information; Simple and efficient; Multimodal dataset	Three branches complex network; Lack of details in some places
	Takashima et al.[69]	Using CCA instead of distance	Unclear details

	Surís et al.[63]	Metric learning;
		Using fewer parameters;
			Static images

	Zeng et al.[66]	Consider mismatching pairs; Exploit negative examples	Complex network
	Chen et al.[68]	Deal with remote sensing data; Low memory and fast retrieval properties	Lack of remote sensing data

	Arsha et al.[65]	Curriculum learning;
		Applied value;
		Low data cost
			Low accuracy for multiple samples

Audio-visual speech recognition	Petridis et al.[77]	Simultaneously obtain feature and classification	Lack of audio information
	Wand et al.[78]	LSTM; Simple method	Word-level
	Chung et al.[84]	Audio and visual information; LRS dataset	The dataset is not guaranteed to be clean
	Zhang et al.[87]	Novel FBP; State-of-the-art	The experimental part is too simple

	Shillingford et al.[79]	Sentence-level;
		LipNet;
		CTC loss
			No audio information

	Zhou et al.[88]	Anti-noise; Simple and effective	Insufficient innovation
	Tao et al.[89]	Novel idea; Good performance	Insufficient contributions
	Makino et al.[81]	Large audio-visual dataset	Lead to low practical value
	Trigeorgis et al.[83]	Audio information; The algorithm is robust;	Noise is not considered
	Afouras et al.[86]	Study noise in audio; LRS2-BBC Dataset	Complex network

learning a robust similarity measure for cross-modality matching by metric learning.

3.1.2 Audio-image retrieval

The cross-modality retrieval task aims to discover the relationship between different modalities. Given one sample in the source modality, the proposed model can retrieve the corresponding sample with the same identity in the target modality. For audio-image retrieval as an example, the aim is to return a relevant piano sound, given a picture of a girl playing the piano. Compared with the previously considered voice and face matching, this task is more coarse-grained.

Unlike other retrieval tasks such as the text-image task[59−61] or the sound-text task[62], the audio-visual retrieval task mainly focuses on subspace learning. Surís et al.[63] proposed a new joint embedding model that mapped two modalities into a joint embedding space and then directly calculated the Euclidean distance between

Table 4 A quantitative study on correspondence learning

Category	Method	Dataset	Result
Voice-face matching	Nagrani et al.[49]	VGGFace[168] and Voxceleb[155]	V-F: 0.81
	Wen et al.[52]	VGGFace[168] and Voxceleb[155]	V-F: 0.84
	Hoover et al.[55]	LibriVox	Accuracy: 0.71
Audio-visual retrieval	Surís et al.[63]	Subset of YouTube-8M[166]	Audio-video recall: 0.631
	Nagrani et al.[65]	Voxceleb[155]	–
Audio-visual speech recognition	Chung et al.[84]	LRS[84]	–
	Trigeorgis et al.[83]	RECOLA	MSE: 0.684
	Afouras et al.[86]	LRS2-BBC	–

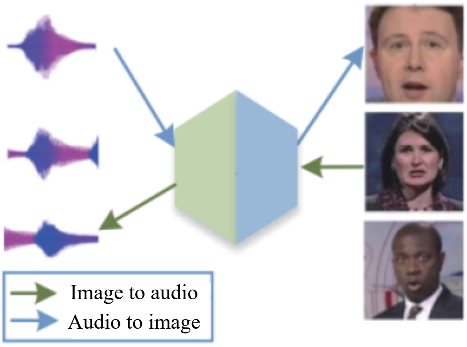


Fig. 4 Demonstration of audio-to-image retrieval (The blue arrows) and image-to-audio retrieval (The green arrows).

them. Surís et al. also leveraged cosine similarity to ensure that the two modalities in the same space were as close as possible while not overlapping. Note that the designed architecture would have a large number of parameters due to the existence of a large number of fully connected layers.

Hong et al.[64] proposed a joint embedding model that relied on pre-trained networks and used CNNs to replace fully connected layers to reduce the number of parameters to some extent. The video and music were fed to the pre-trained network and then aggregated, followed by a two-stream network trained via the inter-modal ranking loss. In addition, to preserve modality-specific characteristics, the researchers proposed a novel soft intra-modal structure loss. However, the resulting network was very complex and difficult to apply in practice. To solve this problem, Nagrani et al.[65] proposed a cross-modality self-supervised method to learn the embedding of audio and visual information from a video and significantly reduced the complexity of the network. For sample selection, Nagrani et al.[65] designed a novel curriculum learning schedule to further improve performance. In addition, the resulting joint embedding could be efficiently and effectively applied in practical applications.

Different from the above works only considering the matching pairs, Zeng et al.[66] further focused the mismatching pairs and proposed a novel deep triplet neural network with cluster-based canonical correlation analysis in a two-stream architecture. Rather than designing a model base on a two-stream structure, Sanguineti et al.[67] introduced an extra model named acoustic images, which contained abundant information. They aligned three modalities in time and space and took advantage of such correlation to learn more powerful audio-visual representations via knowledge distillation. Different from the approaches focused on face and audio, Chen et al.[68] proposed a deep image-voice retrieval (DIVR) to deal with remote sensing images. During the training process, they followed the idea of triplet loss. Moreover, they minimize the distance between hash-like codes and hash codes to reduce quantization error.

Music-emotion retrieval is an interesting topic in audio-image retrieval task. Takashima et al.[69] proposed a deep canonical correlation analysis (DeepCCA) by maximizing the correlation between two modalities in projection space via CCA rather than distance computing.

3.2 Audio-visual speech recognition

The recognition of the content of a given speech clip (for example, predicting the emotion based on the given speech[70]) has been studied for many years, yet despite great achievements, researchers are still aiming for satisfactory performance in challenging scenarios. Due to the correlation between audio and vision, combining these two modalities tends to offer more prior information. For example, one can predict the scene where the conversation took place, which provides a strong prior for speech recognition, as shown in Fig. 5.

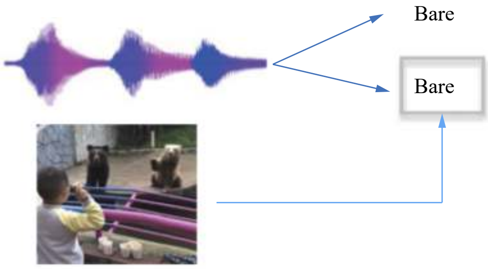


Fig. 5 Demonstration of audio-visual speech recognition

Earlier efforts on audio-visual fusion models usually consisted of two steps: 1) extracting features from the image and audio signals and 2) combining the features for joint classification[71–73]. Later, taking advantage of deep learning, feature extraction was replaced with a neural network encoder[74–76]. Several recent studies have shown a tendency to use an end-to-end approach to visual speech recognition. These studies can be mainly divided into two groups. They either leverage the fully connected layers and LSTM to extract features and model the temporal information[77, 78] or use a 3D convolutional layer followed by a combination of CNNs and LSTMs[79, 80]. However, LSTM is normally extensive. To this end, Makino et al.[81] proposed a large-scale system on the basis of a recurrent neural network transducer (RNN-T) architecture to evaluate the performance of RNN model.

Instead of a two-step strategy, Petridis et al.[77] introduced a new audio-visual model that is simultaneously extracting features directly from pixels and classifying speech, followed by a bidirectional LSTM module to fuse audio and visual information. To this end, Wand et al.[78] presented a word-level lip-reading system using LSTM. However, this work only conduct experiments on a lab-controlled dataset[82]. In contrast to previous methods, Assael et.al[79] proposed an end-to-end LipNet model based on sentence-level sequence prediction, which consisted of spatial-temporal convolutions, a recurrent network, and a model trained via the connectionist temporal classification (CTC) loss. Experiments showed that lip-reading outperformed the two-step strategy.

However, the limited information in the visual modality may lead to a performance bottleneck. To combine both audio and visual information for various scenes, especially in noisy conditions, Trigeorgis et al.[83] introduced an end-to-end model to obtain a context-aware feature from the raw temporal representation.

Chung et al.[84] presented a Watch, Listen, Attend, and Spell (WLAS) network to explain the influence of audio on the recognition task. The model took advantage of the dual attention mechanism and could operate on a single or combined modality. To speed up the training and avoid overfitting, the researchers also used a curriculum learning strategy. To analyze an in-the-wild dataset, Nussbaum-Thom et al.[85] proposed another model based on residual networks and a bidirectional GRU[48]. However, they did not take the ubiquitous noise in the audio into account. To solve this problem, Afouras et al.[86] proposed a model for performing speech recognition tasks. The researchers compared two common sequence prediction types: connectionist temporal classification and sequence-to-sequence (seq2seq) methods in their models. In the experiment, they observed that the model using seq2seq could perform better according to word error rate (WER) when it was only provided with silent videos. For pure-audio or audio-visual tasks, the two methods behaved similarly. In a noisy environment, the performance of the seq2seq model was worse than that of the corresponding CTC model, suggesting that the CTC model could better handle background noises.

Recent works introduced attention mechanisms to highlight some significant information contained in audio or visual representations. Zhang et al.[87] proposed a factorized bilinear pooling to learn the feature of respective modalities via an embedded attention mechanism, and then to integrate complex association between audio and video for the audio-video emotion recognition task. Zhou et al.[88] focused on the feature of respective modalities by multimodal attention mechanism to exploit the importance of both modalities to obtain a fused representation. Compared with the previous works which focused on the feature of each modal, Tao et al.[89] paid more attention to the network and proposed a cross-modal discriminative network called VFNet to establish the relationship between audio and face by cosine loss.

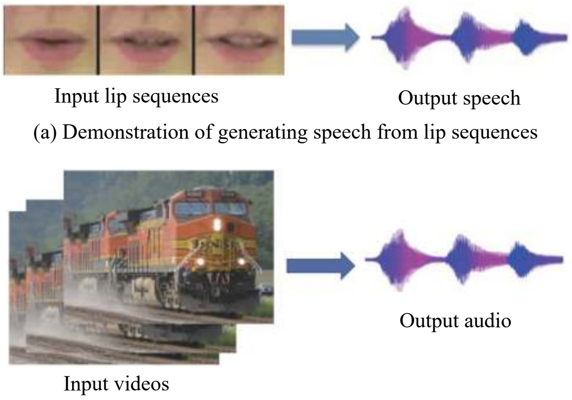
3.3 Discussions

The representation learning between modalities is crucial in audio-visual correspondence learning. One can add more supplementary information (e.g., mutual information, temporal information) or adjust the structure of the network such as the use of RNN and LSTM, increasing the modal structure or input pretreatment, etc., to obtain better representation.

4 Audio and visual generation

The previously introduced retrieval task shows that the trained model is able to find the most similar audio or visual counterpart. While humans can imagine the scenes corresponding to sounds and vice versa, researchers have tried to endow machines with this kind of imagination for many years. Following the invention and advances of generative adversarial networks (GANs[90], a generative model based on adversarial strategy), image or video generation has emerged as a topic. It involves several subtasks, including generating images or video from a potential space[91], cross-modality generation[92, 93], etc. These applications are also relevant to other tasks, e.g., domain adaptation[94, 95]. Due to the difference between audio and visual modalities, the potential correlation between them is nonetheless difficult for machines to discover. Generating sound from a visual signal or vice versa, therefore, becomes a challenging task.

In this section, we will mainly review the recent development of audio and visual generation, i.e., generating audio from visual signals or vice versa. Visual signals here mainly refer to images, motion dynamics, and videos. Section 4.1 mainly focuses on recovering the speech from the video of the lip area (Fig. 6(a)) or generating sounds that may occur in the given scenes (Fig. 6(a)). In contrast, Section 4.2 will examine generating images from a given audio (Fig. 7(a)), body motion generation (Fig. 7(b)), and talking face generation (Fig. 7(c)). The brief advantages and disadvantages are shown in Tables 5–7.



(a) Demonstration of generating speech from lip sequences

(b) Demonstration of video-to-audio generation

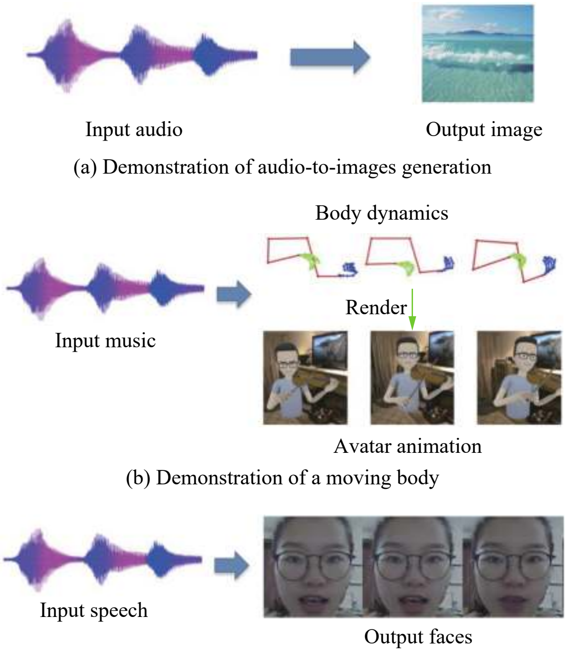
Fig. 6 Demonstration of visual-to-audio generation

4.1 Vision-to-audio generation

Many methods have been explored to extract audio information from visual information, including predicting sounds from visually observed vibrations and generating audio via a video signal. We divide the visual-to-audio generation tasks into two categories: generating speech from lip video and synthesizing sounds from general videos without scene limitations.

4.1.1 Lip sequence to speech

There is a natural relationship between speech and lips. Separately from understanding the speech content by observing lips (lip-reading), several studies have tried to reconstruct speech by observing lips. Le Cornu et al.[96] attempted to predict the spectral envelope from visual features, combining it with artificial excitation signals, and synthesizing audio signals in a speech production model. Ephrat and Peleg[97] proposed an end-to-end model based on a CNN to generate audio features for each silent video frame based on its adjacent frames. The waveform was therefore reconstructed based on the learned features to produce understandable speech.



(a) Demonstration of audio-to-images generation

(b) Demonstration of a moving body

(c) Demonstration of a talking face

Fig. 7 Demonstration of talking face generation and moving body generation.

Using temporal information to improve speech reconstruction has been extensively explored. Ephrat et al.[98] proposed leveraging the optical flow to capture the temporal motion at the same time. Le Cornu et al.[99] leveraged recurrent neural networks to incorporate temporal information into the prediction.

4.1.2 General video to audio

When a sound hits the surfaces of some small objects, the latter will vibrate slightly. Therefore, Davis et al.[100] utilized this specific feature to recover the sound from vibrations observed passively by a high-speed camera. Note

Table 5 Summary of recent approaches to video-to-audio generation

Category	Method	Ideas & strengths	Weaknesses
Lip sequence to speech	Le Cornu et al.[96]	Reconstruct intelligible speech only from visual speech features	Applied to limited scenarios
	Ephrat et al.[98]	Compute optical flow between frames	Applied to limited scenarios
	Le Cornu et al.[99]	Reconstruct speech using a classification approach combined with feature-level temporal information	Cannot apply to real-time conversational speech
General video to audio	Davis et al.[100]	Recover real-world audio by capturing vibrations of objects	Requires a specific device; Can only be applied to soft objects
	Owens et al.[101]	Use LSTM to capture the relation between material and motion	For a lab-controlled environment only
	Zhou et al.[102]	Leverage a hierarchical RNN to generate in-the-wild sounds	Monophonic audio only
	Morgado et al.[12]	Localize and separate sounds to generate spatial audio from 360° video	Expensive 360° videos are required
	Zhou et al.[104]	A unified model to generate stereophonic audio from mono data	

Table 6 A quantitative study on video-to-audio generation

Category	Method	Dataset	Result
Lip sequence to speech	Le Cornu et al.[96]	GRID[82]	–
	Ephrat et al.[98]	GRID[82] and TCD TIMIT[145]	PESQ: 1.922 (on GRID S4)
	Le Cornu et al.[99]	GRID[82]	–
General video to audio	Davis et al.[100]	Videos they collected	SSNR: 28.7
	Owens et al.[101]	Videos they collected	ERR: 0.21
	Zhou et al.[102]	VEGAS	Flow at category level: 0.603

Table 7 Summary of recent studies of audio-to-visual generation

Category	Method	Ideas & strengths	Weaknesses
Audio to image	Wan et al.[105]	Combined many existing techniques to form a GAN	Relative low quality
	Qiu and Kataoka[106]	Generated images related to music	Relative low quality
	Chen et al. [92]	Generated both audio-to-visual and visual-to-audio models	The models were independent
	Wen et al.[112]	Explore the relationship between two modalities	
	Hao et al. [107]	Proposed a cross-modality cyclic GAN	Generated images only
	Li et al.[108]	A teacher-student for speech-to-image generation	
	Wang et al. [109]	Relation information is leveraged	
Audio to motions	Alemi et al.[120]	Generated dance movements from music via real-time GrooveNet	Constrained to the given dataset
	Lee et al.[121]	Generated a choreography system via an autoregressive network	
	Shlizerman et al.[122]	Applied a target delay LSTM to predict body keypoints	
	Tang et al.[123]	Developed a music-oriented dance choreography synthesis method	
	Yalta et al.[124]	Produced weak labels from motion directions for motion-music alignment	
Talking face	Kumar et al.[125] and Supasorn et al.[127]	Generated keypoints by a time-delayed LSTM	Need retraining for different identities
	Jamaludin et al.[128]	Developed an encoder-decoder CNN model suitable for more identities	For a lab-controlled environment only
	Jalalifar et al.[129]	Combined RNN and GAN and applied keypoints	
	Vougioukas et al. [130]	Applied a temporal GAN for more temporal consistency	
	Chen et al.[132]	Applied optical flow	Generated lips only
	Eskimez et al.[137]	3D talking face landmarks; New training method	Mass of time to train the model
	Eskimez et al.[136]	Emotion discriminative loss	Heavy burden on the network and need lots of time
	Zhou et al.[133]	Disentangled information	Lacked realism
	Zhu et al.[93]	Asymmetric mutual information estimation to capture modality coherence	Suffered from the zoom-in-and-out condition
	Chen et al.[134]	Dynamic pixelwise loss	Required multistage training
	Wiles et al.[135]	Self-supervised model for multimodality driving	Relative low quality

that it should be easy for suitable objects to vibrate, which is the case for a glass of water, a pot of plants, or a box of napkins. We argue that this work is similar to the previously introduced speech reconstruction studies[96−99] since all of them use the relation between visual and sound context. In speech reconstruction, the visual part concentrates more on lip movement, while in this work, it focuses on small vibrations.

Owens et al.[101] observed that when different materials were hit or scratched; they emitted a variety of sounds. Thus, the researchers introduced a model that learned to synthesize sound from a video in which objects made of different materials were hit with a drumstick at different angles and velocities. The researchers demonstrated that their model could not only identify different sounds originating from different materials but also learn the pattern of interaction with objects (different actions applied to objects result in different sounds). The model leveraged an RNN to extract sound features from video frames and subsequently generated waveforms through an instance-based synthesis process.

Although Owens et al.[101] could generate sound from various materials, the approach they proposed still could not be applied to real-life applications since the network was trained by videos shot in a lab environment under strict constraints. To improve the result and generate sounds from in-the-wild videos, Zhou et al.[102] designed an end-to-end model. It was structured as a video encoder and a sound generator to learn the mapping from video frames to sounds. Afterwards, the network leveraged a hierarchical RNN[103] for sound generation. Specifically, the authors trained a model to directly predict raw audio signals (waveform samples) from input videos. They demonstrated that this model could learn the correlation between sound and visual input for various scenes and object interactions.

The previous efforts we have mentioned focused on monophonic audio generation, while Morgado et al.[12] attempted to convert monophonic audio recorded by a 360° video camera into spatial audio. Performing such a task of audio specialization requires addressing two primary issues: source separation and localization. Therefore, the researchers designed a model to separate the sound sources from mixed-input audio and then localize them in the video. Another multimodality model was used to guide the separation and localization since the audio and video were complementary. To generate stereophonic audio from mono data, Zhou et al.[104] proposed a sep-stereo framework that integrates stereo generation and source separation into a unified framework.

4.2 Audio to vision

In this section, we provide a detailed review of audio-to-visual generation. We first introduce audio-to-images generation, which is easier than video generation since it does not require temporal consistency between the generated images.

4.2.1 Audio to image

To generate images of better quality, Wan et al.[105] put forward a model that combined the spectral norm, an auxiliary classifier, and a projection discriminator to form the researchers′ conditional GAN model. The model could output images of different scales according to the volume of the sound, even for the same sound. Instead of generating real-world scenes of the sound that had occurred, Qiu and Kataoka[106] suggested imagining the content from music. They proposed a model features by feeding the music and images into two networks and learning the correlation between those features, and finally generating images from the learned correlation.

4.2.2 Speech to image generation

Several studies have focused on audio-visual mutual generation. Chen et al.[92] were the first to attempt to solve this cross-modality generation problem using conditional GANs. The researchers defined a sound-to-image (S2I) network and an image-to-sound (I2S) network that generated images and sounds, respectively. Instead of separating S2I and I2S generation, Hao et al.[107] combined the respective networks into one network by considering a cross-modality cyclic generative adversarial network (CMCGAN) for the cross-modality visual-audio mutual generation task. Following the principle of cyclic consistency, CMCGAN consisted of four subnetworks: audio-to-visual, visual-to-audio, audio-to-audio, and visual-to-visual.

Most recently, some studies tried to generate images conditioned on the speech description. Li et al.[108] proposed a speech encoding to learn the embedding features of speech, which is trained with a pre-trained image encoder using teacher-student learning strategy to obtain better generalization capability. Wang et al.[109] leveraged a speech embedding network to learn speech embeddings with the supervision of corresponding visual information from images. A relation-supervised densely-stacked generative model is then proposed to synthesize images conditioned on the learned embeddings. Furthermore, some studies have tried to reconstruct facial images from speech clips. Duarte et al.[110] synthesized facial images containing expressions and poses through the GAN model. Moreover, Duarte et al.[110] enhanced their model′s generation quality by searching for the optimal input audio length. To better learn normalized faces from speech, Oh et al.[111] explored a reconstructive model. The researchers trained an audio encoder by learning to align the feature space of speech with a pre-trained face encoder and decoder.

Different from the above methods, Wen et al.[112] proposed an unsupervised approach to reconstruct a face from audio. Specifically, they proposed a novel framework base on GANs, which reconstructed a face via an audio vector captured by the voice embedding and the generated face and identity are distinguished by discriminator and classifier, respectively.

4.2.3 Body motion generation

Instead of directly generating videos, numerous studies have tried to animate avatars using motions. The motion synthesis methods leveraged multiple techniques, such as dimensionality reduction[113, 114], hidden Markov models[115], Gaussian processes[116], and neural networks[117−119].

Alemi et al.[120] proposed a real-time GrooveNet based on conditional restricted Boltzmann machines and recurrent neural networks to generate dance movements from music. Lee et al.[121] utilized an autoregressive encoder-decoder network to generate a choreography system from music. Shlizerman et al.[122] further introduced a model that used a target delay LSTM to predict body landmarks. The latter was further used as agents to generate body dynamics. The key idea was to create an animation from the audio that was similar to the action of a pianist or a violinist. In summary, the entire process generated a video of artists′ performance corresponding to the input audio.

Although previous methods could generate body motion dynamics, the intrinsic beat information of the music has not been used. Tang et al.[123] proposed a music-oriented dance choreography synthesis method that extracted a relation between acoustic and motion features via an LSTM-autoencoder model. Moreover, to achieve better performance, the researchers improved their model with a masking method and temporal indexes. Providing weak supervision, Yalta et al.[124] explored producing weak labels from motion direction for motion-music alignment. The authors generated long dance sequences via a conditional autoconfigured deep RNN that was fed by an audio spectrum.

4.2.4 Talking face generation

Exploring audio-to-video generation, many researchers showed great interest in synthesizing people′s faces from speech or music. This has many applications, such as animating movies, teleconferencing, talking agents, and enhancing speech comprehension while preserving privacy. Earlier studies of talking face generation mainly synthesized a specific identity from the dataset based on audio of an arbitrary speech. Kumar et al.[125] attempted to generate key points synced to audio by utilizing a time-delayed LSTM[126] and then generated the video frames conditioned on the key points by another network. Furthermore, Supasorn et al.[127] proposed a teeth proxy to improve the visual quality of teeth during generation.

Subsequently, Jamaludin et al.[128] attempted to use an encoder-decoder CNN model to learn the correspondences between raw audio and videos. Combining recurrent neural network (RNN) and GAN[90], Jalalifar et al.[129] produced a sequence of realistic faces synchronized with the input audio by two networks. One was an LSTM network used to create lip landmarks out of audio input. The other was a conditional GAN (cGAN) used to generate the resulting faces conditioned on a given set of lip landmarks. Instead of applying cGAN,[130] proposed using a temporal GAN[131] to improve the synthesis quality. However, the above methods were only applicable to synthesizing talking faces with identities limited to those in a dataset.

The synthesis of talking faces of arbitrary identities has recently drawn significant attention. Chen et al.[132] considered correlations among speech and lip movements while generating multiple lip images. The researchers used the optical flow to better express the information between the frames. The fed optical flow represented not only the information of the current shape but also the previous temporal information.

A frontal face photo usually has both identity and speech information. Assuming this, Zhou et al.[133] used an adversarial learning method to disentangle different types of information of one image during generation. The disentangled representation had a convenient property that both audio and video could serve as the source of speech information for the generation process. As a result, it was possible to not only output the features but also express them more explicitly while applying the resulting network.

Most recently, to discover the high-level correlation between audio and video, Zhu et al.[93] proposed a mutual information approximation to approximate mutual information between modalities. Chen et al.[134] applied landmark and motion attention to generating talking faces and further proposed a dynamic pixel-wise loss for temporal consistency. Facial generation is not limited to specific audio or visual modalities since the crucial point is whether there is a mutual pattern between these different modalities. Wiles et al.[135] put forward a self-supervising framework called X2Face to learn the embedded features and generate target facial motions. It could produce videos from any input as long as the embedded features were learned.

Different from the above works, Eskimez et al.[136] proposed a supervised system (fed with a speech utterance, face image, emotion label and noise) to generate talking face and focused on emotion to improve the authenticity of the results. As intermediate information in talking face, generating landmarks from audio has attracted more attention in recent years. Eskimez et al.[137] proposed to generate 3D talking face landmarks from audio in a noisy environment. They exploited active shape model (ASM) coefficients of face landmarks to smooth video frames and introduced speech enhancement to cope with noise in the background.

4.3 Discussions

Audio-visual generation is an important yet challenging task among these fields. The challenge mainly derives from the big gap between audio and visual modalities. In order to narrow this gap, some scholars introduced extra information for their model, including landmarks, keypoints, mutual information, and optical flow, etc. More common approaches are changing network structure base on power GAN or other generative models such as cross-modality cycle generative adversarial network, GrooveNet, condition GAN, etc. Another effective approach is pre-processing the model's input, such as aligning the feature space, and animating avatars from motions.

5 Audio-visual representation learning

Representation learning aims to discover the pattern representation from data automatically. It is motivated by the fact that the choice of data representation usually greatly impacts the performance of machine learning[15]. However, real-world data such as images, videos, and audio are not amenable to defining specific features algorithmically. Additionally, the quality of data representation usually determines the success of machine learning algorithms. Bengio et al.[15] assumed the reason for this to be that different representations could better explain the laws underlying data, and the recent enthusiasm for AI has motivated the design of more powerful representation learning algorithms to achieve these priors.

In this section, we will review a series of audio-visual learning methods ranging from single-modality[138] to dual-modality representation learning[16, 17, 139−141]. The basic pipeline of such studies is shown in Fig. 8, and the strengths and weaknesses are shown in Tables 8 and 9.

5.1 Single-Modality representation learning

Naturally, to determine whether audio and video are related to each other, researchers focus on determining if they are from the same video or synchronized in the same video. Aytar et al.[138] exploited the natural synchronization between video and sound to learn an acoustic representation of a video. The researchers proposed a student-teacher training process that used an unlabeled video as a bridge to transfer discernment knowledge from a sophisticated visual identity model to the sound modality. Although the proposed approach managed to learn audio-modality representation in an unsupervised manner, discovering audio and video representations simultaneously remained to be solved.

5.2 Learning an audio-visual representation

The information concerning modality tends to be noisy in the corresponding audio and images, while we

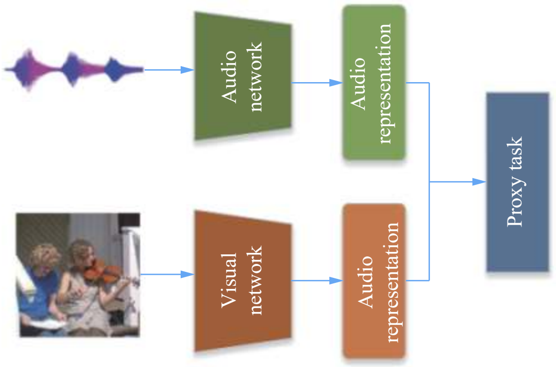


Fig. 8 Basic pipeline of representation learning

Table 8 Summary of recent audio-visual representation learning studies

Type	Method	Ideas & strengths	Weaknesses
Single modality	Aytar et al.[138]	Student-teacher training procedure with natural video synchronization	Only learned the audio representation
Dual modalities	Leidal et al.[140]	Regularized the amount of information encoded in the semantic embedding	Focused on spoken utterances and handwritten digits
	Arandjelovic et al.[16, 139]	Proposed the AVC task	Considered only audio and video correspondence
	Korbar et al.[17]	Proposed the AVTS task with curriculum learning	The sound source has to feature in the video
	Parekh et al.[144]	Use video labels for weakly supervised learning	Leverage the prior knowledge of event classification
	Hu et al.[141]	Disentangle each modality into a set of distinct components	Require a predefined number of clusters

Table 9 A quantitative study on audio-visual representation learning studies

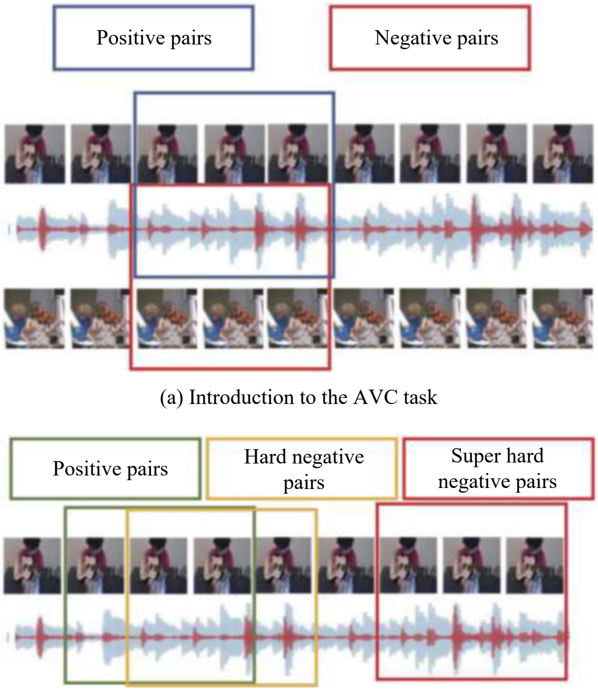
Type	Method	Dataset	Result
Single modality	Aytar et al.[138]	DCASE, ESC-50 and ESC-10	Classification accuracy: 0.88 (on DCASE)
Dual modalities	Leidal et al.[140]	TIDIGITs and MNIST	–
	Arandjelovic et al.[16]	Flickr-SoundNet[138] and Kinetics[161]	Accuracy: 0.74 (on Kinetics)
	Korbar et al.[17]	Kinetics[161] and AudioSet[165]	Accuracy: 0.78 (on Kinetics)
	Parekh et al.[144]	Subset of AudioSet[165]	Recall: 0.694

only require semantic content rather than the exact visual content. Leidal et al.[140] explored unsupervised learning of the semantic embedded space, which required a close distribution of the related audio and image. The researchers proposed a model to map an input to vectors of the mean. The logarithm of variance of a diagonal Gaussian distribution, and the sample semantic embeddings were drawn from these vectors.

To learn the audio and video′s semantic information by simply watching and listening to a large number of unlabeled videos, Arandjelovic et al.[16] introduced an audio-visual correspondence learning task (AVC) for training two (visual and audio) networks from scratch, as shown in Fig. 9(a). In this task, the corresponding audio and visual pairs (positive samples) were obtained from the same video, while mismatched (negative) pairs were extracted from different videos. To solve this task, Arandjelovic and Zisserman[16] proposed an L^3-Net that detected whether the semantics in visual and audio fields were consistent. Although this model was trained without additional supervision, it could learn representations of dual modalities effectively.

Exploring the proposed audio-visual coherence (AVC) task, Arandjelovic and Zisserman[139] continued to investigate AVE-Net that to find the most similar visual area to the current audio clip. Owens and Efros[142] proposed adopting a model similar to that of [16] but used a 3D convolution network for the videos instead, which could capture the motion information for sound localization.

In contrast to previous AVC task-based solutions, Korbar et al.[17] introduced another proxy task called audio-visual time synchronization (AVTS) that further considered whether a given audio sample and video clip were synchronized or not. In the previous AVC tasks, negative samples were obtained as audio and visual samples from different videos. However, exploring AVTS, the researchers trained the model using harder negative samples representing unsynchronized audio and visual segments sampled from the same video, forcing the model to learn the relevant temporal features. At this time, not only the semantic correspondence was enforced between the video and the audio, but more importantly, the synchronization between them was also achieved. The researchers applied the curriculum learning strategy[143] to this task and divided the samples into four categories: positives (the corresponding audio-video pairs), easy negatives (audio and video clips originating from different videos), difficult negatives (audio and video clips originating from the same video without overlap), and super-difficult negatives (audio and video clips that partly overlap), as shown in Fig. 9(b).



(a) Introduction to the AVC task

(b) Introduction to the AVTS task

Fig. 9 Introduction to the representation task

The above studies rely on two latent assumptions: 1) The sound source should be present in the video, and 2) only one sound source is expected. However, these assumptions limit the applications of the respective approaches to real-life videos. Therefore, Parekh et al.[144] leveraged class-agnostic proposals from both video frames to model the problem as a multiple-instance learning task for audio. As a result, the classification and localization problems could be solved simultaneously. The researchers focused on localizing salient audio and visual components using event classes in a weakly supervised manner. This framework was able to deal with the difficult case of asynchronous audio-visual events. To leverage more detailed relations between modalities, Hu et al.[141] recommended a deep coclustering model that extracted a set of distinct components from each modality. The model continually learned the correspondence between such representations of different modalities and further introduced K-means clustering to distinguish concrete objects or sounds.

5.3 Discussions

Representation learning between audio and visual is an emerging topic in deep learning. For single modality representation learning, existing efforts usually train an audio network to correlate with visual outputs. The visual networks are pre-trained with fixed parameters acting as a teacher. In order to learn audio and visual representation simultaneously, some efforts usually use the natural audio-visual correspondence in videos. However, this weak constraint cannot enforce models to produce precise information. Therefore, some efforts were proposed to solve this dilemma by adding more constraints such as class-agnostic proposals, corresponded or not, negative or positive samples, etc. Moreover, some works tend to exploit more precise constraints, for example, synchronized or not, harder negative samples, asynchronous audio-visual events, etc. By adding these constraints, their model can achieve better performance.

6 Recent public audio-visual datasets

Many audio-visual datasets ranging from speech-related to event-related data have been collected and released. We divide datasets into two categories: audio-visual speech datasets that record human faces with the corresponding speech, and audio-visual event datasets consisting of musical instrument videos and real events′ videos. In this section, we summarize the information of recent audio-visual datasets (Table 10 and Fig. 10).

6.1 Audio-visual speech datasets

Constructing datasets containing audio-visual corpora is crucial to understanding audio-visual speech. The datasets are collected in lab-controlled environments where volunteers read the prepared phrases or sentences, or in-the-wild environments of TV interviews or talks.

6.1.1 Lab-controlled environment

Lab-controlled speech datasets are captured in specific environments, where volunteers are required to read the given phases or sentences. Some of the datasets only contain videos of speakers that utter the given sentences; these datasets include GRID[82], TCD TIMIT[145], and VidTIMIT[146]. Such datasets can be used for lip reading, talking face generation, and speech reconstruction. Development of more advanced datasets has continued: e.g., Livingstone et al.[147] offered the RAVDESS dataset that contained emotional speeches and songs. The items in it are also rated according to emotional validity, intensity, and authenticity.

Some datasets such as Lombard Grid[148] and OuluVS[149, 150] focus on multiview videos. In addition, a dataset named SEWA[151] offers rich annotations, including answers to a questionnaire, facial landmarks, LLD (low-level descriptors) features, hand gestures, head gestures, transcript, valence, arousal, liking or disliking, template behaviors, episodes of agreement or disagreement, and episodes of mimicry. MEAD[152] is a large-scale, high-quality emotional audio-visual dataset that contains 60 actors and actresses talking with eight different emotions at three different intensity levels. This large-scale emotional dataset can be applied to many fields, such as conditional generation, cross-modal understanding, and expression recognition.

6.1.2 In-the-wild environment

The above datasets were collected in lab environments; as a result, models trained on those datasets are difficult to apply in real-world scenarios. Thus, researchers have tried to collect real-world videos from TV interviews, talks, and movies and released several real-world datasets, including LRW, LRW variants[84, 153, 154], Voxceleb and its variants[155, 156], AVA-ActiveSpeaker[157], and AVSpeech[11]. The LRW dataset consists of 500 sentences[153], while its variant contains 1 000 sentences[84, 154], all of which were spoken by hundreds of different speakers. VoxCeleb and its variants contain over 100 000 utterances of 1 251 celebrities[155] and over a million utterances of 6 112 identities[156].

AVA-ActiveSpeaker[157] and AVSpeech[11] datasets contain even more videos. The AVA-ActiveSpeaker[157] dataset consists of 3.65 million human-labeled video frames (approximately 38.5 h). The AVSpeech[11] dataset contains approximately 4 700 h of video segments from a total of 290 000 YouTube videos spanning a wide variety of people, languages, and face poses. The details are reported in Table 10.

6.2 Audio-visual event datasets

Another audio-visual dataset category consists of music or real-world event videos. These datasets are different from the aforementioned audio-visual speech datasets in not being limited to facial videos.

6.2.1 Music-related datasets

Most music-related datasets were constructed in a lab environment. For example, ENST-Drums[158] merely contains drum videos of three professional drummers specializing in different music genres. The C4S dataset[159] consists of 54 videos of 9 distinct clarinetists, each performing three different classical music pieces twice (4.5 h in total).

The URMP[160] dataset contains a number of multi-instrument musical pieces. However, these videos were recorded separately and then combined. To simplify the use of the URMP dataset, Chen et al.[92] further proposed the Sub-URMP dataset that contains multiple video frames and audio files extracted from the URMP dataset.

6.2.2 Real events-related datasets

More and more real-world audio-visual event datasets have recently been released, that consisting numerous videos uploaded to the Internet. The datasets often comprise hundreds or thousands of event classes and the corresponding videos. Representative datasets include the following.

Kinetics-400[161], Kinetics-600[162] and Kinetics-700[163] contain 400, 600 and 700 human action classes with at least 400, 600 and 700 video clips for each action, respectively. Each clip lasts approximately 10 s and is taken from a distinct YouTube video. The actions cover a broad range of classes, including human-object interactions such as playing instruments, as well as human-human interactions such as shaking hands. The AVA-Actions dataset[164] densely annotated 80 atomic visual actions in 43 015 min of movie clips, where actions were loc-

Table 10 Summary of speech-related audio-visual datasets. These datasets can be used for all tasks related to speech we have mentioned above. Note that the length of a speech dataset denotes the number of video clips, while for music or real event datasets, the length represents the total number of hours of the dataset.

Category	Dataset	Environment	Classes	Length*	Year
Speech	GRID[82]	Lab	34	33 000	2006
	Lombard Grid[148]	Lab	54	54 000	2018
	TCD TIMIT[145]	Lab	62	–	2015
	Vid TIMIT[146]	Lab	43	–	2009
	RAVDESS[147]	Lab	24	–	2018
	SEWA[151]	Lab	180	–	2017
	OuluVS[149]	Lab	20	1 000	2009
	OuluVS2[150]	Lab	52	3 640	2016
	MEAD[152]	Lab	60	281 400	2020
	Voxceleb[155]	Wild	1 251	154 516	2017
	Voxceleb2[156]	Wild	6 112	1 128 246	2018
	LRW[153]	Wild	~1 000	500 000	2016
	LRS[84]	Wild	~1 000	118 116	2017
	LRS3[154]	Wild	~1 000	74 564	2017
	AVA-ActiveSpeaker[157]	Wild	–	90 341	2019
Music	C4S[159]	Lab	–	4.5	2017
	ENST-Drums[158]	Lab	–	3.75	2006
	URMP[160]	Lab	–	1.3	2019
Real event	YouTube-8M[166]	Wild	3 862	350 000	2016
	AudioSet[165]	Wild	632	4 971	2016
	Kinetics-400[161]	Wild	400	850*	2018
	Kinetics-600[162]	Wild	600	1 400*	2018
	Kinetics-700[163]	Wild	700	1 806*	2018

Fig. 10 Demonstration of audio-visual datasets

alized in space and time, resulting in 1.58 M action labels with multiple labels corresponding to a certain person.

AudioSet[165], a more general dataset, consists of an expanding ontology of 632 audio event classes and a collection of 2 084 320 human-labeled 10-second sound clips. The clips were extracted from YouTube videos and cover a wide range of human and animal sounds, musical instruments and genres, and common everyday environmental sounds. YouTube-8M[166] is a large-scale labeled video dataset that consists of millions of YouTube video IDs with high-quality machine-generated annotations from a diverse vocabulary of 3 800+ visual entities.

7 Discussions

AVL is a foundation of the multimodality problem that integrates the two most important perceptions of our daily life. Despite great efforts focused on AVL, there is still a long way to go for real-life applications. In this section, we briefly discuss the key challenges and the potential research directions in each category.

7.1 Challenges

The heterogeneous nature of the discrepancy in AVL determines its inherent challenges. Audio tracks use a level of electrical voltage to represent analog signals, while the visual modality is usually represented in the RGB color space; the large gap between the two poses a major challenge to AVL. The essence of this problem is to understand the relation between audio and vision, which is also the basic challenge of AVL.

Audio-visual separation and localization is a longstanding problem in many real-life applications. Regardless of the previous advances in speaker-related or recent object-related separation and localization, the main challenges are failing to distinguish the timbre of various objects and exploring ways of generating different objects′ sounds. Addressing these challenges requires us to carefully design the models or ideas (e.g., the attention mechanism) for dealing with different objects. **Audio-visual correspondence learning** has vast potential applications, such as those in criminal investigations, medical care, transportation, and other industries. Many studies have tried to map different modalities into the shared feature space. However, it is challenging to obtain satisfactory results since extracting clear and effective information from ambiguous input, and target modalities remains difficult. Therefore, sufficient prior information (the specific patterns people usually focus on) has a significant impact on obtaining more accurate results. **Audio and vision generation** focuses on empowered machine imagination. In contrast to the conventional discriminative problem, the task of cross-modality generation is to fit a mapping between probability distributions. Therefore, it is usually a many-to-many mapping problem that is difficult to learn. Moreover, despite the large difference between audio and visual modalities, humans are sensitive to the difference between real-world and generated results, and subtle artifacts can be easily noticed, making this task more challenging. Finally, **audio-visual representation learning** can be regarded as a generalization of other tasks. As we discussed before, both audio represented by electrical voltage and vision represented by the RGB color space are designed to be perceived by humans while not making it easy for a machine to discover the common features. The difficulty stems from having only two modalities and lacking explicit constraints. Therefore, the main challenge of this task is to find a suitable constraint. Unsupervised learning as a prevalent approach to this task provides a well-designed solution, while not having external supervision makes it difficult to achieve our goal. The challenge of the weakly supervised approach is to find correct implicit supervision.

7.2 Directions for future research

AVL has been an active research field for many years[20, 21] and is crucial to modern life. However, there are still many open questions in AVL due to the challenging nature of the domain itself and people′s increasing demands.

First, from a macro perspective, as AVL is a classic multimodality problem, its primary issue is to learn the mapping between modalities, specifically to map the attributes in audio and the objects in an image or a video. We think that mimicking the human learning process, e.g., by following the ideas of the attention mechanism and a memory bank, may improve the performance of learning this mapping. Furthermore, the second most difficult goal is to learn logical reasoning. Endowing a machine with the ability to reason is not only important for AVL but also an open question for the entire AI community. Instead of directly empowering a machine with the full logic capability, which is a long way to go from the current development state, we can simplify this problem and consider fully utilizing the prior information and constructing the knowledge graph. Building a comprehensive knowledge graph and leveraging it in specific areas properly may help machine thinking.

As to each task we have summarized before, Sections 2 and 3 can be referred to as the problem of understanding, while Sections 4 and 5 can be referred to as generation and representation learning, respectively. Significant advances in understanding and generation tasks such as lip-reading, speaker separation, and talking face generation have recently been achieved for human faces. The domain of faces is comparatively simple yet important since the scenes are normally constrained, and it has a sizable amount of available useful prior information. For example, consider a 3D face model. These faces usually have neutral expressions, while the emotions that are the basis of the face have not been studied well. Furthermore, apart from faces, the more complicated in-the-wild scenes with more conditions are worth considering. Adapting models to the new varieties of audio (stereoscopic audio) or vision (3D video and AR) also leads in a new direction. The datasets, especially large and high-quality ones that can significantly improve the performance of machine learning, are fundamental to the research community[167]. However, collecting a dataset is laborintensive and time-intensive[168]. However, collecting a dataset is labor-intensive and time-intensive. Small-sample learning also benefits the application of AVL. Learning representations, which is a more general and basic form of other tasks, can also mitigate the dataset problem. While recent studies lacked sufficient prior information or supervision to guide the training procedure, exploring suitable prior information may allow models to learn better representations.

Finally, many studies focus on building more complex networks to improve performance, and the resulting networks generally entail unexplainable mechanisms. To make a model or an algorithm more robust and explainable, it is necessary to learn the essence of the earlier explainable algorithms to advance AVL.

8 Conclusions

The desire to better understand the world from the human perspective has drawn considerable attention to audio-visual learning in the deep learning community. This paper provides a comprehensive review of recent audio-visual learning advances categorized into four research areas: audio-visual separation and localization, audio-visual correspondence learning, audio and visual generation, and audio-visual representation learning. Furthermore, we present a summary of datasets commonly used in audio-visual learning. The discussion section identifies the key challenges of each category, followed by potential research directions.

Acknowledgments

This work was supported by National Key Research and Development Program of China (No. 2016YFB1001001), Beijing Natural Science Foundation (No. JQ18017), and National Natural Science Foundation of China (No. 61976002).

Open Access

References

[1] R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, M. Ekelid. Speech recognition with primarily temporal cues. *Science*, vol. 270, no. 5234, pp. 303–304, 1995. DOI: 10.1126/science.270.5234.303.

[2] G. Krishna, C. Tran, J. G. Yu, A. H. Tewfik. Speech recognition with no speech or with noisy speech. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, Brighton, UK, pp. 1090−1094, 2019. DOI: 10.1109/ICASSP.2019.8683453.

[3] R. He, W. S. Zheng, B. G. Hu. Maximum correntropy criterion for robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1561–1576, 2011. DOI: 10.1109/TPAMI.2010.220.

[4] C. Y. Fu, X. Wu, Y. B. Hu, H. B. Huang, R. He. Dual variational generation for low shot heterogeneous face recognition. In *Proceedings of Advances in Neural Information Processing Systems*, Vancouver, Canada, pp. 2670–2679, 2019.

[5] S. G. Tong, Y. Y. Huang, Z. M. Tong. A robust face recognition method combining lbp with multi-mirror symmetry for images with various face interferences. *International Journal of Automation and Computing*, vol. 16, no. 5, pp. 671–682, 2019. DOI: 10.1007/s11633-018-1153-8.

[6] A. X. Li, K. X. Zhang, L. W. Wang. Zero-shot fine-grained classification by deep feature learning with semantics. *International Journal of Automation and Computing*, vol. 16, no. 5, pp. 563–574, 2019. DOI: 10.1007/s11633-019-1177-8.

[7] Y. F. Ding, Z. Y. Ma, S. G. Wen, J. Y. Xie, D. L. Chang, Z. W. Si, M. Wu, H. B. Ling. AP-CNN: Weakly supervised attention pyramid convolutional neural network for fine-grained visual classification. *IEEE Transactions on Image Processing*, vol. 30, pp. 2826–2836, 2021. DOI: 10.1109/TIP.2021.3055617.

[8] D. L. Chang, Y. F. Ding, J. Y. Xie, A. K. Bhunia, X. X. Li, Z. Y. Ma, M. Wu, J. Guo, Y. Z. Song. The devil is in the channels: Mutual-channel loss for fine-grained image classification. *IEEE Transactions on Image Processing*, vol. 29, pp. 4683–4695, 2020. DOI: 10.1109/TIP.2020.2973812.

[9] A. Gabbay, A. Ephrat, T. Halperin, S. Peleg. Seeing through noise: Visually driven speaker separation and enhancement. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, Calgary, Canada, pp. 3051−3055, 2018. DOI: 10.1109/ICASSP.2018.8462527.

[10] T. Afouras, J. S. Chung, A. Zisserman. The conversation: Deep audio-visual speech enhancement. In *Proceedings of the 19th Annual Conference of the International Speech*

Communication Association, Hyderabad, India, pp. 3244−3248, 2018. DOI: 10.21437/Interspeech.2018-1400.

[11] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, M. Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics*, vol. 37, no. 4, Article number 112, 2018. DOI: 10.1145/3197517.3201357.

[12] P. Morgado, N. Vasconcelos, T. Langlois, O. Wang. Self-supervised generation of spatial audio for 360° video. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 360−370, 2018. DOI: 10.5555/3326943.3326977.

[13] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville. Improved training of Wasserstein GANs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, USA, pp. 5769−5779, 2017. DOI: 10.5555/3295222.3295327.

[14] T. Karras, S. Laine, T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 4396−4405, 2019. DOI: 10.1109/CVPR.2019.00453.

[15] Y. Bengio, A. Courville, P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013. DOI: 10.1109/TPAMI.2013.50.

[16] R. Arandjelovic, A. Zisserman. Look, listen and learn. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Venice, Italy, pp. 609−617, 2017. DOI: 10.1109/ICCV.2017.73.

[17] B. Korbar, D. Tran, L. Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Montreal, Canada pp. 7774−7785, 2018. DOI: 10.5555/3327757.3327874.

[18] Y. Z. Isik, J. Le Roux, Z. Chen, S. Watanabe, J. R. Hershey. Single-channel multi-speaker separation using deep clustering. In *Proceedings of Interspeech 2016*, ISCA, San Francisco, USA, pp. 545−549, 2016. DOI: 10.21437/Interspeech.2016-1176.

[19] Y. Luo, Z. Chen, N. Mesgarani. Speaker-independent speech separation with deep attractor network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787−796, 2018. DOI: 10.1109/TASLP.2018.2795749.

[20] T. Darrell, J. W. Fisher III, P. Viola. Audio-visual segmentation and "the cocktail party effect". In *Proceedings of the 3rd International Conference on Multimodal Interfaces*, Springer, Beijing, China, pp. 32−40, 2000. DOI: 10.1007/3-540-40063-X_5.

[21] J. W. Fisher III, T. Darrell, W. T. Freeman, P. Viola. Learning joint statistical models for audio-visual fusion and segregation. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, Denver, USA, pp. 742−748, 2000. DOI: 10.5555/3008751.3008859.

[22] B. C. Li, K. Dinesh, Z. Y. Duan, G. Sharma. See and listen: Score-informed association of sound tracks to players in chamber music performance videos. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, New Orleans, USA, pp. 2906−2910, 2017. DOI: 10.1109/ICASSP.2017.7952688.

[23] J. Pu, Y. Panagakis, S. Petridis, M. Pantic. Audio-visual object localization and separation using low-rank and sparsity. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, New Orleans, USA, pp. 2901−2905, 2017. DOI: 10.1109/ICASSP.2017.7952687.

[24] S. Hochreiter, J. Schmidhuber. Long short-term memory. *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735.

[25] R. Lu, Z. Y. Duan, C. S. Zhang. Listen and look: Audio–visual matching assisted speech source separation. *IEEE Signal Processing Letters*, vol. 25, no. 9, pp. 1315–1319, 2018. DOI: 10.1109/LSP.2018.2853566.

[26] G. Morrone, S. Bergamaschi, L. Pasa, L. Fadiga, V. Tikhanoff, L. Badino. Face landmark-based speaker-independent audio-visual speech enhancement in multi-talker environments. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, Brighton, UK, pp. 6900−6904, 2019. DOI: 10.1109/ICASSP.2019.8682061.

[27] J. Hershey, J. Movellan. Audio-vision: Using audio-visual synchrony to locate sounds. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, Denver, USA, pp. 813−819, 1999. DOI: 10.5555/3009657.3009772.

[28] H. L. van Trees. *Optimum Array Processing: Part IV of Detection, Estimation and Modulation Theory*, New York, USA: Wiley-Interscience, 2002.

[29] A. Zunino, M. Crocco, S. Martelli, A. Trucco, A. Del Bue, V. Murino. Seeing the sound: A new multimodal imaging device for computer vision. In *Proceedings of IEEE International Conference on Computer Vision Workshop*, IEEE, Santiago, Chile, pp. 693−701, 2015. DOI: 10.1109/ICCVW.2015.95.

[30] R. H. Gao, R. Feris, K. Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 36−54, 2018. DOI: 10.1007/978-3-030-01219-9_3.

[31] R. Z. Gu, S. X. Zhang, Y. Xu, L. W. Chen, Y. X. Zou, D. Yu. Multi-modal multi-channel target speech separation. *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 530–541, 2020. DOI: 10.1109/JSTSP.2020.2980956.

[32] L. Y. Zhu, E. Rahtu. Separating sounds from a single image. [Online], Available: https://arxiv.org/abs/2007.07984, 2020.

[33] H. Izadinia, I. Saleemi, M. Shah. Multimodal analysis for identification and segmentation of moving-sounding objects. *IEEE Transactions on Multimedia*, vol. 15, no. 2, pp. 378–390, 2013. DOI: 10.1109/TMM.2012.2228476.

[34] A. Senocak, T. H. Oh, J. Kim, M. H. Yang, I. S. Kweon. Learning to localize sound source in visual scenes. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 4358−4366, 2018. DOI: 10.1109/CVPR.2018.00458.

[35] E. Hoffer, N. Ailon. Deep metric learning using triplet network. In *Proceedings of the 3rd International Workshop on Similarity-Based Pattern Recognition*, Springer, Copenhagen, Denmark, pp. 84−92, 2015. DOI: 10.1007/

978-3-319-24261-3_7.

[36] Y. Wu, L. C. Zhu, Y. Yan, Y. Yang. Dual attention matching for audio-visual event localization. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Korea, pp. 6291–6299, 2019. DOI: 10.1109/ICCV.2019.00639.

[37] Y. P. Tian, J. Shi, B. C. Li, Z. Y. Duan, C. L. Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 252–268, 2018. DOI: 10.1007/978-3-030-01216-8_16.

[38] R. Sharma, K. Somandepalli, S. Narayanan. Crossmodal learning for audio-visual speech event localization. [Online], Available: https://arxiv.org/abs/2003.04358, 2020.

[39] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, A. Torralba. The sound of pixels. In *Proceedings of 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 587–604, 2018. DOI: 10.1007/978-3-030-01246-5_35.

[40] H. Zhao, C. Gan, W. C. Ma, A. Torralba. The sound of motions. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Korea, pp. 1735–1744, 2019. DOI: 10.1109/ICCV.2019.00182.

[41] A. Rouditchenko, H. Zhao, C. Gan, J. McDermott, A. Torralba. Self-supervised audio-visual co-segmentation. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, Brighton, UK, pp. 2357–2361, 2019. DOI: 10.1109/ICASSP.2019.8682467.

[42] S. Parekh, A. Ozerov, S. Essid, N. Q. K. Duong, P. Pérez, G. Richard. Identify, locate and separate: Audio-visual object extraction in large video collections using weak supervision. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, IEEE, New Paltz, USA, pp. 268–272, 2019. DOI: 10.1109/WASPAA.2019.8937237.

[43] X. C. Sun, H. Jia, Z. Zhang, Y. Z. Yang, Z. Y. Sun, J. Yang. Sound localization and separation in three-dimensional space using a single microphone with a metamaterial enclosure. [Online], Available: https://arxiv.org/abs/1908.08160, 2019.

[44] K. Sriskandaraja, V. Sethu, E. Ambikairajah. Deep siamese architecture based replay detection for secure voice biometric. In *Proceedings of the 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India, pp. 671–675, 2018. DOI: 10.21437/Interspeech.2018-1819.

[45] R. Białobrzeski, M. Kośmider, M. Matuszewski, M. Plata, A. Rakowski. Robust Bayesian and light neural networks for voice spoofing detection. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, pp. 1028–1032, 2019. DOI: 10.21437/Interspeech.2019-2676.

[46] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, A. M. Gomez. A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, pp. 1068–1072, 2019. DOI: 10.21437/Interspeech.2019-2212.

[47] X. Wu, R. He, Z. N. Sun, T. N. Tan. A light CNN for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018. DOI: 10.1109/TIFS.2018.2833032.

[48] J. Chung, C. Gulcehre, K. Cho, Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. [Online], Available: https://arxiv.org/abs/1412.3555, 2014.

[49] A. Nagrani, S. Albanie, A. Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 8427–8436, 2018. DOI: 10.1109/CVPR.2018.00879c.

[50] A. Torfi, S. M. Iranmanesh, N. M. Nasrabadi, J. Dawson. 3D convolutional neural networks for audio-visual recognition. [Online], Available: https://arxiv.org/abs/1706.05739, 2017.

[51] K. Simonyan, A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 568–576, 2014. DOI: 10.5555/2968826.2968890.

[52] Y. D. Wen, M. Al Ismail, W. Y. Liu, B. Raj, R. Singh. Disjoint mapping network for cross-modal matching of voices and faces. In *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, USA, 2019.

[53] S. Ioffe, C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, Lille, France, pp. 448–456, 2015.

[54] C. Lippert, R. Sabatini, M. C. Maher, E. Y. Kang, S. Lee, O. Arikan, A. Harley, A. Bernal, P. Garst, V. Lavrenko, K. Yocum, T. Wong, M. F. Zhu, W. Y. Yang, C. Chang, T. Lu, C. W. H. Lee, B. Hicks, S. Ramakrishnan, H. B. Tang, C. Xie, J. Piper, S. Brewerton, Y. Turpaz, A. Telenti, R. K. Roby, F. J. Och, J. C. Venter. Identification of individuals by trait prediction using whole-genome sequencing data. In *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 38, pp. 10166–10171, 2017. DOI: 10.1073/pnas.1711125114.

[55] K. Hoover, S. Chaudhuri, C. Pantofaru, M. Slaney, I. Sturdy. Putting a face to the voice: Fusing audio and visual signals across a video to determine speakers. [Online], Available: https://arxiv.org/abs/1706.00079, 2017.

[56] S. W. Chung, J. S. Chung, H. G. Kang. Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, Brighton, UK, pp. 3965–3969, 2019. DOI: 10.1109/ICASSP.2019.8682524.

[57] R. Wang, H. B. Huang, X. F. Zhang, J. X. Ma, A. H. Zheng. A novel distance learning for elastic cross-modal audio-visual matching. In *Proceedings of IEEE International Conference on Multimedia & Expo Workshops*, IEEE, Shanghai, China, pp. 300–305, 2019. DOI: 10.1109/ICMEW.2019.00-70.

[58] A. H. Zheng, M. L. Hu, B. Jiang, Y. Huang, Y. Yan, B. Luo. Adversarial-metric learning for audio-visual cross-modal matching. *IEEE Transactions on Multimedia*, 2021. DOI: 10.1109/TMM.2021.3050089.

[59] R. K. Srihari. Combining text and image information in content-based retrieval. In *Proceedings of International Conference on Image Processing*, IEEE, Washington, USA, pp. 326–329, 1995. DOI: 10.1109/ICIP.1995.529712.

[60] L. R. Long, L. E. Berman, G. R. Thoma. Prototype client/server application for biomedical text/image retriev-

al on the Internet. In *Proceedings of Storage and Retrieval for Still Image and Video Databases IV*, SPIE, San Jose, USA, vol. 2670, pp. 362–372, 1996. DOI: 10.1117/12.234775.

[61] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM International Conference on Multimedia*, ACM, Firenze, Italy, pp. 251–260, 2010. DOI: 10.1145/1873951.1873987.

[62] Y. Aytar, C. Vondrick, A. Torralba. See, hear, and read: Deep aligned representations. [Online], Available: https://arxiv.org/abs/1706.00932, 2017.

[63] D. Surís, A. Duarte, A. Salvador, J. Torres, X. Giró-i-Nieto. Cross-modal embeddings for video and audio retrieval. In *Proceedings of European Conference on Computer Vision Workshop*, Springer, Munich, Germany, pp. 711–716, 2019. DOI: 10.1007/978-3-030-11018-5_62.

[64] S. Hong, W. Im, H. S. Yang. Content-based video-music retrieval using soft intra-modal structure constraint. [Online], Available: https://arxiv.org/abs/1704.06761, 2017.

[65] A. Nagrani, S. Albanie, A. Zisserman. Learnable PINs: Cross-modal embeddings for person identity. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 73–89, 2018. DOI: 10.1007/978-3-030-01261-8_5.

[66] D. H. Zeng, Y. Yu, K. Oyama. Deep triplet neural networks with cluster-CCA for audio-visual cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 16, no. 3, Article number 76, 2020. DOI: 10.1145/3387164.

[67] V. Sanguineti, P. Morerio, N. Pozzetti, D. Greco, M. Cristani, V. Murino. Leveraging acoustic images for effective self-supervised audio representation learning. In *Proceedings of the 16th European Conference on Computer Vision*, Springer, Glasgow, UK, pp. 119–135, 2020. DOI: 10.1007/978-3-030-58542-6_8.

[68] Y. X. Chen, X. Q. Lu, S. Wang. Deep cross-modal image–voice retrieval in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 10, pp. 7049–7061, 2020. DOI: 10.1109/TGRS.2020.2979273.

[69] N. Takashima, F. Li, M. Grzegorzek, K. Shirahama. Cross-modal music-emotion retrieval using DeepCCA. *Information Technology in Biomedicine*, E. Pietka, P. Badura, J. Kawa, W. Wieclawek, Eds., Cham, Germany: Springer, pp. 133–145, 2021. DOI: 10.1007/978-3-030-49666-1_11.

[70] I. Kansizoglou, L. Bampis, A. Gasteratos. An active learning paradigm for online audio-visual emotion recognition. *IEEE Transactions on Affective Computing*, 2019. DOI: 10.1109/TAFFC.2019.2961089.

[71] S. Dupont, J. Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000. DOI: 10.1109/6046.865479.

[72] S. Petridis, M. Pantic. Prediction-based audiovisual fusion for classification of non-linguistic vocalisations. *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 45–58, 2016. DOI: 10.1109/TAFFC.2015.2446462.

[73] G. Potamianos, C. Neti, G. Gravier, A. Garg, A. W. Senior. Recent advances in the automatic recognition of audiovisual speech. In *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003. DOI: 10.1109/JPROC.2003.817150.

[74] D. Hu, X. L. Li, X. Q. Lu. Temporal multimodal learning in audiovisual speech recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 3574–3582, 2016. DOI: 10.1109/CVPR.2016.389.

[75] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, Bellevue, USA, pp. 689–696, 2011. DOI: 10.5555/3104482.3104569.

[76] H. Ninomiya, N. Kitaoka, S. Tamura, Y. Iribe, K. Takeda. Integration of deep bottleneck features for audio-visual speech recognition. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association*, Dresden, Germany, pp. 563–567, 2015.

[77] S. Petridis, Z. W. Li, M. Pantic. End-to-end visual speech recognition with LSTMS. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, New Orleans, USA, pp. 2592–2596, 2017. DOI: 10.1109/ICASSP.2017.7952625.

[78] M. Wand, J. Koutník, J. Schmidhuber. Lipreading with long short-term memory. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, Shanghai, China, pp. 6115–6119, 2016. DOI: 10.1109/ICASSP.2016.7472852.

[79] Y. M. Assael, B. Shillingford, S. Whiteson, N. de Freitas. LipNet: Sentence-level lipreading. [Online], Available: https://arxiv.org/abs/1611.01599v1, 2016.

[80] T. Stafylakis, G. Tzimiropoulos. Combining residual networks with LSTMs for lipreading. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, pp. 3652–3656, 2017. DOI: 10.21437/Interspeech.2017-85.

[81] T. Makino, H. Liao, Y. Assael, B. Shillingford, B. Garcia, O. Braga, O. Siohan. Recurrent neural network transducer for audio-visual speech recognition. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, IEEE, Singapore, pp. 905–912, 2019. DOI: 10.1109/ASRU46091.2019.9004036.

[82] M. Cooke, J. Barker, S. Cunningham, X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006. DOI: 10.1121/1.2229005.

[83] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, S. Zafeiriou. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, Shanghai, China, pp. 5200–5204, 2016. DOI: 10.1109/ICASSP.2016.7472669.

[84] J. S. Chung, A. Senior, O. Vinyals, A. Zisserman. Lip reading sentences in the wild. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, pp. 3444–3453, 2017. DOI: 10.1109/CVPR.2017.367.

[85] M. Nussbaum-Thom, J. Cui, B. Ramabhadran, V. Goel. Acoustic modeling using bidirectional gated recurrent convolutional units. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association*, San Francisco, USA, pp. 390–394, 2016. DOI: 10.21437/Interspeech.2016-212.

[86] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, A. Zisserman. Deep audio-visual speech recognition. [Online], Available: https://arxiv.org/abs/1809.02108, 2018.

[87] Y. Y. Zhang, Z. R. Wang, J. Du. Deep fusion: An attention guided factorized bilinear pooling for audio-video emotion recognition. In *Proceedings of International Joint Conference on Neural Networks*, IEEE, Budapest, Hungary, pp. 1–9, 2019. DOI: 10.1109/IJCNN.2019.8851942.

[88] P. Zhou, W. W. Yang, W. Chen, Y. F. Wang, J. Jia. Modality attention for end-to-end audio-visual speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, Brighton, UK, pp. 6565–6569, 2019. DOI: 10.1109/ICASSP.2019.8683733.

[89] R. J. Tao, R. K. Das, H. Z. Li. Audio-visual speaker recognition with a cross-modal discriminative network. In *Proceedings of the 21st Annual Conference of the International Speech Communication Association*, Shanghai, China, pp. 2242–2246, 2020. DOI: 10.21437/Interspeech.2020-1814.

[90] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 2672–2680, 2014. DOI: 10.5555/2969033.29691250.

[91] M. Arjovsky, S. Chintala, L. Bottou. Wasserstein GAN. [Online], Available: https://arxiv.org/abs/1701.07875, 2017.

[92] L. L. Chen, S. Srivastava, Z. Y. Duan, C. L. Xu. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia*, ACM, Mountain View, USA, pp. 349–357, 2017. DOI: 10.1145/3126686.3126723.

[93] H. Zhu, H. B. Huang, Y. Li, A. H. Zheng, R. He. Arbitrary talking face generation via attentional audio-visual coherence learning. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, Yokohama, Japan, pp. 2362–2368, 2020. DOI: 10.24963/ijcai.2020/327.

[94] L. H. Wei, S. L. Zhang, W. Gao, Q. Tian. Person transfer GAN to bridge domain gap for person re-identification. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 79–88, 2018. DOI: 10.1109/CVPR.2018.00016.

[95] S. W. Huang, C. T. Lin, S. P. Chen, Y. Y. Wu, P. H. Hsu, S. H. Lai. AugGAN: Cross domain adaptation with GAN-based data augmentation. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 731–744, 2018. DOI: 10.1007/978-3-030-01240-3_44.

[96] T. Le Cornu, B. Milner. Reconstructing intelligible audio speech from visual speech features. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association*, Dresden, Germany, pp. 3355–3359, 2015.

[97] A. Ephrat, S. Peleg. Vid2speech: Speech reconstruction from silent video. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, New Orleans, USA, pp. 5095–5099, 2017. DOI: 10.1109/ICASSP.2017.7953127.

[98] A. Ephrat, T. Halperin, S. Peleg. Improved speech reconstruction from silent video. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Venice, Italy, pp. 455–462, 2017. DOI: 10.1109/ICCVW.2017.61.

[99] T. Le Cornu, B. Milner. Generating intelligible audio speech from visual speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 9, pp. 1751–1761, 2017. DOI: 10.1109/TASLP.2017.2716178.

[100] A. Davis, M. Rubinstein, N. Wadhwa, G. J. Mysore, F. Durand, W. T. Freeman. The visual microphone: Passive recovery of sound from video. *ACM Transactions on Graphics*, vol. 33, no. 4, Article number 79, 2014. DOI: 10.1145/2601097.2601119.

[101] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, W. T. Freeman. Visually indicated sounds. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 2405–2413, 2016. DOI: 10.1109/CVPR.2016.264.

[102] Y. P. Zhou, Z. W. Wang, C. Fang, T. Bui, T. L. Berg. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 3550–3558, 2018. DOI: 10.1109/CVPR.2018.00374.

[103] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. C. Courville, Y. Bengio. SampleRNN: An unconditional end-to-end neural audio generation model. In *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France, 2017.

[104] H. Zhou, X. D. Xu, D. H. Lin, X. G. Wang, Z. W. Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *Proceedings of the 16th European Conference on Computer Vision*, Springer, Glasgow, UK, pp. 52–69, 2020. DOI: 10.1007/978-3-030-58610-2_4.

[105] C. H. Wan, S. P. Chuang, H. Y. Lee. Towards audio to scene image synthesis using generative adversarial network. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, Brighton, UK, pp. 496–500, 2019. DOI: 10.1109/ICASSP.2019.8682383.

[106] Y. Qiu, H. Kataoka. Image generation associated with music data. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, Salt Lake City, USA, pp. 2510–2513, 2018.

[107] W. L. Hao, Z. X. Zhang, H. Guan. CMCGAN: A uniform framework for cross-modal visual-audio mutual generation. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, the 30th Innovative Applications of Artificial Intelligence, and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence*, New Orleans, USA, pp. 6886–6893, 2018.

[108] J. G. Li, X. F. Zhang, C. M. Jia, J. Z. Xu, L. Zhang, Y. Wang, S. W. Ma, W. Gao. Direct speech-to-image translation. *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 517–529, 2020. DOI: 10.1109/JSTSP.2020.2987417.

[109] X. S. Wang, T. T. Qiao, J. H. Zhu, A. Hanjalic, O. Scharenborg. Generating images from spoken descriptions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 850–865, 2021. DOI: 10.1109/TASLP.2021.3053391.

[110] A. Duarte, F. Roldan, M. Tubau, J. Escur, S. Pascual, A. Salvador, E. Mohedano, K. McGuinness, J. Torres, X. Giro-i-Nieto. Wav2Pix: Speech-conditioned face generation using generative adversarial networks. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, Brighton, UK, pp. 8633–8637, 2019. DOI: 10.1109/ICASSP.2019.8682970.

[111] T. H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, W. Matusik. Speech2Face: Learning the face behind a voice. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 7531–7540, 2019 . DOI: 10.1109/CVPR.2019.00772.

[112] Y. D. Wen, B. Raj, R. Singh. Face reconstruction from voice using generative adversarial networks. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, Vancouver, Canada, pp. 5266–5275, 2019.

[113] A. A. Samadani, E. Kubica, R. Gorbet, D. Kulic. Perception and generation of affective hand movements. *International Journal of Social Robotics*, vol. 5, no. 1, pp. 35–51, 2013. DOI: 10.1007/s12369-012-0169-4.

[114] J. Tilmanne, T. Dutoit. Expressive gait synthesis using PCA and Gaussian modeling. In *Proceedings of the 3rd International Conference on Motion in Games*, Springer, Utrecht, The Netherlands, pp. 363–374, 2010. DOI: 10.1007/978-3-642-16958-8_34.

[115] M. Brand, A. Hertzmann. Style machines. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, ACM, New Orleans, USA, pp. 183–192, 2000. DOI: 10.1145/344779.344865.

[116] J. M. Wang, D. J. Fleet, A. Hertzmann. Multifactor Gaussian process models for style-content separation. In *Proceedings of the 24th International Conference on Machine Learning*, ACM, Corvalis, USA, pp. 975–982, 2007. DOI: 10.1145/1273496.1273619.

[117] G. W. Taylor, G. E. Hinton. Factored conditional restricted Boltzmann machines for modeling motion style. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM, Montreal, Canada, pp. 1025–1032, 2009. DOI: 10.1145/1553374.1553505.

[118] L. Crnkovic-Friis, L. Crnkovic-Friis. Generative choreography using deep learning. In *Proceedings of the 7th International Conference on Computational Creativity*, Paris, France, pp. 272–277, 2016.

[119] D. Holden, J. Saito, T. Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics*, vol. 35, no. 4, Article number 138, 2016. DOI: 10.1145/2897824.2925975.

[120] O. Alemi, J. Françoise, P. Pasquier. GrooveNet: Real-time music-driven dance movement generation using artificial neural networks. In *Proceedings of the 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining Workshop on Machine Learning for Creativity*, ACM, Halifax, Canada, pp. 26, 2017.

[121] J. Lee, S. Kim, K. Lee. Listen to dance: Music-driven choreography generation using autoregressive encoder-decoder network. [Online], Available: https://arxiv.org/abs/1811.00818, 2018.

[122] E. Shlizerman, L. Dery, H. Schoen, I. Kemelmacher-Shlizerman. Audio to body dynamics. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 7574–7583, 2018. DOI: 10.1109/CVPR.2018.00790.

[123] T. R. Tang, J. Jia, H. Y. Mao. Dance with melody: An LSTM-autoencoder approach to music-oriented dance synthesis. In *Proceedings of the 26th ACM International Conference on Multimedia*, ACM, Seoul, Republic of Korea, pp. 1598–1606, 2018. DOI: 10.1145/3240508.3240526.

[124] N. Yalta, S. Watanabe, K. Nakadai, T. Ogata. Weakly-supervised deep recurrent neural networks for basic dance step generation. In *Proceedings of International Joint Conference on Neural Networks*, IEEE, Budapest, Hungary, 2019. DOI: 10.1109/IJCNN.2019.8851872.

[125] R. Kumar, J. Sotelo, K. Kumar, A. de Brébisson, Y. Bengio. ObamaNet: Photo-realistic lip-sync from text. [Online], Available: https://arxiv.org/abs/1801.01442, 2017.

[126] A. Graves, J. Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, vol. 18, no. 5–6, pp. 602–610, 2005. DOI: 10.1016/j.neunet.2005.06.042.

[127] S. Suwajanakorn, S. M. Seitz, I. Kemelmacher-Shlizerman. Synthesizing Obama: Learning lip sync from audio. *ACM Transactions on Graphics*, vol. 36, no. 4, Article number 95, 2017. DOI: 10.1145/3072959.3073640.

[128] A. Jamaludin, J. S. Chung, A. Zisserman. You said that?: Synthesising talking faces from audio *International Journal of Computer Vision*, vol. 127, no. 11–12, pp. 1767–1779, 2019. DOI: 10.1007/s11263-019-01150-y.

[129] S. A. Jalalifar, H. Hasani, H. Aghajan. Speech-driven facial reenactment using conditional generative adversarial networks. [Online], Available: https://arxiv.org/abs/1803.07461, 2018.

[130] K. Vougioukas, S. Petridis, M. Pantic. End-to-end speech-driven facial animation with temporal GANs. In *Proceedings of British Machine Vision Conference*, Newcastle, UK, 2018.

[131] M. Saito, E. Matsumoto, S. Saito. Temporal generative adversarial nets with singular value clipping. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Venice, Italy, pp. 2849–2858, 2017. DOI: 10.1109/ICCV.2017.308.

[132] L. Chen, Z. H. Li, R. K. Maddox, Z. Y. Duan, C. L. Xu. Lip movements generation at a glance. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 538–553, 2018. DOI: 10.1007/978-3-030-01234-2_32.

[133] H. Zhou, Y. Liu, Z. W. Liu, P. Luo, X. G. Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence, the 31st Innovative Applications of Artificial Intelligence Conference, the 9th AAAI Symposium on Educational Advances in Artificial Intelligence*, Honolulu, USA, pp. 9299–9306, 2019. DOI: 10.1609/aaai.v33i01.33019299.

[134] L. L. Chen, R. K. Maddox, Z. Y. Duan, C. L. Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 7824–7833, 2019. DOI: 10.1109/CVPR.2019.00802.

[135] O. Wiles, A. S. Koepke, A. Zisserman. X2Face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 690–706, 2018. DOI: . DOI: 10.1007/978-3-030-01261-8_41.

[136] S. E. Eskimez, Y. Zhang, Z. Y. Duan. Speech driven talking face generation from a single image and an emotion condition. [Online], Available: https://arxiv.org/abs/2008.03592, 2020.

[137] S. E. Eskimez, R. K. Maddox, C. L. Xu, Z. Y. Duan. Noise-resilient training method for face landmark generation from speech. *IEEE/ACM Transactions on Audio,*

Speech, and Language Processing, vol. 28, pp. 27–38, 2020. DOI: 10.1109/TASLP.2019.2947741.

[138] Y. Aytar, C. Vondrick, A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS, Barcelona, Spain, pp. 892–900, 2016. DOI: 10.5555/3157096.3157196.

[139] R. Arandjelovic, A. Zisserman. Objects that sound. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 451–466, 2018. DOI: 10.1007/978-3-030-01246-5_27.

[140] K. Leidal, D. Harwath, J. Glass. Learning modality-invariant representations for speech and images. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, IEEE, Okinawa, Japan, pp. 424–429, 2017. DOI: 10.1109/ASRU.2017.8268967.

[141] D. Hu, F. P. Nie, X. L. Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 9240–9249. DOI: 10.1109/CVPR.2019.00947.

[142] A. Owens, A. A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 639–658, 2018. DOI: 10.1007/978-3-030-01231-1_39.

[143] Y. Bengio, J. Louradour, R. Collobert, J. Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM, Montreal, Canada, pp. 41–48, 2009. DOI: 10.1145/1553374.1553380.

[144] S. Parekh, S. Essid, A. Ozerov, N. Q. K. Duong, P. Pérez, G. Richard. Weakly supervised representation learning for unsynchronized audio-visual events. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, Salt Lake City, USA, pp. 2518–2519, 2018.

[145] N. Harte, E. Gillen. TCD-TIMIT: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015. DOI: 10.1109/TMM.2015.2407694.

[146] C. Sanderson, B. C. Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In *Proceedings of the 3rd International Conference on Advances in Biometrics*, Springer, Alghero, Italy, pp. 199–208, 2009. DOI: 10.1007/978-3-642-01793-3_21.

[147] S. R. Livingstone, F. A. Russo. The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS One*, vol. 13, no. 5, Article number e0196391, 2018. DOI: 10.1371/journal.pone.0196391.

[148] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, G. J. Brown. A corpus of audio-visual Lombard speech with frontal and profile views. *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. EL523–EL529, 2018. DOI: 10.1121/1.5042758.

[149] G. Y. Zhao, M. Barnard, M. Pietikainen. Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1254–1265, 2009. DOI: 10.1109/TMM.2009.2030637.

[150] I. Anina, Z. H. Zhou, G. Y. Zhao, M. Pietikäinen. OuluVs2: A multi-view audiovisual database for non-rigid mouth motion analysis. In *Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, IEEE, Ljubljana, Slovenia, pp. 1–5, 2015. DOI: 10.1109/FG.2015.7163155.

[151] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller, K. Star, E. Hajiyev, M. Pantic. SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 1022–1040, 2021. DOI: 10.1109/TPAMI.2019.2944808.

[152] K. S. Y. Wang, Q. Y. Wu, L. S. Song, Z. Q. Yang, W. Wu, C. Qian, R. He, Y. Qiao, C. C. Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *Proceedings of the 16th European Conference on Computer Vision*, Springer, Glasgow, UK, pp. 700–717, 2020. DOI: 10.1007/978-3-030-58589-1_42.

[153] J. S. Chung, A. Zisserman. Lip reading in the wild. In *Proceedings of the 13th Asian Conference on Computer Vision*, Springer, Taipei, China, pp. 87–103, 2017. DOI: 10.1007/978-3-319-54184-6_6.

[154] J. S. Chung, A. Zisserman. Lip reading in profile. In *Proceedings of British Machine Vision Conference 2017*, BMVA Press, London, UK, 2017.

[155] A. Nagrani, J. S. Chung, A. Zisserman. VoxCeleb: A large-scale speaker identification dataset. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, pp. 2616–2620, 2017. DOI: 10.21437/Interspeech.2017-950.

[156] J. S. Chung, A. Nagrani, A. Zisserman. VoxCeleb2: Deep speaker recognition. In *Proceedings of the 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India, pp. 1086–1090, 2018. DOI: 10.21437/Interspeech.2018-1929.

[157] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. H. Xi, C. Pantofaru. Supplementary material: AVA-ActiveSpeaker: An audio-visual dataset for active speaker detection. In *Proceedings of IEEE/CVF International Conference on Computer Vision Workshop*, IEEE, Seoul, Korea, pp. 3718–3722, 2019. DOI: 10.1109/ICCVW.2019.00460.

[158] O. Gillet, G. Richard. ENST-drums: An extensive audio-visual database for drum signals processing. In *Proceedings of the 7th International Conference on Music Information Retrieval*, Victoria, Canada, pp. 156–159, 2006.

[159] A. Bazzica, J. C. van Gemert, C. C. S. Liem, A. Hanjalic. Vision-based detection of acoustic timed events: A case study on clarinet note onsets. [Online], Available: https://arxiv.org/abs/1706.09556, 2017.

[160] B. C. Li, X. Z. Liu, K. Dinesh, Z. Y. Duan, 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. DOI: 10.1109/TMM.2018.2856090.

[161] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, A. Zisserman. The kinetics human action video dataset. [Online], Available: https://arxiv.org/abs/1705.06950, 2017.

[162] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, A. Zisserman. A short note about kinetics-600. [Online], Available: https://arxiv.org/abs/1808.01340, 2018.

[163] J. Carreira, E. Noland, C. Hillier, A. Zisserman. A short note on the kinetics-700 human action dataset. [Online],

Available: https://arxiv.org/abs/1907.06987, 2019.

[164] C. H. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Q. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, J. Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 6047–6056, 2018. DOI: 10.1109/CVPR.2018.00633.

[165] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, New Orleans, USA, pp. 776–780, 2017. DOI: 10.1109/ICASSP.2017.7952261.

[166] J. Lee, A. Natsev, W. Reade, R. Sukthankar, G. Toderici. The 2nd youtube-8m large-scale video understanding challenge. In *Proceedings of European Conference on Computer Vision*, Springer, Munich, Germany, pp. 193–205, 2019. DOI: 10.1007/978-3-030-11018-5_18.

[167] C. Sun, A. Shrivastava, S. Singh, A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Venice, Italy, pp. 843–852, 2017. DOI: 10.1109/ICCV.2017.97.

[168] O. M. Parkhi, A. Vedaldi, A. Zisserman. Deep face recognition. In *Proceedings of British Machine Vision Conference*, Swansea, UK, 2015.

Hao Zhu received the B. Eng. degree from Anhui Polytechnic University, China in 2018. He is currently a master student in Research Center of Cognitive Computing, Anhui University, China. He is also a joint master student in Center for Research on Intelligent Perception and Computing (CRIPAC), Institute of Automation, Chinese Academy of Sciences (CAS), China.

His research interests include deepfakes generation, computer vision, and pattern recognition.

E-mail: haozhu96@gmail.com

ORCID iD: 0000-0003-2155-1488

Man-Di Luo received the B. Eng. degree in automation engineering from University of Electronic Science and Technology of China, China in 2017, and the B. Sc. and M. Sc. degrees in electronic engineering from Katholieke University Leuven, Belgium in 2017 and 2018. She is currently a Ph. D. degree candidate in computer application technology at University of Chinese Academy of Sciences, China.

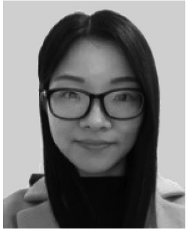
Her research interests include biometrics, pattern recognition, and computer vision.

E-mail: luomandi2019@ia.ac.cn

Rui Wang received the B. Sc. degree in computer science and technology from Hefei University, China in 2018. He is currently a master student in Department of Computer Science and Technology, Anhui University, China. He is also an intern at Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, China.

His research interests include style-transfer, computer vision, and pattern recognition.

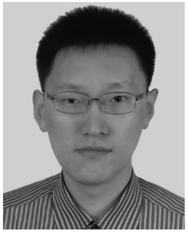
E-mail: rui.wang@cripac.ia.ac.cn



Ai-Hua Zheng received B. Eng. degree in computer science and technology from Anhui University of China, China in 2006 and 2008, and received the Ph. D. degree in computer science from University of Greenwich, UK in 2012. She visited University of Stirling, UK and Texas State University, USA during June to September in 2013 and September 2019 to August 2020 respectively. She is currently an associate professor and a Ph. D. supervisor in Anhui University, China.

Her research interests include vision based artificial intelligence and pattern recognition, especially on person/vehicle re-identification, audio visual computing, motion detection and tracking.

E-mail: ahzheng214@foxmail.com



Ran He received the B. Eng. and M. Sc. degrees in computer science from Dalian University of Technology, China in 2001 and 2004, and the Ph. D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CASIA), China in 2009. In September 2010, he joined NLPR, where he is currently a full professor. He is the Fellow of International Association for Pattern Recognition (IAPR). He serves as the editor board member of *Pattern Recognition*, and serves on the program committee for several conferences. His work won IEEE SPS Young Author Best Paper Award (2020), IAPR ICPR Best Scientific Paper Award (2020), IAPR/IEEE ICB Honorable Mention Paper Award (2019), IEEE ISM Best Paper Candidate (2016) and IAPR ACPR Best Poster Award (2013).

His research interests include information theoretic learning, pattern recognition, and computer vision.

Email: rhe@nlpr.ia.ac.cn (Corresponding author)

ORCID iD: 0000-0002-3807-991X