Deep Learning for Visual Speech Analysis: A Survey

Changchong Sheng, Gangyao Kuang, Liang Bai, Chenping Hou, Yulan Guo, Xin Xu, Matti Pietikäinen, and Li Liu*

Abstract—Visual speech, referring to the visual domain of speech, has attracted increasing attention due to its wide applications, such as public security, medical treatment, military defense, and film entertainment. As a powerful AI strategy, deep learning techniques have extensively promoted the development of visual speech learning. Over the past five years, numerous deep learning based methods have been proposed to address various problems in this area, especially automatic visual speech recognition and generation. To push forward future research on visual speech, this paper aims to present a comprehensive review of recent progress in deep learning methods on visual speech analysis. We cover different aspects of visual speech, including fundamental problems, challenges, benchmark datasets, a taxonomy of existing methods, and state-of-the-art performance. Besides, we also identify gaps in current research and discuss inspiring future research directions.

Index Terms—Deep Learning, Visual Speech, Lip Reading, Speech Perception, Computer Vision, Computer Graphics

1 Introduction

Human speech is by nature bimodal: visual and audio. Visual speech refers to the visual domain of speech, *i.e.*, the movements of the lips, tongue, teeth, jaw, *etc.*, and other facial muscles of the lower face that are naturally produced during talking [1], while audio speech refers to the acoustic waveform pronounced by the speaker. Speech perception is intrinsically bimodal, as shown several decades ago by the famous McGurk effect [2] that human speech perception depends not only on auditory information, but also on visual cues like lip movements. Therefore, undoubtedly, visual speech contributes to human speech perception, especially for people who are hearing-impaired or hard of hearing or when acoustic information is corrupted.

As a fundamental and challenging topic in computer vision and multimedia, automatic Visual Speech Analysis (VSA) has received increasing attention in recent years, due to the important role it plays in a wide variety of applications many of which are newly emerging. VSA embraces two fundamental closely-related and formal-dual problems: Visual Speech Recognition (VSR) or Lip Reading, Visual Speech Generation (VSG) or Lip Sequence Generation. Significant progress has been witnessed in this field due to the recent boom of deep learning. Typical academia and practical applications of VSA include multimodal speech recognition and enhancement, speaker recognition and verification [3], medical assistance, security, forensic, video compression, entertainment, human-computer interaction, emotion understanding [4, 5], *etc.*

To give some application examples, in speech recognition and enhancement, visual speech can be treated as a complementary signal to increase the accuracy and robustness of current audio speech recognition and separation under various unfavorable acoustic conditions [6, 7, 8, 9]. In the medical domain, solving the VSR task can also help the hearing impaired [10] and people with vocal cord lesions. In public security, VSA can be applied to face forgery detection [11] and liveness detection [12]. In human-computer interaction, visual speech can serve as a new type of interactive information, improving the diversity and robustness of interactions [13, 14]. In the entertainment domain, VSG technology plays a crucial role in personalized 3D talking avatars generation [15] in virtual gaming and realizing high-fidelity photo-realistic talking videos generation for movie post-production like visual dubbing [16]. In addition, VSR can be used to transcribe archival silent films.

The core of VSA lies in visual speech representation learning and sequence modeling. In the era dominated by traditional VSA methods, shallow representations of visual speech such as visemes [17, 18], mouth geometry descriptors [19], linear transformation features [20], statistical representations [21], and sequence modeling like Gaussian process dynamical models [22], hidden Markov models (HMMs) [23], decision tree models [24] were widely used in solving VSA tasks. Since the significant breakthroughs [25] of deep neural networks (DNNs) in the image classification task, most computer vision and natural language problems have focused explicitly on deep learning methods, including VSA. In 2016, deep learning based VSA methods [26, 27] have vastly outperformed traditional approaches, bringing the VSA into the deep learning era. Meanwhile, the emergence of large-scale VSA datasets [27, 28, 29, 30, 31] promoted the further development of deep learning based VSA research. In this paper, we mainly focus on the deep learning based VSA approaches. The milestones of VSA technologies from 2016 to the present are shown in Fig. 1, including representative deep VSR and VSG methods and related audio-visual datasets.

Although the recent promising progress brought by deep learning in the past several years, the VSA technology is still in its early stages and incapable of performing at a level sufficient for real-world applications. This is certainly not because of a lack of effort by researchers, as there have been many excellent works on VSA [6, 28, 32, 33, 34, 35]. Therefore, it is of great importance to systematically review the recent developments in the field, identify the main challenges and open problems preventing its advancement, and define promising future directions. However, the large part of VSA research remains rather scattered, and no such systematic surveys exist. This motivates this survey which will fill this gap.

1.1 The Scope of this Survey

The main objective of this survey is to provide a comprehensive overview of current deep learning based VSA methods, in particular, VSR and VSG and related applications, main challenges, benchmark datasets, methods, and State of the Art

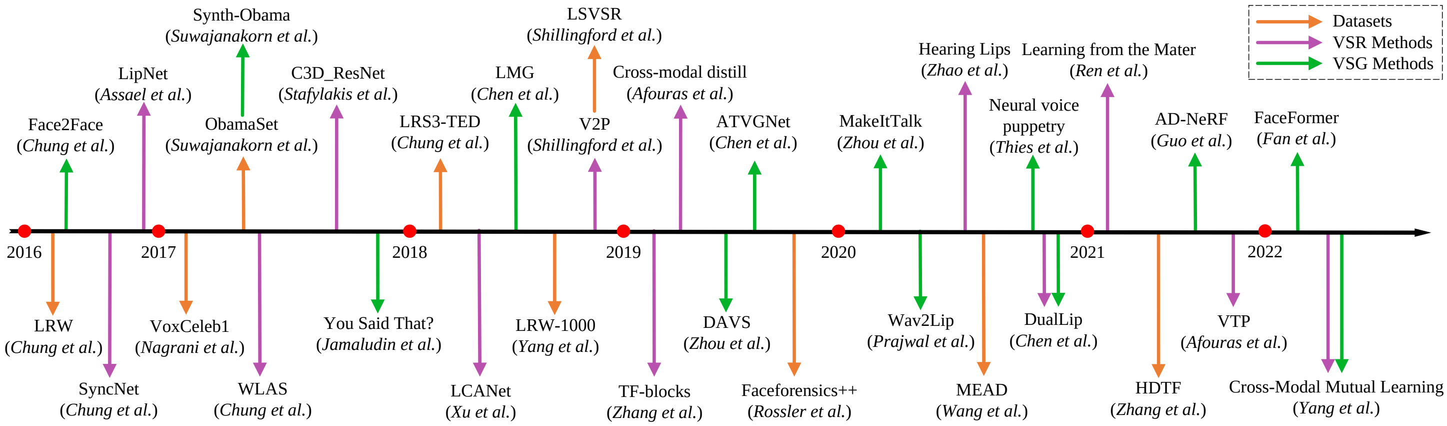

Fig. 1: Chronological milestones on visual speech analysis from 2016 to the present, including representative VSR and VSG methods, and audio-visual datasets. Handcrafted feature engineering methods dominated VSA until a transition took place in 2016 with the introduction of related deep networks.

(SOTA) results, together with existing gaps and promising future research directions.

There are mainly three reasons that we comprehensively overview VSR and VSG together. First, as the most fundamental problems in VSA, VSR and VSG cover most aspects of visual speech analysis. Other VSA-related tasks, such as speech enhancement, speaker verification, face forgery detection, *etc.*, can be seen as extended applications of VSR and VSG. Second, because VSR and VSG are formal-dual and mutually promoted, the dual learning [36] and generative adversarial learning mechanisms [37] are widely adopted in many existing VSA works [32, 38, 39, 40, 41]. Therefore, we intend to provide a side-by-side perspective for readers to know the evolution of VSR and VSG. Third, VSR and VSG have common core technical points, such as visual speech representation learning approaches and contextual sequence modeling approaches. We hope it would be helpful for readers to have an accessible understanding of the cross-task transferability of these methods.

1.2 Differences with Related Surveys

Several surveys on VSA [42, 43, 44, 45, 46] have been published. However, they have only partially reviewed specific VSA tasks. For example, [43, 44, 46] conducted reviews on VSR, and [42, 45] focused on VSG. We give a brief conclusion related surveys and then emphasize our new contributions.

In 2014, Zhou *et al.* [46] summarized three central questions of visual feature extraction for VSR: speaker dependency, head pose variation, and efficiently encoding of spatio-temporal information. Then they reviewed mainstream visual features extraction and dynamic audio-visual speech fusion methods of VSR from the view of question-solving, which brought a new perspective for readers to know the developments of VSR. In 2015, Mattheyses *et al.* [45] gave an extensive and comprehensive overview of audio-visual speech synthesis with great effort. We advocate readers to refer to [45] for the development history of VSG before 2015. As [45, 46] provided comprehensive surveys on traditional VSR and VSG methods, in this paper, we mainly focus on the recent advances caused by deep learning technologies. Adriana *et al.* [44] summarized VSR datasets according to the differences in recognition tasks and reviewed traditional and deep learning based methods of VSR. They mainly focused on the existing datasets, and the analysis of VSR methods for different recognition tasks on each dataset. However, the research reviewed in [44] is mainly pre-2018, prior to the recent striking success. Recently, Fenghour *et al.* [43] conducted a survey reviewing deep learning driven VSR methods, including audio-visual datasets, feature extraction, classification networks and classification schemes. However, some essential advances of VSR were omitted, such as self-supervised learning methods [47, 48, 49, 50], cross-modal knowledge distillation methods [34, 51, 52], graph neural networks backbone architectures [53, 54], *etc.* Chen *et al.* [42] conducted a thoughtful analysis across several representative identity-independent VSG methods and designed a performance evaluation benchmark for VSG. However, their core contributions are well-defined standards of evaluation metrics rather than the comprehensive discussion and overview of VSG methods.

Now we are in a place to summarize our main contributions in this paper.

- To the best of our knowledge, this is the ***first*** survey paper to systematically and comprehensively review deep learning methods for visual speech analysis, covering two fundamental problems, *i.e.*, visual speech recognition and visual speech generation.
- Problem definition, main challenges, benchmark datasets and testing protocols are summarized for each problem, and notably, the relationship among different VSA problems is also identified.
- We propose a taxonomy to group the prominent methods. In addition, performance comparisons, merits, and demerits of representative approaches and their underlying connections are also analyzed.
- Open issues and promising directions in this field are provided.

The remainder of this paper is structured as follows. The problem definitions and main challenges of VSA are summarized in Section 2. In Section 3, we review the audio-visual datasets and evaluation metrics and compare dataset attributes from multiple perspectives. Section 4 illustrates the general framework and representative methods for VSR. Section 5 provides a comprehensive survey of existing methods for VSG. A taxonomy of VSR and VSG methods is illustrated in Fig. 2. In Section 6, we conclude the paper and discuss the possible promising future research directions.

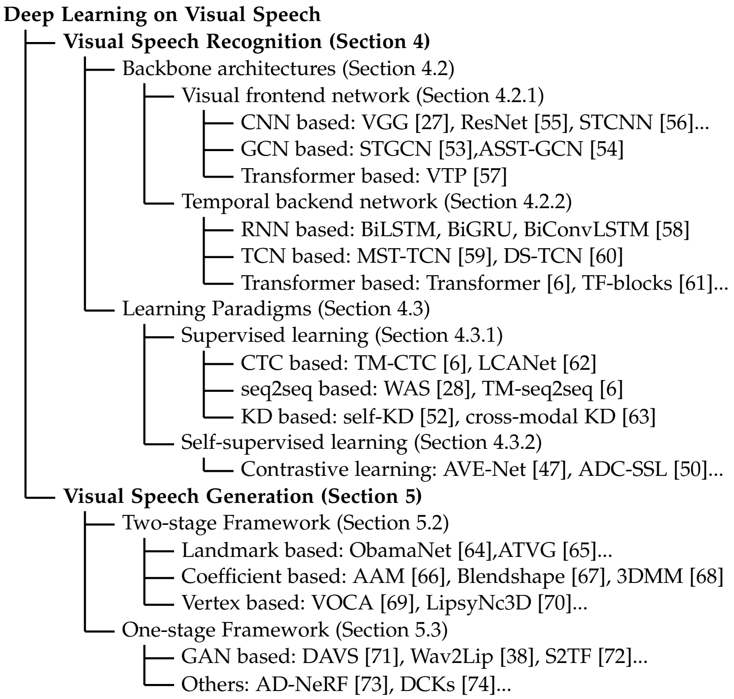
Deep Learning on Visual Speech
Visual Speech Recognition (Section 4)
Backbone architectures (Section 4.2)
Visual frontend network (Section 4.2.1)
CNN based: VGG [27], ResNet [55], STCNN [56]...
GCN based: STGCN [53],ASST-GCN [54]
Transformer based: VTP [57]
Temporal backend network (Section 4.2.2)
RNN based: BiLSTM, BiGRU, BiConvLSTM [58]
TCN based: MST-TCN [59], DS-TCN [60]
Transformer based: Transformer [6], TF-blocks [61]...
Learning Paradigms (Section 4.3)
Supervised learning (Section 4.3.1)
CTC based: TM-CTC [6], LCANet [62]
seq2seq based: WAS [28], TM-seq2seq [6]
KD based: self-KD [52], cross-modal KD [63]
Self-supervised learning (Section 4.3.2)
Contrastive learning: AVE-Net [47], ADC-SSL [50]...
Visual Speech Generation (Section 5)
Two-stage Framework (Section 5.2)
Landmark based: ObamaNet [64],ATVG [65]...
Coefficient based: AAM [66], Blendshape [67], 3DMM [68]
Vertex based: VOCA [69], LipsyNc3D [70]...
One-stage Framework (Section 5.3)
GAN based: DAVS [71], Wav2Lip [38], S2TF [72]...
Others: AD-NeRF [73], DCKs [74]...

Fig. 2: A taxonomy of representative visual speech recognition and generation methods.

2 Background

2.1 The Problems

Visual speech analysis can be divided into two fundamental problems: recognition and generation. As shown in Fig. 3, the two problems are formal-dual and have a reverse pipeline.

Visual speech recognition (VSR), also known as automatic lip reading, involves designing algorithms to infer the text content according to the speaker's mouth movements. Given a talking face video, a VSR system first crops the video and gets the mouth-centered cropped video. And then, it decodes the cropped video into a specific type of text (words, phrases or sentences). According to recognition targets, VSR mainly includes two types: word-level and sentence-lvel. The word-level VSR aims to classify the input video into one of a set of predefined word categories, while the sentence-level VSR tries to predict consecutive sentences from the input video. More specifically, a VSR system mainly consists of two sub-problems: visual speech representation learning and recognition. The extraction of discriminative visual speech features plays a relatively more important role since even the best recognizer will fail to achieve good results on poor visual speech features.

As a dual task of VSR, the goal of visual speech generation (VSG) is to synthesize a photo-realistic, high-quality talking video that corresponds to the driving source (*e.g.*, a piece of reference audio or text) and the target identity. Specifically, a VSG system first extracts speech representations from the driving source and then fuses the learned speech representations with the target identity to output continual talking frames. From the perspective of the learning target, the goal of VSG is more subjective and diverse than that of VSR, making VSG a more challenging problem than VSR.

2.2 Main Challenges

Despite several years of development, most VSA methods have not been capable of meeting real-world requirements due to various challenges. As illustrated in Fig. 4(a), to systematically present the challenges in VSA, we classify the main difficulties from recognition-related, and generation-related and discuss the challenges of audio-visual datasets. In Fig. 4(b)&(c), we provide some instances of typical challenges for intuitively understanding.

2.2.1 Recognition-related Challenges

From the perspective of representation learning, the ideal of visual speech recognition is to extract speech-related features with strong distinctiveness and robustness. However, both of the two goals suffer from severe practical challenges. Recognition-related challenges mainly stem from (1) the vast range of intra-class variations and (2) the inter-class similarities.

Intra-class variations can be organized into two types: visual speech intrinsic factors and other recognition-irrelevant factors. In terms of visual speech intrinsic factors, as shown in Fig. 4(b1), a word can produce quite different visual dynamics due to different contexts, continuous reading, speech emotion changing, speaking rate, *etc*. Meanwhile, many types of speech-irrelevant interference information dramatically impact visual speech recognition, such as speaker difference, head pose movement, facial expression, imaging condition, and so on. As illustrated in Fig. 4(b3), although the two speakers say the same word "*after*", their lip motions look different. Early research on visual speech recognition only considered frontal talking face videos in laboratory environments, limiting its availability in practical applications. Nowadays, visual speech recognition in-the-wild has attracted more and more attention, many audio-visual datasets were collected from various realistic scenes (TV News, public speech video...). Fig. 4(b4) shows an example video from the LRS3 dataset, a speaker is talking with dramatic head pose changes. it is hard to eliminate interference of these irrelevant motion information under the unconstrained environment. In addition to head pose changes, illumination, noise corruption, poor resolution *etc* also bring great difficulties to visual speech recognition.

Besides intra-class variations, visual speech recognition also suffers from inter-class similarities. As we know, the phoneme is the smallest recognizable unit of sound in a language that serves to distinguish words from one another. Similarly, viseme is the smallest recognizable unit of visual speech. The number of visemes is much smaller than that of phonemes. There are about three times as many phonemes as visemes in English. Therefore, several phonemes map onto a few visemes. Some phonemes, such as [*p*] and [*b*] [*k*] and [*g*], [*t*] and [*d*], *etc* have almost the same visual characteristics, so they are almost indistinguishable without considering the context in visual domain. We define this phenomenon as visual ambiguity, the leading cause of inter-class similarities. Fig. 4(b2) demonstrates an instance of word-level visual ambiguity. The other challenge of inter-class similarities stems from thousands of word classes. Subtle differences (*e.g*. various forms of words) in different word classes make the problem more difficult.

2.2.2 Generation-Related Challenges

Different from visual speech recognition, visual speech generation requires not only speech-related information but also identity-related information. As shown in Fig. 4(a), generation-related challenges mainly come from (1) information coupling, (2) diversity targets, and (3) evaluation validity.

A talking face video contains many types of coupled information, such as various motion-related information and

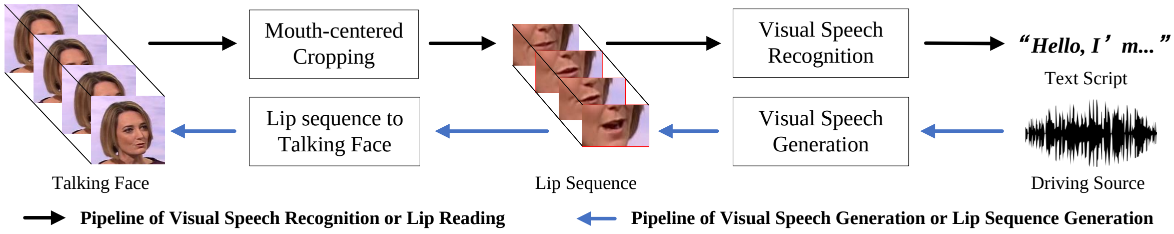
Mouth-centered Cropping
Visual Speech Recognition
"Hello, I' m..."
Text Script
Lip sequence to Talking Face
Visual Speech Generation
Talking Face
Lip Sequence
Driving Source
Pipeline of Visual Speech Recognition or Lip Reading
Pipeline of Visual Speech Generation or Lip Sequence Generation

Fig. 3: The two formal-dual fundamental problems of visual speech analysis. Top part: Visual speech recognition or lip reading; Bottom part: Visual speech generation or lip sequence generation.

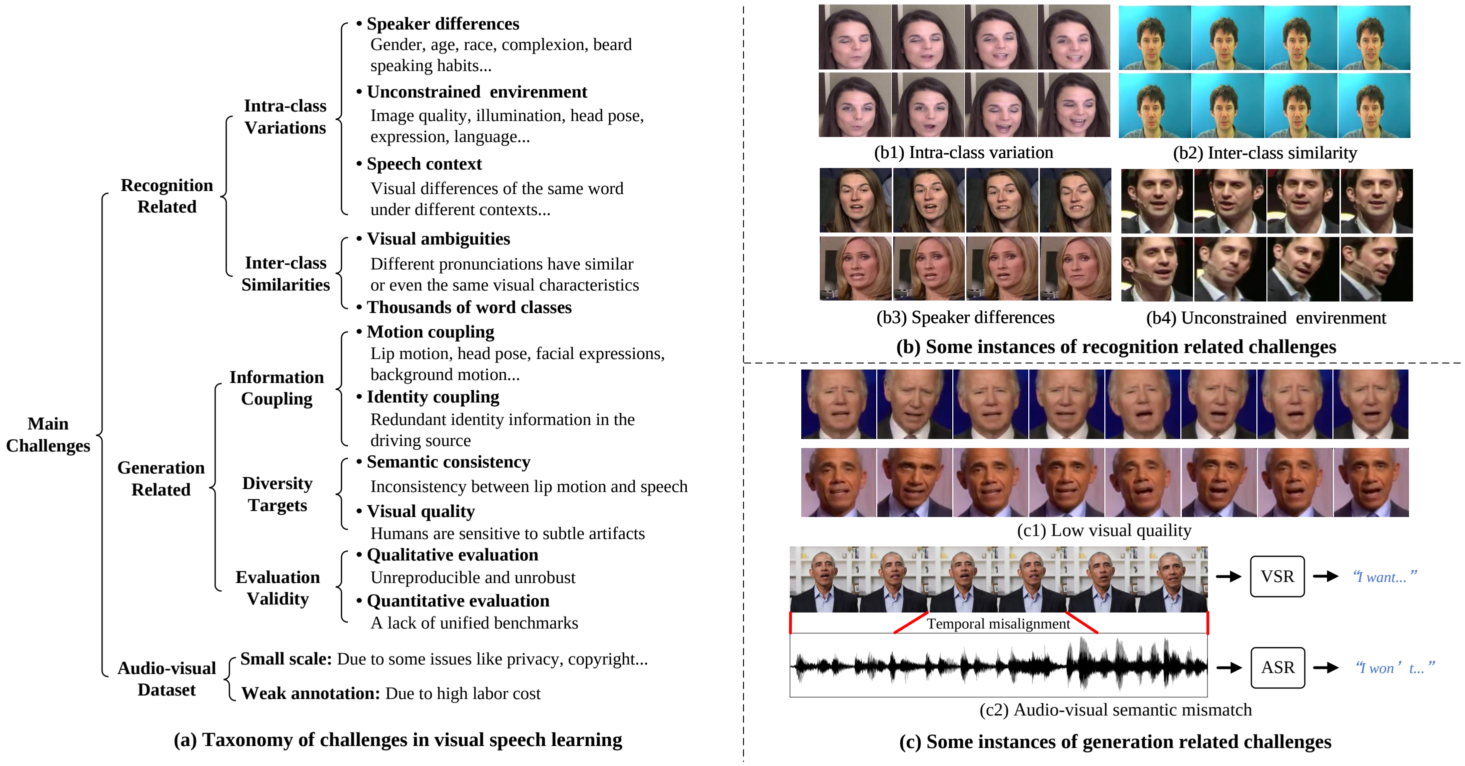
Main Challenges
Recognition Related
Intra-class Variations
• Speaker differences
Gender, age, race, complexion, beard speaking habits...
• Unconstrained envirenment
Image quality, illumination, head pose, expression, language...
• Speech context
Visual differences of the same word under different contexts...
Inter-class Similarities
• Visual ambiguities
Different pronunciations have similar or even the same visual characteristics
• Thousands of word classes
Generation Related
Information Coupling
• Motion coupling
Lip motion, head pose, facial expressions, background motion...
• Identity coupling
Redundant identity information in the driving source
Diversity Targets
• Semantic consistency
Inconsistency between lip motion and speech
• Visual quality
Humans are sensitive to subtle artifacts
Evaluation Validity
• Qualitative evaluation
Unreproducible and unrobust
• Quantitative evaluation
A lack of unified benchmarks
Audio-visual Dataset
Small scale: Due to some issues like privacy, copyright...
Weak annotation: Due to high labor cost
(a) Taxonomy of challenges in visual speech learning
(b1) Intra-class variation
(b2) Inter-class similarity
(b3) Speaker differences
(b4) Unconstrained envirenment
(b) Some instances of recognition related challenges
(c1) Low visual quaility
Temporal misalignment
VSR
"I want..."
ASR
"I won' t..."
(c2) Audio-visual semantic mismatch
(c) Some instances of generation related challenges

Fig. 4: Main Challenges of visual speech analysis. (a) A taxonomy of main challenges. (b) Some practical examples of different challenges. (b1) The upper and lower lines are the respectively different visual dynamics of the word *"wind"* under different contexts; (b2) The upper video instance refers to the word *"place"*, while the lower video refers to the word *"please"*. However, their visual dynamics are very similar; (b3) Two people speak the word *"after"* respectively, with a noticeable difference in their lip motions; (b4) An example of real-time changes in the head pose of a speaker during talking.

identity-related information. For motion coupling, motions that occur on a talking face video can be categorized into two types: intrinsic motions (head pose, facial expression, lip motion, *etc.*) and extrinsic motions (camera motion, background motion, *etc.*). All of these various motions are highly coupled. The motion coupling challenge stems not only from disentangling lip motion from all of these speech-irrelevant motions but also from integrating the synthesized lip sequence into a given identity image. The other coupling issue of visual speech generation is identity coupling. As illustrated in Fig. 4(c1), people may feel eerie and uncomfortable while observing these images due to the identity of generated face subtlely changed. This phenomenon, also known as the "uncanny valley effect" [100], occurs when people observe a synthetic face that's almost human-like but not quite perfect. Generally, the driving source contains rich information about the source identity. Therefore, the critical challenge is how to remove the identity information from the driving source to avoid corruption in the process of the target identity synthesis. Besides, most existing methods are only adaptive to specific target identity since different speakers have significant differences in appearance, speaking habits, *etc.* So, the lack of identity generalization is also an important challenge.

Semantic consistency and visual quality are the most desired properties of an excellent VSG method. Semantic consistency represents that the synthesized lip sequence should be synchronized and speech-consistent with the driving source. As shown in Fig. 4(c2), semantic consistency mainly involves two demands: temporal alignment and speech matching. However, the synchronous speech mapping between the driving source and the generated talking video is difficult to realize due to intrinsic differences in temporal resolution and speech characteristics of different data modalities. As for visual quality, There are two difficulties: (1) There is a lack of explicit training objectives since the fidelity and visual quality of the generated lip motion sequences are difficult to define quantitatively. (2) Because humans are sensitive to subtle artifacts, integrating the generated lip sequence into the whole face without people-oriented perceptual errors is a complicated issue.

TABLE 1: Statistics of commonly used audio-visual datasets.

Dataset Name	#Hours	#Vocab.	#Utter.	#Subj.	Image size FPS	Environment	Data Type	Year	Highlight	Download Link
AVICAR [75]	∼33	26† 13‡ 1317	59k	86	720×480 30	Car-driving	4-view face-centered videos	2004	Recorded in a car environment with various noise conditions; Consists of four scripts: isolated digits, isolated letters, phone numbers, and sentences	[76]
GRID [77]	∼28	51	33k	33	720×576 25	Lab-controlled	3-second face-centered videos	2006	Each sentence consists of a six-word sequence of the specific form	[78]
MODALITY [79]	∼31	182	5880	35	1920×1080 100	Lab-controlled	Stereoscopic RGB-D face-centered videos	2015	Command-like sentences; High resolution with varying noise conditions	[80]
OuluVS2 [81]	∼2	N/A	2120	53	1920×1080 30	Lab-controlled	5-view face-centered videos	2015	Three types of utterances; High recording quality; Five views from 0-90	[82]
IBM AV-ASR [83]	∼40	∼10.4k	N/A	262	704×480 30	Lab-controlled	Face-centered videos	2015	Large vocabulary size; Recorded in clean, studio conditions	Unpublic
LRW [27]	∼111	500	∼539K	1k+	256×256 25	In-the-wild	1.2-second face-centered videos	2016	Collected from British television programs; Each video corresponds to a word category	[84]
LRS2-BBC [6]	∼225	∼62.8k	∼144.5k	1k+	160×160 25	In-the-wild	Face-centered videos	2017	Collected from British television; Large-scale; Open-world; Sentence-level lip reading	[85]
VoxCeleb1 [30]	∼352	N/A	∼153.5k	1k+	N/A	In-the-wild	Public YouTube videos	2017	Large-scale; In-the-wild; Mainly for speaker identification and verification	[86]
ObamaSet [87]	∼14	N/A	N/A	1	N/A	In-the-wild	Public YouTube videos	2017	Focuses on Barack Obama; Collected from Obama's weekly presidential addresses; High quality	[88]
LRS3-TED [89]	∼475	∼71.1k	∼151.8k	5k+	224×224 25	In-the-wild	Face-centered videos	2018	Larger-scale; Along with the corresponding subtitles and word alignment boundaries	[90]
VoxCeleb2 [29]	∼2.4k	N/A	∼1.1m	6k+	N/A	In-the-wild	Public YouTube videos	2018	Significantly larger scale; A wide range of different ethnicities, accents, professions, and ages	[91]
LSVSR [92]	∼3.9k	∼127k	∼2.9m	N/A	N/A	In-the-wild	Face-centered videos	2018	The largest existing visual speech recognition dataset; Extracted from public YouTube videos	Unpublic
LRW-1000 [31]	∼57	1k	∼718K	2k+	N/A	In-the-wild	Mouth-centered videos	2019	The first large-scale Mandarin audio-visual speech dataset; collected from TV programs	[93]
Faceforensics++ [94]	∼5.7	N/A	∼1k	1k	N/A	In-the-wild	Manipulated talking videos	2019	Commonly used for facial forgery detection.	[95]
VOCASET [69]	N/A	N/A	255	12	5023 vertices 60	Lab-controlled	3d face mesh	2019	Higher quality 3D scans as well as alignments of the entire head	[96]
MEAD [97]	∼39	N/A	N/A	60	1920×1080 30	Lab-controlled	7-view face-centered videos	2020	Multi-view Emotional Audio-visual Dataset	[98]
HDTF [68]	∼15.8	N/A	10k+	300+	N/A	In-the-wild	Face-centered videos	2021	Higher video resolution than previous in-the-wild datasets	[99]

[1] †: Alphabets, ‡: Digits.

Besides the aforementioned difficulties, efficiently evaluating visual speech generation methods is another challenge. Existing evaluations on this problem, including qualitative and quantitative metrics, have many limitations. For example, qualitative metrics like user study are unreproducible and unstable. As for quantitative metrics, although there are a dozen of evaluation metrics, some of them are not appropriate and even mutually contradictory.

2.2.3 *Dataset-Related Challenges*

In addition to the above problem-oriented difficulties and challenges, audio-visual dataset-related issues also significantly impact the progress of VSA. Since that most of the current deep learning methods are data-driven, the importance of datasets is self-evident. However, existing audio-visual datasets suffer from small scale and weak annotations due to privacy protection and high labor cost. A potential research direction is to realize cross-modal self-supervised visual speech learning [49, 50, 101] based on unlabeled audio-visual data. Despite this, the issue of dataset scale limitation remains to be resolved.

3 Datasets and Evaluation Metrics

Datasets have played an important role throughout the history of visual speech research, especially in the big data era. First, benchmark datasets serve as a common platform for measuring and comparing performances of competing VSA algorithms; Second, as a typical data-driven learning strategy, deep learning technologies have made significant progress in many audio-visual learning tasks. It is worth noting that the large amounts of annotated data play a crucial role in their success; Third, datasets also further push the field towards increasingly complicated and challenging problems. Therefore, in this section, we first review the existing commonly used datasets for VSA with motivations, statistics, highlights, and the download links, then introduce evaluation metrics of different tasks, and finally discuss the future trends in audio-visual datasets.

3.1 Datasets

There are dozens of commonly used audio-visual datasets built for VSA. The statistics, highlights and download links are summarized in Table 1, and some selected sample images are shown in Fig. 5. We divide these datasets into two types: controlled and uncontrolled environments. We introduce them briefly in the following.

3.1.1 *Datasets under controlled environments*

As we can see from Table 1, before 2015, visual speech research mainly focused on controlled environments. Controllable factors include recording conditions, equipment, data types, scripts, *etc*. These datasets provide an excellent foundation for visual speech research. Next, we review some representative audio-visual datasets collected under controlled environments.

AVICAR [75] is the most representative public audio-visual dataset recorded in a car-driving environment. As mentioned above, visual speech can contribute to audio-based speech recognition, especially in noisy environments. Motivated by this, AVICAR is collected for modeling bimodal speech in a driving car, as car-driving is a typical acoustic noisy environment.

GRID [77], consisting of high-quality audio and video recordings of 1,000 syntactically identical phrases spoken by 34 talkers, is built for comprehensive audio-visual perceptual analysis and microscopic modeling. Besides speech recognition, it can also support audio-visual speech separation tasks.

MODALITY [79] contains 31 hours of recordings was created to test the robustness of audio-visual speech recognition (AVSR) systems. As for the difference from other datasets, its corpus includes high-resolution, high-framerate stereoscopic video streams from RGB-D cameras.

OuluVS2 [81] is a multi-view audio-visual dataset built for non-rigid mouth motion analysis. It includes 53 speakers uttering three types of utterances. Moreover, it is recorded from five different views spanned between the frontal and profile views. Multiple views of talking mouths simulate a real-world

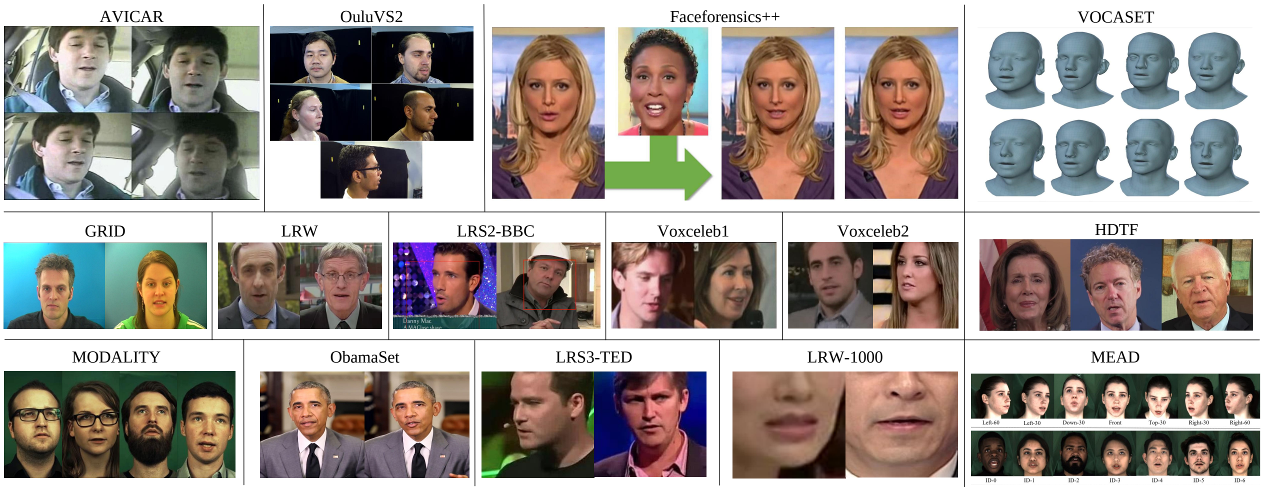

Fig. 5: Some example images from AVICAR, OuluVS2, Faceforensics++, GRID, LRW, LRS2-BBC, VoxCeleb1, VoxCeleb2, MODALITY, ObamaSet, LRS3-TED, LRW-1000, VOCASET, HDTF, MEAD. See Table. 1 for a summary of these datasets.

situation, as users may not face the video camera all the time while talking.

IBM AV-ASR [83] is a large corpus containing 40 hours of audio-visual recordings from 262 speakers in clean, studio conditions. Compared to previous datasets under controlled environments, it has significant advantages in vocabulary and speaker number. However, this dataset is not publicly available.

VOCASET [69] is a 4D face dataset with about 29 minutes of 4D face scans with synchronized audio from 12 speakers (6 females and 6 males), and the 4D face scans are recorded at 60fps. As a representative high-quality 4D face audio-visual dataset, VOCASET greatly promoted the research on 3D VSG.

MEAD [97], namely Multi-view Emotional Audio-visual Dataset, is a large-scale, high-quality emotional audio-visual dataset. Unlike previous datasets, it focuses on natural emotional talking face generation and takes multiple emotion states (eight different emotions at three intensity levels) into consideration.

3.1.2 Datasets under uncontrolled environments

Recently, researchers are gradually shifting their focus to in-the-wild visual speech learning. As a result, many large-scale in-the-wild audio-visual datasets are constructed to promote the research. We introduce some of the audio-visual datasets collected under in-the-wild environments in the following.

LRW [27] is a word-level audio-visual dataset constructed by a multi-stage data automatic collection pipeline. It revolutionary enlarged the dataset scale and speaker number based on the rich data volume of BBC television programs. It contains over 1,000k word instances spoken by over a thousand people. The main objective of LRW is to test speaker-independent word-level lip reading methods.

LRS2-BBC [6] is a sentence-level audio-visual dataset with a similar data collection pipeline and data source as that LRW dataset. It is built for sentence-level lip reading, a more challenging VSR problem than word-level lip reading. All videos in LRS2-BBC are collected from the BBC program, and it contains over 144.5k utterances with a vocabulary size of about 62.8k.

VoxCeleb1 [30] is a large-scale text-independent audio-visual dataset collected from open-source YouTube media. It contains over 100k utterances from 1,251 celebrities. Although it is mainly built for speaker identification, it also can be used for VSG.

ObamaSet [87] is a specific audio-visual dataset focused on the visual speech analysis of former US President Barack Obama. All video samples are collected from his weekly address footage. Unlike previous datasets, it focuses on Barack Obama only and does not provide any human annotations. Therefore, it is only used for Obama-oriented VSG.

LRS3-TED [89] is a large-scale sentence-level audio-visual dataset. Compared to LRS2-TED, it has a larger scale in terms of duration, vocabulary, and number of speakers. It consists of talking face videos from over 400 hours of TED and TEDx videos, the corresponding subtitles, and word alignment boundaries. Besides, it is the largest among existing public available annotated English audio-visual datasets.

VoxCeleb2 [29] is a high-level version of VoxCeleb1 extended on ethnic diversity. In VoxCeleb2, the VoxCeleb1 dataset is re-purposed to serve as a test set for speaker verification. Furthermore, it is currently the largest public available audio-visual dataset.

LSVSR [92] is the largest existing visual speech recognition dataset, consisting of pairs of text and video clips of faces speaking (3,886 hours of video). It is collected from public YouTube videos. But unfortunately, this dataset is not publicly available due to the restricted license.

LRW-1000 [31] is the largest word-level Chinese Mandarin audio-visual dataset. It contains 1k word classes with 718,018 samples from more than 2k individual speakers. Each class corresponds to the syllables of a Mandarin word composed of one or several Chinese characters. All videos are collected from television programs on Chinese TV stations.

Faceforensics++ [94] is an automated benchmark for facial manipulation detection. Different from existing audio-visual datasets, all videos have been manipulated based on DeepFakes [102], Face2Face [103], FaceSwap [104], NeuralTextures [105] as main methods for facial manipulations. It is commonly used to test forgery video detection methods.

HDTF [68] is a large-scale in-the-wild audio-visual dataset built for talking face generation. It consists of about 362 different high-resolution videos collected online. Due to the high quality of origin videos, the cropped face-centered videos also have higher visual quality than that of previous datasets like LRW and LRS2-BBC.

In addition to the datasets listed in table 1, there are several audio-visual datasets recorded in different languages. For example, Spanish language dataset VLRF [106], Australian English dataset [107], Russian language dataset

HAVRUS [108] *etc.* also promoted the research of VSA on various languages.

Considering that datasets play a crucial role in VSA, we would like to give a summary and discussion of datasets to help readers to know the development of VSA. Compared with early audio-visual datasets, recent ones have been improved in the number of subjects, dataset scale, recording conditions and script diversity, data quality, *etc*. Due to the privacy protection laws (*e.g.*, General Data Protection Regulation (GDPR) in Europe Union), some of existing large-scale datasets [83, 92] are not public available. An intuitive solution is to automatically collect available data from online media (*e.g.*, Youtube, BBC, or other online television programs). However, existing audio-visual data auto-collection algorithms may cause a large amount of low-quality data. Therefore, an optimized auto-collection algorithm is crucial for VSA datasets in the future.

3.2 Evaluation Metrics

3.2.1 Evaluation Metrics on VSR

The word-level VSR task is essentially a multi-class classification problem. Therefore, classification accuracy is the most common evaluation metric for classification models because of its simplicity and efficiency. Besides, $Top-k$ accuracy, namely the standard accuracy of the actual class being equal to any of the k most probable classes predicted by the classification model, is also widely used in VSR.

As for the sentence-level task, Character Error Rate (CER) and Word Error Rate (WER) [109], also known as average character-level and word-level edit distances, are the most commonly used evaluation metrics. CER is defined as $\mathrm{CER} = (S + D + I)/N$, where S, D and I are the numbers of substitutions, deletions, and insertions respectively to get from the reference to the hypothesis, and N is the number of characters in the reference. This metric imposes smaller penalties where the predicted string is similar to the ground truth. For example, if the ground truth is "*about*" and the model prediction is "*above*", then $\mathrm{CER} = 0.4$. WER and CER are calculated in the same way. The difference lies in whether the formula is applied to character-level or word-level. Besides, BLEU [110], a modified form of n-gram precision to compare a candidate sentence to one or more reference sentences, is sometimes adopted.

3.2.2 Evaluation Metrics on VSG

Appropriate evaluation for VSG continues to be an open problem, and many recent works have explored various evaluation metrics on VSG. We categorize those metrics based on three learning targets, *i.e.*, identity preservation, visual quality, audio-visual semantic consistency.

Identity Preservation. One of the most important goals of VSG is to preserve the target identity as much as possible during video generation, as humans are quite sensitive to subtle appearance changes in synthesized videos. Since identity is a semantic concept, direct evaluation is not feasible. To evaluate how well the generated video preserves the target identity, existing works usually use the embedding distance of the generated video frames and the ground truth image to measure the identity-preserving performance. For example, Vougioukas *et al.* [40] adopted the average content distance (ACD) [111] to measure the average Euclidean distance of target image representation, obtained using OpenFace [112], and the representations of generated frames. Besides, Zakharov *et al.* [113] used the cosine similarity between embedding vectors of the ArcFace network [114] for measuring identity mismatch.

Visual Quality. To evaluate the quality of the synthesized video frames, reconstruction error measurement (*e.g.*, Mean Squared Error) is a natural evaluation way. However, reconstruction error only focuses on pixel-wise alignments and ignores global visual quality. Therefore, existing works usually adopt Peak Signal-to-Noise Ratio (PSNR) and Structure Similarity Index Measure (SSIM) to evaluate the global visual quality of generated frames. More recently, Prajwal *et al.* [38] introduced Fréchet Inception Distance (FID) to measure the distance between synthetic and real data distributions, as FID is more consistent with human perception evaluation. Besides, Cumulative Probability Blur Detection (CPBD) [115], a non-reference measure, is also widely used to evaluate the loss of sharpness during video generation.

Audio-visual Semantic Consistency. Semantic consistency of the generated video and the driving source mainly contains audio-visual synchronization and speech consistency. For, audio-visual synchronization, Landmark Distance (LMD) [116] computes the Euclidean distance of the lip region landmarks between the synthesized video frames and ground truth frames. The other synchronization evaluation metric is to use a pre-trained audio-to-video synchronisation network [48] to predict the offset of generated frames and the ground truth. For the speech consistency, Chen *et al.* [42] proposed a lip-synchronization evaluation metric, *i.e.*, Lip-Reading Similarity Distance (LRSD), which measures the Euclidean distance of semantic-level speech embeddings obtained by lip reading networks. For better evaluation of speech consistency, lip reading results (accuracy, CER, or WER) comparisons of the generated frames and ground truth are also used as consistency evaluation metrics.

In addition to the above objective metrics, subjective metrics like user study are also widely used in VSG.

4 Visual Speech Recognition

4.1 The Overall Framework

Visual Speech Recognition (VSR), also known as lip reading, aims to decode speech from speakers' mouth movements. An essential preprocessing of VSR is mouth-centered region of interest (ROI) cropping. A talking face video contains a large amount of redundant information (such as pose, illumination, gender, skin color, *etc.*) unrelated to the VSR task. To reduce redundant information, it is necessary to crop mouth-centered videos from the raw input video. However, defining the size of mouth-centered ROI is still an open problem. Koumparoulis *et al.* [117] proved that the selection of ROI will significantly affect the final recognition performance, but it is still unable to determine the optimal ROI.

As shown in Fig. 6, a VSR system usually contains three sub-modules. The first sub-module is visual feature extraction, intending to extract compact and effective visual feature vectors from mouth-cropped videos. Then, the second sub-module is temporal context aggregation, aiming to aggregate temporal context information for better text script decoding and recognition. The above two sub-modules are also the cores of deep learning based VSR methods. This paper will summarize and discuss existing deep networks for visual feature extraction and temporal context aggregation in Section. 4.2. The last sub-module is text decoding, *i.e.*, converting the feature representations to text.

The rest of this section is organized as follows. Section. 4.2 presents our taxonomy of deep representation learning networks

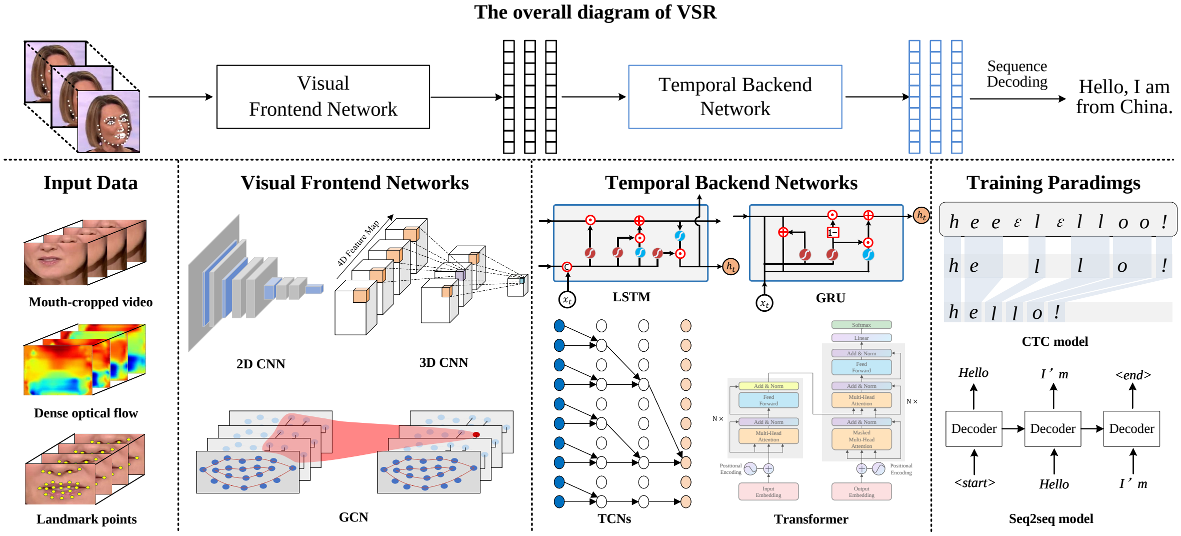

Fig. 6: The overall diagram of VSR and various visual frontend networks, temporal backbone networks, and trainging paradigms.

for VSR. And then, we review and discuss various visual speech representation learning paradigms for VSR (supervised learning and unsupervised learning) in Section. 4.3. Section. 4.4 provides a comprehensive summary for readers to know the progress and limitations of existing VSR methods.

4.2 Backbone Architectures

Before the era of deep learning, representation learning for VSR had already been explored for a long time. From the feature engineering perspective, traditional feature extraction methods can be categorized into three types: appearance-based, shape-based, and motion-based [8]. Although simple and explainable, traditional representation learning methods usually do not work well, especially in uncontrolled environments. This paper mainly focuses on summarizing and discussing representation learning methods driven by deep learning technologies. Considering the significant difference between deep representation learning and traditional feature extraction, we introduce a novel taxonomy strategy based on two independent parts: visual frontend network and temporal backend network.

4.2.1 Visual frontend network

As shown in Fig. 6, there are mainly three types of input data: mouth-centered videos, dense optical flow, and landmark points. Among them, mouth-centered videos and dense optical flow are regular grid data, so CNNs are the most suitable and commonly used backbone architectures for them. On the other hand, as landmark points are irregular data, some existing works [53, 54, 118] adopted Graph Convolution Networks (GCNs) to extract visual features from landmark points. Next, we review these backbone architectures.

CNN-based Architectures. CNNs have been becoming one of the most common architectures in the field of deep learning. Since AlexNet [119] was proposed in 2012, researchers have invented a variety of deeper, wider, and lighter CNN models [120]. Representative CNN architectures, such as VGG [121], ResNet [122], MobileNet [123], DenseNet [124], ShuffleNet [125] *etc*, have been widely used in learning visual representation for VSR.

The first end-to-end deep visual representation learning for word-level VSR was proposed by Chung *et al*. [27]. Based on the VGG-M backbone network, they compared different image sequence input (Multiple Towers *vs.* Early Fusion) and temporal fusion (2D CNNs *vs.* 3D CNNs) architectures and discussed their pros and cons. The experimental results showed that the 2D CNNs are superior to their 3D counterparts by a large margin. However, the above conclusion was not rigorous enough, as the ablation study is insufficient, and word-level VSR datasets have a very short-term dependency. In 2017, Assael *et al*. [56] proposed LipNet, the first end-to-end sentence-level VSR model. LipNet extracts visual features using a 3-layer Spatio-temporal Convolution Neural Network (STCNN, also known as 3D CNN). The experimental results confirm the intuition that extracting spatio-temporal features using STCNN is better than aggregating spatial-only features. Considering 3D CNNs are more capable of capturing the dynamics of the mouth region while 2D CNNs are more efficient in time and memory, Stafylakis *et al*. [55] proposed to combine 3D CNNs and 2D CNNs for visual feature extraction. In specific, the proposed visual backbone network consists of a shadow 3D CNN and 2D ResNet. The 3D CNN has just one layer to aggregate short-term temporal information on lip movements. Due to the considerable performance of the model, plenty of VSR models [50, 52, 59, 61, 126] adopted it as the backbone network for visual features extraction. Recently, Feng *et al*. [127] improved this architecture by integrating the Squeeze-and-Extract [128] module. Besides VGG and ResNet, researchers have also adopted other representative 2D CNN architectures, including DenseNet [58], ShuffleNet [52], MobileNet [129], *etc*.

GCN-based Architectures. Considering CNNs are not suitable for irregular grid data, researchers proposed to utilize Graph Convolution Networks (GCNs) to extract visual features from the facial landmark points [130]. Liu *et al*. [53] proposed the first end-to-end GCN model (ST-GCN) that extracts shape-based visual features by learning the lip landmark points and their relationships. They firstly proposed lip graph connection relations and defined the graph adjacency matrices based on the manifold distance of nodes. Then, they combined the image features and shape features to extract more discriminative visual features. However, the lip graph connection relations do not naturally exist, and the intuition-guided predefined lip graph restricts the representation ability of shape-based features. Motivated by this, Sheng *et al*. [54] proposed an Adaptive Semantic-Spatial-Temporal Graph Convolution Network (ASST-GCN). Unlike [53], the ASST-GCN parameterizes

TABLE 2: The pros and cons of various visual frontend network architectures and temporal backend network architectures.

Architectures	Available input	Pros	Cons
2D CNNs	mouth video	high memory/time efficiency	poor at capturing temporal correlation
3D CNNs	mouth video / optical flow	powerful at short-term spatio-temporal modeling	high memory/time cost
3D + 2D CNNs	mouth video / optical flow	high memory/time efficiency; strong discrimination	not good at capturing subtle lip dynamics
Visual transformers	mouth video	high robustness; strong discrimination	high memory/time cost
GCNs	lip landmark points	high computation efficiency; semantic preserving	low accuracy ; low robustness
RNNs	visual features	relatively good generalization	short-term dependency; serial computing
Transformers	visual features	long-term dependency; parallel computation	overfitting on small datasets; hard to converage
TCNs	visual features	adaptive to multi-scale patterns; high memory efficiency	short-term dependency

graph connections and automatically learns adaptive graph connections. Besides, they introduced two graph structures, *i.e.*, semantic graph and spatial-temporal graph, making graph parameters can be adaptively learned with other parameters in the network training. Existing works show that image-based features are more discriminative than landmark-based features. Sheng *et al.* [54] concluded the reason for this. The accuracy and coordinates resolution of landmark point detection significantly influence its feature discrimination. However, facial landmark detection is challenging, especially in uncontrolled environments. Since the complementarity between image and landmark features, the combination of CNNs and GCNs is often widely adopted [53, 54, 118].

Visual Transformer-based Architectures. Inspired by the significant success of transformer architectures in the field of NLP, researchers have recently applied transformers to computer vision (CV) tasks [131]. Recently, transformers have been showing they are potential alternatives to CNNs. Afouras *et al.* [57] designed an end-to-end visual transformer-based pooling mechanism that learns to track and aggregate the lip movement representations. The proposed visual backbone network can reduce the need for complicated preprocessing, improving the robustness of visual representation. The ablation study clearly shows that the visual transformer-based pooling mechanism significantly boosts the performance of VSR.

Based on the above backbone architectures, some works further improved visual representation by utilizing two-stream networks. For example, Weng *et al.* [132] successfully migrated the two-stream (the raw grayscale video stream and the dense optical flow stream) I3D model to VSR, and achieve comparable performance on word-level VSR. However, dense optical flow and 3D convolution calculation is very time consuming, resulting in low feature extraction efficiency. Wang *et al.* [58] utilized 2D CNNs and 3D CNNs to extract both frame-wise spatial features and short-term spatio-temporal features, and then fused the features with an adaptive mask to obtain strong, multi-grained visual features.

4.2.2 Temporal backend network

The temporal backend network built upon visual features aims to further aggregate context information. In traditional VSR, classical statistical models (*e.g.*, Hidden Markov Model, HMM) are commonly used for temporal information aggregation.

RNN-based Architectures. In the field of deep learning, Recurrent Neural Networks (RNNs) are representative network structures used to learn sequence data. The typical RNN structures (*e.g.*, LSTM and GRU) are shown in Fig. 6, and their basic structures are similar to that of HMM, in which the dependencies between the observed state sequences are described by the transformation of the hidden state sequence. Compared to HMM, RNNs have a more powerful representation ability due to the nonlinear transformation during hidden state transitions. Bidirectional RNNs (BiRNNs) are variations of basic RNNs, which attempt to aggregate context information from previous timesteps as well as future timesteps. Many works [28, 50, 55, 56, 127, 132, 133, 134, 135] have adopted RNN-based (BiLSTM or BiGRU) network architectures as the temporal backend network in VSR. Beyond above fundamental RNN structures, various modifications [58, 136] have been made to improve feature learning for VSR. For example, Wang *et al.* [58] utilized BiConvLSTM [137] as temporal backend network. ConvLSTM is a convolutional counterpart of conventional fully connected LSTM, which models temporal dependency while preserving spatial information. Wang *et al.*integrated the attention mechanism into the model to further improve the BiConvLSTM architecture.

Transformer-based Architectures. Compared to RNN-based architectures, Transformers [138] have significant advantages in long-term dependency and parallel computation. However, transformers usually suffer from some drawbacks. First, transformers are more prone to overfitting than RNNs and TCNs in small-scale datasets. Second, transformers are limited in some specific tasks (*e.g.*, word-level VSR tasks) with short-term context. Therefore, transformers are more suitable for sentence-level VSR tasks, rather than word-level VSR tasks. [6] is the first work introducing transformers to VSR. Based on the basic transformer architecture, the authors proposed two types of backend models: TM-seq2seq and TM-CTC. The difference between the two models lies in the training target. The experiments clearly showed that the transformer-based backend network performs much better than the RNN-based backend network in the sentence-level VSR task. Since the basic transformer pays no extra attention to short-term dependency, Zhang *et al.* [61] proposed multiple Temporal Focal blocks (TF-blocks), helping features to look around their neighbors and capturing more short-term temporal dependencies. The results demonstrated that the short-term dependency is as crucial as the long-term dependency in sentence-level VSR.

TCN-based Architectures. In the context of deep sequence models, RNNs and Transformers have a high demand for memory and computation ability. Temporal Convolutional Networks (TCNs) are another type of deep sequence model, and various improvements have been applied to the basic TCN to make them more appropriate for VSR. For example, Afouras *et al.* [60] used depth-wise separable convolution (DS-TCN) for sentence-level VSR. However, the performance of DS-TCN does not work as well as transformers, as TCN-based models have poor ability on capturing long-term dependency. To enable temporal backend network capturing multi-scale temporal patterns, Martinez [59] proposed to utilize multi-scale TCN (MS-TCN) structure, which achieved SOTA results (87.9% Acc) on the word-level LRW dataset.

Table. 2 summarizes the general pros and cons of various visual frontend networks and temporal backend networks, and the available inputs of the corresponding visual frontend network. As we know, most of the existing VSR models are derived from general backbone models used in other fields (*e.g.*, action recognition [139, 140], audio speech recognition [137], *etc.*), and few are designed explicitly for VSR. Therefore, more attention should be paid to the particular structure adaptive to the properties of VSR in the future.

4.3 Learning Paradigms

4.3.1 Supervised learning for VSR

There are two mainstream VSR tasks: word-level and sentence-level. With a limited number of word categories, the former is to recognize isolated words from the input videos (*i.e.*, talking face video classification), usually trained with multi-classification cross-entropy loss. The latter is to make unconstrained sentence-level sequence prediction. However, due to the unconstrained word categories and video frame length, it is much more complicated than the word-level VSR task.

Supervised learning of end-to-end sentence-level VSR tasks (sentence prediction) can be divided into two types. Given the input sequence, the first type uses a neural network as an emission model, which outputs the likelihood of each output symbol (*e.g.*, phonemes, characters, words). These methods generally employ a second phase of decoding using HMM. A popular version of this variant is the Contortionist Temporal Classification (CTC) [141], where the model predicts frame-wise labels and then looks for the optimal alignment between the frame-wise predictions and the output sequence. The main weakness of CTC is that the output labels are not conditioned on each other (it assumes each unit is conditional independent), and hence a language model is needed as a post-processing step. Different from the basic CTC, Xu *et al.* [62] proposed LCANet that feeds the encoded spatio-temporal features into a cascaded attention CTC decoder. The introduction of an attention mechanism improves the defect of the conditional independence assumption CTC in hidden neural layers. Another assumption of this approach is that it assumes a monotonic ordering between input and output sequences, which is suitable for VSR but not for machine translation.

The second type is sequence-to-sequence (seq2seq) models that first read the whole input sequence before predicting the output sentence. A number of works adopted this approach for speech recognition [142]. Chan *et al.* [143] proposed an elegant seq2seq method to transcribe audio signal to characters. Seq2seq models decode an output symbol at time t (*e.g.*, phonemes, characters, words) conditioned on previous outputs $1, ..., t-1$. Thus, unlike CTC-based models, the model implicitly learns a language model over output symbols, and no further processing is required. However, it has been shown [144] that it is beneficial to incorporate an external language model in the decoding of seq2seq models as well. Chung *et al.* [28] proposed the WAS (Watch, Attend and Spell) model, which is a classical seq2seq VSR model. With the help of attention mechanism, WAS model is more capable of capturing long-term dependency.

Based on the transformer backbone architecture, Afouras *et al.* [6] have deeply analyzed the pros and cons of the CTC model and the seq2seq model for VSR. Generally, the seq2seq model performs well than the CTC model in the sentence-level VSR task. But, the seq2seq model needs more training time and inference time. Besides, the CTC model generalizes better and adapts faster as the sequence lengths are increased.

Besides the above label-level supervised learning paradigms, feature-level supervised learning is also widely explored in VSR. Knowledge distillation [145] (KD) technology is the key to feature-level supervised learning. For example, Ma *et al.* [52] proposed a multi-stage self-KD training framework for the word-level VSR task. Like label smoothing, KD can provide an extra supervisory signal with inter-class similarity information. Some works [51, 63, 146] utilized cross-modal KD to train a robust VSR model by distilling from a well-trained ASR model. With the help of the ASR model, KD technology can significantly speed up the training of the VSR model. Meanwhile, combining CTC loss and KD loss can further improve the performance of VSR.

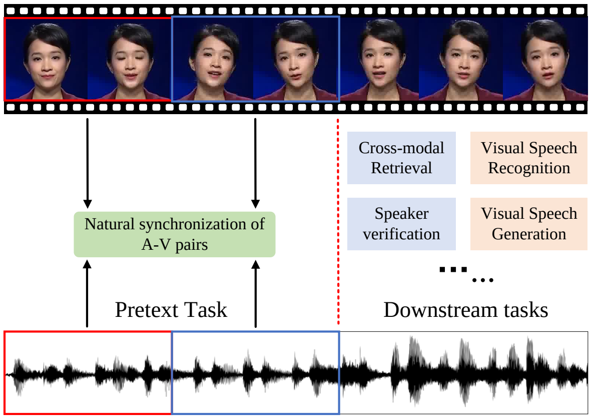

Fig. 7: The general motivation and available downstream tasks of self-supervised learning on visual speech.

4.3.2 Unsupervised learning for VSR

Unsupervised learning for VSR aims to learn discriminative visual representations without access to manual annotation. Among multiple unsupervised learning frameworks, cross-modal self-supervised learning is dominated in VSR. Despite the remarkable progress witnessed in the past decade, the successes of supervised deep learning rely heavily on vast manually annotated training data, which has severe limitations in many real-world applications, including the VSR task. Firstly, supervised learning is restricted to relatively narrow domains primarily defined by the labeled training data and thus leads to limited generalization ability. Secondly, a large amount of accurately labeled data like a large-scale annotated dataset for VSR is costly to gather. Recently, self-supervised learning has received growing attention due to its high label efficiency and good generalization.

Recent advances in cross-modal self-supervised learning have shown that the corresponding audio can serve as a supervisory signal to learn effective visual representations for VSR. As shown in Fig. 7, audio-visual self-supervised learning aims to extract efficient representations from the co-occurring A-V data pairs without any extra annotation. Based on the natural synchronization property of audio and video, existing methods mainly adopt contrastive learning to achieve this goal. Chung *et al.* [149] are the first to train an A-V synchronization model in an end-to-end manner with margin-based [150] pairwise contrastive loss. Besides VSR, they have proved that the trained network work can effectively be finetuned to other tasks like speaker detection. With the same training strategy, Korbar *et al.* [151] broadened the scope of the study to encompass arbitrary human activities rather than lip movements. Except for margin-based loss, L1 loss and binary classification loss [47, 152, 153, 154] are also widely used for A-V representations learning. Those works have proved the learned A-V representations can be further transferred to more downstream tasks, such as visualizing the locations of sound sources, action recognition,

TABLE 3: The comparison of representative VSR methods.

Task type	Method	Frontend network	Backend network	Experimental settings				Performance (Dataset)	Highlights
				Learning paradigm	Extra datasets	Extra LM	output symbol		
Word-level	Chung *et al*. [27]	VGG-M	/	/	/	/	word	61.1% (LRW) 25.7%(LRW-1000)	Discussed various temporal fusion ways for word-level VSR networks
	Stafylakis *et al*. [55]	C3D-ResNet34	BiLSTM	/	/	/	word	83.5% (LRW) 38.2% (LRW-1000)	Proposed the most widely used visual frontend network, *i.e.*, C3D_ResNet
	Wang *et al*. [58]	ResNet34 3D-DenseNet52	BiConvLSTM	/	/	/	word	83.3% (LRW) 36.9% (LRW-1000)	Introduced the BiConvLSTM architecture for VSR
	Liu *et al*. [53]	C3D-ResNet34 ST-GCN	BiGRU	/	/	/	word	84.25% (LRW)	Firstly utilized the GCN-based network in VSR
	Martinez *et al*. [59]	C3D-ResNet18	MS-TCN	/	/	/	word	85.3% (LRW) 41.4% (LRW-1000)	Improved performance by multi-scale TCN; Adaptive to varying input length
	Sheng *et al*. [54]	C3D-ResNet18 ASST-GCN	MSTCN	/	/	/	word	85.7% (LRW)	Introduced lip semantic encoding; No need for predefined lip graph
	Ma *et al*. [52]	C3D-ResNet18	MS-TCN	Multi-stage KD	/	/	word	**87.7%** (LRW) 43.2% (LRW-1000)	Improved generalization with the help of KD and achieved SOTA results on LRW
	Feng *et al*. [127]	SE-C3D-ResNet18	BiGRU	/	/	/	word	85.0% (LRW) **48.0%** (LRW-1000)	Introduced Squeeze-and-Extract module; Achieved SOTA results on LRW-1000
	Yang *et al*. [34]	C3D-ResNet18	ResNet18	Cross-modal mutual learning	/	/	word	**88.5**%† (LRW) **50.5**%† (LRW-1000)	Proposed a unified framework for audio-visual speech recognition and synthesis
Sentence-level	Assael *et al*. [56]	ST-CNN	BiGRU	CTC loss	/	√	character	1.9% CER (GRID) 4.8% WER (GRID)	The first end-to-end sentence-level VSR model
	Xu *et al*. [62]	C3D + HighwayNet	BiGRU	CTC loss	/	/	character	1.3% CER (GRID) 2.9% WER (GRID)	Compensated the defect of the CTC approach
	Afouras *et al*. [6]	C3D-ResNet18	Transformer	CTC loss	LRW, MVLRS, LRS2	√	character	54.7% CER (LRS2) 66.3% WER (LRS3)	Deeply analyzed the pros and cons of the CTC model and the seq2seq model
				seq2seq loss				48.3% CER (LRS2) 58.9% WER (LRS3)	
	Shillingford *et al*. [92]	ST-CNN	BiLSTM	CTC loss	LSVSR	√	phoneme	28.3% CER (LSVSR) 40.9% WER (LSVSR) 55.1 WER (LRS3)	Adopted phoneme as the output symbol and proposed the largest dataset LSVSR
	Zhang *et al*. [61]	C3D-ResNet18	TF-blocks	seq2seq loss	LRW, LRS2	/	character	1.3% WER (GRID) 51.7% WER (LRS2) 60.1% WER (LRS3)	Integrated causal convolution into transformer for VSR
	Makino *et al*. [147]	ST-CNN	RNN-T (BiLSTM)	seq2seq loss	YT-31khrs	√	character	33.6% WER (LRS3)	Proposed an RNN-T based VSR system and collectd a large dataset from YouTube
	Ma *et al*. [148]	C3D-ResNet18	Conformer + Transformer	CTC loss + seq2seq loss	LRW, LRS2, LRS3	√	character	37.9% WER (LRS2) 43.3% WER (LRS3)	Proposed a hybrid CTC/Attention model and achieved SOTA results on LRS3
	Afouras *et al*. [57]	3DCNN+VTP	Transformer	seq2seq loss	LRS2, LRS3, MVLRS, TEDx	√	sub-word	**22.6%** WER (LRS2) **30.7%** WER (LRS3)	Introduced sub-word as the output symbol and replaced the 2DCNN with the visual transformer

†: audio data is used in the training.

audio-visual source separation, *etc*Recently, Chung *et al*. [101] reformulated the contrastive task as a multi-way matching task and demonstrated that using multiple negative samples can improve the performance. Considering existing methods only exploit the natural synchronization of the video and the corresponding audio, Sheng *et al*. [50] proposed a novel self-supervised learning framework called Adversarial Dual-Contrast Self-Supervised Learning (ADC-SSL), to go beyond previous methods by explicitly forcing the visual representations disentangled from speech-unrelated information. To achieve this goal, they combine contrastive learning and adversarial training by three pretext tasks: A-V synchronization, identity discrimination, and modality classification.

4.4 Summary and performance comparison

We have witnessed significant progress in various aspects of visual speech recognition. In this subsection, we will compare existing VSR methods on representative datasets and summarize the main issues of VSR.

4.4.1 Performance Comparison

In this section, we compare the existing deep learning-based VSR methods. Due to a large number of methods proposed for VSR, it is not possible to list and compare all of them. Thus, we select representative works and several milestone methods. Table. 3 summarizes the performance and some experimental settings of some representative VSR methods on large-scale commonly used benchmark datasets, including LRW [27], LRW-1000 [31], GRID [77], LRS2 [6] and LRS3 [89].

As for the word-level VSR task, various visual frontend networks have been designed to boost the performance, such as VGG-M, C3D-ResNet, ST-GCN, ASST-GCN,*etc*. Among them, the C3D-ResNet architecture is the most widely used. [55] provided the baseline (C3D-ResNet34 + BiLSTM, 83.5%) on the LRW dataset. Many subsequent works inherited this structure and further improved the performance by introducing some tricks, such as label smoothing, weight decay, dropout, Squeeze-and-Extract module, two-stream, multi-stage KD, *etc*. As for temporal backend networks, RNN-based models and TCN-based models have similar performance. Based on C3D-ResNet18 + MSTCN, Ma *et al*. [52] improved the SOTA to 87.7% on LRW. Recently, more and more works [34, 47, 49, 50, 101, 149] tried to improve visual representations by utilizing extra audio information in the training stage rather than the design of network architectures, as audio signals can provide more fine-grained supervision than text annotations. The SOTA results (88.5% on LRW and 50.5% on LRW-1000) was realized based on cross-modal audio-visual mutual learning [34].

As for the sentence-level VSR task, deep learning-based VSR methods have vastly outperformed human lip-readers [56]. As shown in Table. 3, deep VSR models almost reach performance saturation (SOTA result: 1.3% WER) on the simple (constrained recording environment and limited corpus) GRID dataset. Therefore, researchers pay more attention to VSR in unconstrained environments. Motivated by the practical need, we focus more on the large-scale in-the-wild datasets (*e.g.*, LRS2 and LRS3). The fair performance comparison of sentence-level VSR methods is quite hard, as there are too many extra influencing factors. For example, some methods trained the model with extra datasets (even some of them are not public available). Besides, the outputs of the model are generally optimized by extra language models, while language models are trained with existing large-scale text corpus. The introduction of language models can significantly improve the performance, and it is not fair to compare these methods optimized by different language models. Therefore, to make it clearer for readers, we list some representative sentence-level VSR models as well as their experimental settings in Table. 3.

4.4.2 Main issues and facts

Over the last decade, deep learning-based VSR techniques have been significantly developed. However, some issues remain to be solved. We conclude them as follows:

- The cropping preprocessing of raw talking face videos have a significant impact on the recognition results, and how to define the optimal lip ROI for the VSR task is worthy of further exploration.

- In practical applications, real-time is another substantial demand for VSR. However, most existing VSR methods only focus on recognition accuracy but ignore real-time. Therefore, the trade-off between accuracy and real-time should be considered in the future.
- There is no formal robustness analysis of existing VSR methods. As we have mentioned in Section. 2.2.1, VSR faces many challenges, such as speaker differences and unconstrained environments. Existing deep learning-based VSR networks are rarely targeted to solve these problems. Therefore, robustness analysis of VSR methods needs more attention in the future.
- Another serious problem of VSR research is the lack of fair benchmarks for algorithm comparison, especially for the sentence-level VSR task. The performance of VSR is affected by many factors, such as extra language models, multiple training datasets, audio signals, and implementation details. Due to the lack of a unified test platform, a fair comparison of VSR algorithms is not easy to achieve.

5 Visual Speech Generation

Visual Speech Generation (VSG), also known as lip sequence generation, aims to synthesize a lip sequence corresponding to the driving source (a clip of audio or a piece of text).

Traditional VSG methods suffer from severe practical challenges [45], such as complex generation pipelines, constrained applicable environments, over-reliance on fine-grained viseme (phoneme) annotations, *etc*. To realize mapping driving sources to lip dynamics, representative traditional VSG methods mainly adopted cross-modal retrieval approaches [16, 103, 155, 156] and HMM-based approaches [157, 158]. For example, Thies *et al.* [103] introduced a typical image-based mouth synthesis approach that generates a realistic mouth interior by retrieving and warping best-matching mouth shapes from offline samples. However, retrieval-based methods are static text-phoneme-viseme mappings and do not really consider the contextual information of the speech. Meanwhile, retrieval-based methods are pretty sensitive to head pose changes. HMM-based methods also suffer from some drawbacks, such as the limitation of the prior assumptions (*e.g.*, Gaussian Mixture Model (GMM) and its diagonal covariance). As deep learning technologies have extensively promoted the developments of VSG, we focus on reviewing deep learning based VSG methods in this section.

To make the scope of VSG clear for readers, we first explain the relationship and difference between VSG and another hot topic, *i.e.*, Talking Face Generation (TFG) [1] [71, 159].

TFG aims to synthesize a realistic, high-quality talking face video corresponding to the driving source and the target identity. According to the modality of driving sources, TFG can be divided into audio-driven, text-driven, and video-driven. Among them, video-driven TFG mainly focuses on video-oriented face-to-face facial expression transferring rather than visual speech generation. Therefore, video-driven TFG methods will not appear in this paper.

Traditionally, VSG can be viewed as a key sub-component of text-driven (audio-driven) TFG. The other component is video editing, following a specific editing pipeline to output the final synthesized talking face video based on the generated lip sequence. Recently, to reduce manual intervention and simplify the complexity of the overall pipeline, more and more researchers have tried to synthesize full talking face in an end-to-end manner instead of lip sequence. Consequently, the definition boundary between VSG and text-driven (audio-driven) TFG is getting blurred, which means some text-driven (audio-driven) TFG methods are also in our review scope. Therefore, to give a comprehensive survey on VSG, we also review some TFG methods driven by text and audio, as these works implicitly or explicitly involve VSG modules.

5.1 The Overall Pipeline

Given a reference identity (an image or a 3D facial model of the target speaker) and a driving source (a piece of audio or text), the objective of VSG is to generate the final synthesized talking lip (face) videos. Existing VSG approaches have various properties, such as input modalities (text-driven or audio-driven), synthesizing strategies (computer graphics based, image reconstruction based, or hybrid based), speaker generalization (speaker-independent or speaker-dependent), learning paradigms (supervised learning or unsupervised learning), classifying these approaches is not an easy task.

This section provides a novel taxonomy for VSG methods, as shown in Fig. 8(a). In specific, we organize VSG approaches into two frameworks: a) Two-stage frameworks, which include two mapping steps, *i.e.*, driving source to facial parameters and facial parameters to videos; and b) One-stage (Unified) frameworks, having single generation process which is intermediate facial parameters free. Next, we review and analyze current two-stage and one-stage VSG methods as well as their advantages and disadvantages in detail in Section. 5.2 and 5.3, respectively.

5.2 Two-Stage VSG Framework

The two-stage VSG frameworks mainly consist of two steps: a) mapping the driving source to facial parameters using DNNs and b) transforming the learned facial parameters to output videos based on GPU rendering, video editing, or Generative Adversarial Networks (GANs) [37]. According to the data type of facial parameters, existing two-stage VSG approaches can be divided into Landmarks based, Coefficients based, Vertex based and others.

5.2.1 Landmark based Methods

Facial landmark points around facial components capture the rigid and non-rigid facial deformations due to head movements and facial expressions [160]. Facial landmark points are widely used in various facial analysis tasks, including VSG. As a pioneering work, Suwajanakorn *et al.* [87] adopted a simple single-layer LSTM with the time delay mechanism to learn a nonlinear mapping from audio coefficients to lip landmark points. As shown in Fig. 8(b), the model outputs the synthesized talking face video of former US President Barack Obama, following the pipeline of facial texture synthesis, video re-timing, and target video compositing. Beyond computer graphic video generation methods, as shown in Fig. 8(c), Kumar *et al.* [64] proposed the LSTM + UNet architecture, improving the model by replacing the complex video generation pipeline with a pix2pix framework [161]. In this way, there is no need to get

[1] Talking Face Generation is also called talking face synthesis, talking head generation, or talking portraits generation. These concepts are interchangeable, and to be consistent, the expression “Talking Face Generation (TFG)” is adopted in this paper.

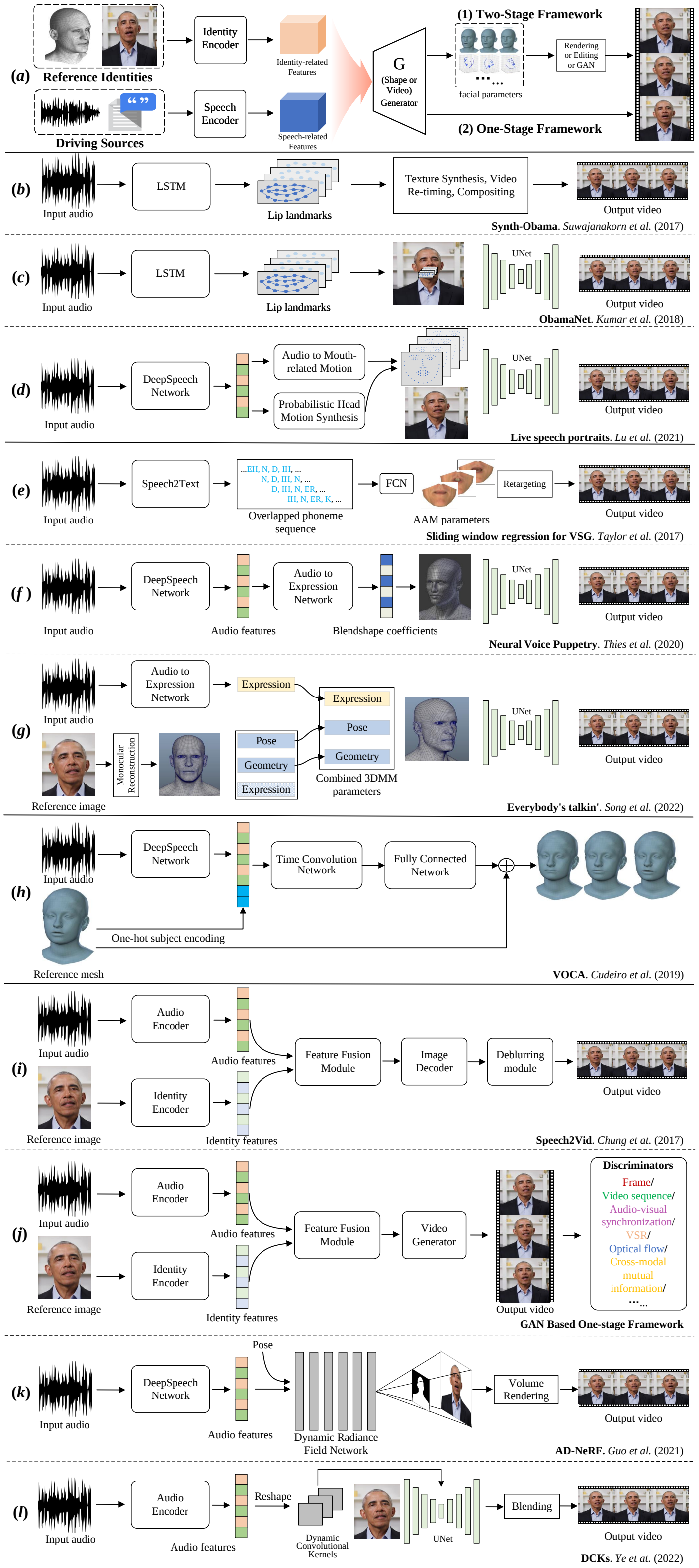

Fig. 8: *(a)*: The overall framework of visual speech generation. *(b)-(h)*: Representative Two-stage VSG methods. *(i)-(l)*: Representative One-stage VSG methods.

involved with details of a face, *e.g.*, synthesizing realistic teeth. However, as the above methods are trained on only Barack Obama with many hours of his weekly address footage, they cannot generalize to new identities or voices. The LSTM + UNet VSG backbone architecture is widely adopted in many subsequent works [15, 97, 162, 163]. Unlike previous methods using audio MFCC features as input, Sinha *et al.* [162, 163] introduced DeepSpeech [164] features instead, as DeepSpeech features are more robust for the variation of speakers.

In 2018, Jalalifar *et al.* [165] proposed LSTM + C-GAN VSG backbone architecture, using the basic conditional generative adversarial network (C-GAN) [166] for generating talking faces given the learned landmarks. As the LSTM network and C-GAN network are mutual independence, this model can reanimate the target face with audio from another person. In 2019, Chen *et al.* [65] proposed a novel LSTM + Convolutional-RNN structure, further considering the correlation between adjacent video frames during generation. Besides, they also propose a novel dynamic pixel-wise Loss to solve the pixel jittering problem in correlated audio-visual regions. Wang *et al.* [167] proposed a three-stage VSG framework. Firstly, they use the 3D Hourglass Network as a motion field generator to predict landmark points based on the input audio, head motions, and the reference image. And then convert the predicted landmark points to dense motion fields. Finally, the synthesized talking video is obtained using a first-order motion model [168]. Recently, they further updated the motion field generator by replacing the 3D Hourglass Network with a self-attention architecture [41].

Besides 2D landmark based approaches, mapping driving source to 3D landmarks is also widely explored. Audio signals contain sich semantic-level information, including speech content, speaker's speaking style, emotion, *etc*. Zhou *et al.* [15] utilized a voice conversion neural network to learn disentangled speech content and identity features. Then, an LSTM-based network is introduced to predict 3D landmarks based on speech content features. Finally, the final synthesized talking face video is realized using a UNet-style generator network. The key insight is to predict 3D landmarks from disentangled audio content features and speaker-aware features, such that they capture controllable lip synchronization and head motion dynamics. As shown in Fig. 8(d), Lu *et al.* [169] introduced extracting the high-level speech information using an autoregressive predictive coding (APC) model [170] and manifold projection for better generalization. Then, an audio to lip-related motion module is designed to predict 3D lip landmarks. Finally, an image-to-image translation network (UNet) is introduced to synthesize video frames.

5.2.2 Coefficient based Methods

2D Coefficient based. Active Appearance Model (AAM) is one of the most commonly used facial coefficient models, representing both the shape and texture variations and their correlation. Fan *et al.* [26] utilized a two-layer BiLSTM network to estimate AAM coefficients of the mouth area based on the overlapped triphone input, which then is transferred to a face image to produce a photo-realistic talking head. The experiments show that the BiLSTM network has superior performance to previous HMM-based approaches. Similarly, as shown in Fig. 8(e), Taylor *et al.* [66] introduced a simple and effective DNN as a sliding window predictor to automatically learn AAM coefficients based on the fixed-length phoneme sequence. Furthermore, the model can be retargeted to drive other face models with the help of an effective retargeting approach. The main practical limitation of AAM coefficients is that the reference face AAM parameterization may cause potential errors when retargeting to a new subject.

3D Coefficient based. Besides 2D facial coefficient models, 3D facial coefficients via principal component analysis (PCA) are more commonly used in VSG [67, 70, 171, 172, 173, 174, 175].

Pham *et al*. [171, 172, 176] proposed utilizing CNN + RNN based backbone architectures to map audio signals to blendshape coefficients [177] of a 3D face. However, these methods rely heavily on the prior 3D facial models of target speakers. Hussen *et al*. [173] finetuned a pretrained DNN-based acoustic model to map driving audios to 3D blendshape coefficients, as they hold the idea that a pretrained acoustic model has better generalization on speaker-independent VSG tasks than a randomly initialized model. As shown in Fig. 8(f), Thies *et al*. [67] proposed a generalized Audio2Expression network and a specialized UNet-based neural face rendering network for audio-driven VSG. The proposed Audio2Expression network aims to estimate temporally stable 3D blenshape coefficients based on the DeepSpeech audio features, using a CNN based backbone architecture and a content-aware filtering network. In this way, the model is able to synthesize talking face videos from an audio sequence from another person.

Besides the 3D blendshape model, Kim *et al*. [178, 179] introduced 3D Morphable Model (3DMM) [180], a more dense 3d face parametric representation, for video-oriented face2face translation. The 3DMM coefficients contain the rigid head pose parameters, facial identity coefficients, expression coefficients, gaze direction parameters for both eyes, and spherical harmonics illumination coefficients. Referring to the 3DMM based face2face translation pipeline mentioned above, [4, 68, 174, 175, 181, 182, 183] converted the driving source from video to audio clips (text scripts) and migrated this pipeline to VSG tasks. These methods have an approximate framework, as shown in Fig. 8(g). The flowchart of this framework generally follows four steps: 1) Train a network to map the driving source to the facial expression coefficients, as visual speech information is implicit in facial expression coefficients. 2) Use a pretrained deep face reconstruction model to get the 3DMM coefficients of the reference identity image. 3) Combine the 3DMM coefficients from the reference identity image and the predicted facial expression coefficients to get hybrid 3DMM coefficients. 4) Synthesize talking videos using GPU rendering or a generation network.

Following the above flowchart, Song *et al*. [181] designed a novel Audio2Expression network. They empirically find that source identity information embedded in speech features will degrade the performance of mapping speech to mouth movement. Therefore, they explicitly add an ID-Removing sub-network to remove the identity information from the driving audio. Meanwhile, a UNet-style generation network is introduced to complete the mouth region guided by mouth landmarks. Yi *et al*. [174] proposed an LSTM-based network to map audio MFCC features to facial expression and head pose, as they argue that audio and head pose are correlated in a short period. Besides, they propose a memory-augmented GAN to refine these synthesized video frames into real ones. Wu *et al*. [182] proposed an arbitrary talking style imitation VSG method. During the mapping stage, they introduced an extra style reference video as input and used a deep 3D reconstruction model to get the style code of the reference video. Next, they concatenate audio features with the reconstructed style code to predict the stylized 3DMM coefficients. However, the above 3DMM based models are not able to disentangle visual speech information from other facial expressions like eyebrow and head pose. Therefore, Zhang *et al*. [68] proposed a novel flow-guided VSG framework, including one style-specific animation generator and one flow-guided video generator, to synthesize high visual quality videos. Moreover, the style-specific animation generator successfully disentangles lip dynamics with eyebrow and head pose. Li *et al*. [184] employed a similar framework for text-driven VSG. Ji *et al*. [4] proposed an emotional video portrait (EVP) to achieve audio-driven emotional control for talking face synthesis. Unlike previous methods, they adopt the cross-reconstruction [185] technique in the audio2expression stage to decompose the input audio into disentangled content and emotion embeddings.

5.2.3 *Vertex based Methods*

3D facial vertices are another popularly used 3D face model in VSG. For example, Karras *et al*. [5] used a simple CNN-based architecture to learn a nonlinear mapping from input audios to the 3D vertex coordinates (totally 15,066 vertices) of the target face. To make the synthesized video more natural, they introduce an extra emotion code as an intuitive control for the emotional state of the face puppet. However, the proposed model is specialized for a particular speaker. To overcome this issue, as shown in Fig. 8(h), Cudeiro *et al*. [69] extended the model to multiple subjects. The proposed VOCA model concatenates the Deepspeech audio features and one-hot vector of a speaker and outputs 3D vertex (totally 5023 vertices) displacements instead of vertex coordinates. The critical contribution of VOCA is that the additional identity control parameters can vary the identity-dependent visual dynamics. Based on VOCA, Liu *et al*. [186] proposed a geometry-guided dense perspective network (GDPnet) with two constraints from different perspectives to achieve a more robust generation. Fan *et al*. [187] proposed a Transformer-based autoregressive VSG model named FaceFormer to encode the long-term audio context information and predict a sequence of 3D face vertices.

Richard *et al*. [188] proposed a categorical latent space for VSG that disentangles audio-correlated and audio-uncorrelated (facial expressions like eye blinks, eyebrow) information based on a cross-modality loss. Then, a UNet-style architecture with skip connections is used to predict 3D vertex coordinates. Since the modalities disentanglement mechanism, the plausible motion of uncorrelated regions of the face is controllable, making the synthesized video more photo-realistic. Lahiri *et al*. [70] proposed a speaker-dependent VSR method, which decomposes the audio to talking face mapping problem into the prediction of the 3D face shape and the regressions over the 2D texture atlas. To do so, they first introduced a normalization preprocessing stage to eliminate the effects of head movement and lighting variations. Then, a geometry decoder and an auto-regressive texture synthesis network were trained to learn vertex displacements and the corresponding lip-centered texture, respectively. Finally, a computer graphics based video rendering pipeline is used to generate talking videos for the target speaker.

5.3 One-Stage VSG Frameworks

The two-stage VSG frameworks have been dominated before 2018. Nevertheless, two-stage VSG frameworks suffer from the complex processing pipeline, expensive and time-consuming facial parameter annotations, extra aid technologies like facial landmark detection and monocular 3D face reconstruction *etc*. Therefore, instead of optimizing the individual components of a complex two-stage VSG pipeline, researchers have paid more attention to exploring one-stage (end-to-end) VSG approaches. One-stage VSG pipelines refer to architectures that directly generate talking lip (face) videos from driving source with an end-to-end learning strategy that does not involve any intermediate facial parameters.

Speech2Vid [189] was among the first to explore one-stage VSG frameworks. As shown in Fig. 8(i), it consists of four sub-networks. An audio encoder aims to extract speech features based on the driving audio; An identity encoder aims to extract identity features based on the reference image; And an image decoder tries to output synthesized images based on the fused speech and identity features. The above sub-networks form an autoencoder architecture, and L1 reconstruction loss is used for training. Besides, a separate pretrained deblurring CNN is introduced as a post-processing module to improve image quality. As a pioneer work, Speech2Vid provides a baseline for speaker-independent VSG and greatly motivates the research on one-stage VSG. However, Speech2Vid only uses the L1 reconstruction loss during training, which is not efficient for VSG for the following reasons. 1) The L1 reconstruction loss is operated on the whole face, and spontaneous motion of the face mainly occurs on the upper part of the face, leading to discouraging the visual speech generation. 2) As Speech2Vid is temporal-independent (no knowledge of its previous outputs), it usually produces less coherent video sequences. 3) No consideration of consistency of the generated video with the driving audio.

5.3.1 GAN based Methods

To overcome the above limitations of Speech2Vid, many researchers try to improve VSG performance by utilizing generative adversarial training [37] strategies. As shown in Fig. 8(j), GAN based VSG methods usually consist of three sub-architectures, *i.e.*, encoders, generators, and discriminators.

Taking audio-driven VSG as an example, a piece of audio naturally entangles various information, such as speech, emotion, speaking style, *etc*. As we have emphasized in Section. 2.2.2, information coupling brings enormous challenges to VSR. To ameliorate this issue, Zhou *et al*. [71] proposed a novel VSG framework called Disentangled Audio-Visual System (DAVS). Compared with previous VSG approaches, they focus more on the disentangled speech and identity feature extraction, which is realized based on supervised adversarial training. However, DAVS relies on extra Word-ID labels and Person-ID labels in the training stage. Sun *et al*. [72] improved the model by learning speech and identity features within a self-supervised contrastive learning framework, with no need for extra annotations. Si *et al*. [190] utilized knowledge distillation to disentangle emotion features, identity features, and speech features from the audio input with the help of a pretrained emotion recognition teacher network and a pretrained face recognition teacher network. Recently, some works have tried to encode additional facial controllable dynamics like emotion and head pose into the generation pipeline to generate a more natural-spontaneous talking face. For example, [191, 192] introduce additional emotion encoders, and [193] devise implicit pose encodings into the generation pipeline.

Considering the drawbacks of only using image reconstruction loss, GAN based methods focus on customizing more effective learning goals for VSG. For example, Prajwal *et al*. [38, 194] introduced a simple audio-visual synchronization discriminator for lip-syncing VSG. In addition, Chen *et al*. [116] proposed an audio-visual derivative correlation loss to optimize the consistency of the two modalities in feature spaces and a three-stream GAN discriminator to force talking mouth videos generation depending on the input audio signal.

For temporal-dependent video generation, [40, 195, 196] utilized autoregression style VSG generator networks for talking face generation. Two discriminators, *i.e.*, a frame and sequence discriminator, are used to optimize the generated facial dynamics. Based on [40], Song *et al*. [39] introduced a VSR discriminator further to improve the lip movement accuracy of generated talking videos. The ablation study demonstrated that the additional VSR discriminator helps achieve more obvious lip movement, proving our motivation that VSR and VSG are dual and mutually promoted. Furthermore, Chen *et al*. [32] developed the DualLip system to jointly improve VSR and VSG by leveraging the task duality and demonstrating that both VSR and VSG models can be enhanced with the help of extra unlabeled data. Besides the above learning goals, the optical flow discriminator [197], speech-related facial action units [198], and cross-modal mutual information estimator [199] are also utilized to optimize lip motion and cross-modal consistency of generated talking videos with the driving source.

5.3.2 Other Methods

In addition, some other one-stage VSG schemes have also been proposed. Inspired by the success of the neural radiance field (NeRF) [200], Guo *et al*. [73] proposed the audio-driven neural radiance fields (AD-NeRF) model for VSG. As shown in Fig. 8(k), AD-NeRF takes DeepSpeech audio features as conditional input, learning an implicit neural scene representation function to map audio features to dynamic neural radiance fields for talking face rendering. Furthermore, AD-NeRF models not only the head region but also the upper body via learning two individual neural radiance fields. However, AD-NeRF does not generalize well on mismatched driving audios and speakers. As shown in Fig. 8(l), unlike the previous concatenation-based feature fusion strategy, Ye *et al*. [74] presented a full convolutional neural network with dynamic convolution kernels (DCKs) for cross-modal feature fusion, which extracts features from audio and reshapes features as DCKs of the fully convolutional network. Due to the simple yet effective network architecture, the real-time performance of VSG is significantly improved.

5.4 Summary and Performance Comparison

Visual speech generation is an important and challenging problem in the cross-field of computer vision, computer graphics, and natural language analysis and has received considerable attention in recent five years. Moreover, thanks to remarkable developments in deep learning techniques, the field of VSG has dramatically evolved. In this subsection, we will discuss representative VSG methods on large-scale datasets and summarize the main issues of VSG.

Because VSG approaches have various implementation requirements (driving sources, extra technologies, diverse annotation needs, specific datasets, *etc*.) and configurations (training sets, learning paradigms, lip or whole face generation, background, pose and emotion control, *etc*.), it may be impractical to compare every recently proposed VSG method in a unified and fair manner.

It is nevertheless valuable to integrate some representative VSG methods and their requirements, configurations, and highlights into a table. Therefore, as shown in Table. 4, we summarize the performance and experimental settings of some representative VSG methods tested on large-scale, commonly used benchmark datasets, including GRID and LRW.

To give readers a general understanding of the performance of the VSG method in different frameworks, three commonly used quantitative evaluation metrics, *i.e.*, PSNR, SSIM, and LMD, are listed in Table. 4. It is worth noting that the above three

TABLE 4: The comparison of representative VSG methods.

Framework & Method			Settings				GRID			LRW			Highlights
			Input†	Training Set	Extra Requriment	By-product‡	PSNR	SSIM	LMD	PSNR	SSIM	LMD	
Two-stage	Landmark based	Chen *et al.* [65]	A+I	LRW GRID	Landmark detector	/	32.15	0.83	1.29	30.91	0.81	1.37	Proposed a Convolutional-RNN structure, which utilizes correlation between adjacent frames in the generation stage
		Das *et al.* [163]	A+I	TCD-TIMIT	Landmark detector; DeepSpeech model	Blink motion	29.9	0.83	1.22	/	/	/	Proposed two GAN based networks to learn the motion and texture separately
		Wang *et al.* [167]	A+I	GRID LRW VoxCeleb	Landmark detector; Pretrained image generator and face encoder	Head Motion	30.93	0.91	/	19.53*	0.63*	/	Regressed the head motions in accordance with audio dynamics
	Coefficient based	Song *et al.* [181]	A+V	GRID & a novel dataset	Face reconstruction	Head Motion	32.23	0.97	/	/	/	/	Proposed an Audio ID-Removing Network for pure speech feature learning
		Yi *et al.* [174]	A+I	LRW	Face reconstruction	Head Motion	/	/	/	30.94	0.75	1.58	Proposed a memory-augmented GAN module for rendered frames refining
One-stage	Auto Encoder	Chung *et al.* [189]	A+I	VoxCeleb LRW	Pretrained face encoder	/	29.36	0.74	1.35	28.06	0.46	2.25	promoted the research on one-stage VSG
	GAN Based	Vougioukas *et al.* [195]	A+I	GRID	Pretrained VSR model	/	27.10	0.82	/	23.08	0.76	/	Utilized an autoregressive temporal GAN for more coherent sequences generation
		Chen *et al.* [116]	A+I	GRID LRW	Pretrained FlowNet	/	29.89	0.73	1.18	28.65	0.53	1.92	Proposed a novel correlation loss to synchronize lip movements and input audio
		Prajwal *et al.* [194]	A+V	LRS2	/	/	/	/	/	33.4	0.96	0.60	Proved that a lip synchronization discriminator is quite useful for VSG
		Song *et al.* [39]	A+I	TCD-TIMIT LRW VoxCeleb	Pretrained VSR model	/	/	/	/	27.43	0.92	3.14	Introduced a lip-reading discriminator to guide lip motion generation
		Zhou *et al.* [71]	A+V	LRW	Word and Identity labels	/	/	/	/	26.7	0.88	/	Improved visual quality by using disentangled audio-visual representation learning
		Zhu *et al.* [199]	A+I	GRID LRW	/	/	31.01	0.97	0.78	32.08	0.92	1.21	Improved cross-modality coherence with a novel Asymmetric Mutual Information Estimator (AMIE)
		Chen *et al.* [198]	A+I	GRID TCD-TIMIT	AU Classifier	/	29.84	0.77	/	/	/	/	Used both audio and speech-related facial action units (AUs) as driving information
	Others	Ye *et al.* [74]	A+V	a mixed dataset	Pretrained AudioNet	/	/	/	/	31.98	0.81	1.44	Proposed a novel one-stage VSG paradigm with the introduction of dynamic convolution kernels

∗: Including background regions.
†: A-Audio, I-Image, V-Video.
‡: Additional effects besides VSG.

metrics are most widely used for VSG, even though they are not yet effective and perfect enough. Although many quantitative metrics for VSG were proposed recently, the following issues need further investigation.

- In the early stage of VSG, qualitative evaluations are primarily utilized, such as visualization and user preference studies [16, 66, 87]. However, qualitative evaluations are unstable and unreproducible.
- Many works have attempted to establish VSG evaluation benchmarks, and more than a dozen evaluation metrics have been proposed for VSG. Consequently, existing VSG evaluation benchmarks are not unified. Chen *et al.* [42] have conducted a survey of VSG evaluation and designed a unified benchmark for VSG according to desired properties. To promote VSG development, researchers should pay more effort to VSG evaluation benchmarks.
- The results of quantitative evaluation and qualitative evaluation are sometimes in mutual conflict. For example, some works [39, 167, 195] have observed that both the PSNR and SSIM are negatively affected by introducing image or video discriminators. Nevertheless, these discriminators significantly improve the video realism and visual quality in the user study experiments.
- In practical applications, real-time is another substantial demand for VSG. However, most of the current VSG methods ignore real-time. Therefore, real-time performance is also an important evaluation metric that needs to be considered in the future.

6 Conclusion and Outlooks

In this paper, we have presented a comprehensive review on the deep learning based VSA. We focus on two fundamental questions, *i.e.*, visual speech recognition and visual speech generation, and summarize realistic challenges and current developments, including datasets, evaluation protocols, representative methods, SOTA performance, practical issues, *etc*. We presented a systemic overview of VSR and VSG approaches and discussed their underlying connections, contributions, and shortcomings. Considering that many practical issues discussed in Section. 4.4 and Section. 5.4 remain unresolved, there are still enough opportunities for VSA research and application. We attempt to provide some ideas and discuss potential future research directions in the following.

Advanced sensors for visual speech. There are at least three reasons for developing more advanced sensors for acquiring visual speech data. (1) It is estimated that only 30% to 40% of speech sounds can be noncontact lip read. (2) Noncontact visual speech data suffer from Coupling with expression and head movement, *etc*. (3) Existing VSA systems are developed based on talking face videos. However, these systems fail when the speaker's mouth is covered by a mask (ordinary in the coronavirus disease 2019 (COVID-19) pandemic). As diverse visual speech data provides a promising way to boost the development of VSA, therefore, a potential solution is to develop contact visual speech sensors by capturing the motion of speech-related muscles.

Learning with fewer labels. As we have mentioned above, collecting a large-scale audio-visual dataset is quite costly, and manually labeling is even more consuming. Existing deep learning based VSA approaches usually rely heavily on labelled data, which is a current limitation for VSA research. Recently, some works have explored cross-modal self-supervised learning, knowledge distillation for VSA. However, it is valuable to explore other label-efficient learning paradigms like domain adaptation, active learning, few-shot learning, *etc*.

Multilingual VSA. Existing audio-visual datasets are mostly monolingual. In general, English is the most universal language. However, in some practical scenes like air traffic control (ATC) and international conference, multilingual communication is needed. Although multilingual audio speech recognition has been widely explored, multilingual visual speech recognition has received little attention yet.

Extended applications of VSA. Besides VSR and VSG, there are also some hot topics that VSA can be helpful. One of the most common tasks is audio-visual speech recognition (AVSR), a speech recognition technology that uses visual and audio information. Another typical extended task is audio-visual speech enhancement (AVSE), aiming to separate a speaker's voice given lip regions in the corresponding video by predicting both the magnitude and the phase of the target signal. Besides, for DeepFake detection, VSA can serve as an effective technology for counterfeit talking video detection.

VSA technologies for virtual characters. As an emerging type of internet application and social platform, Metaverse has recently gained a lot of attention. Virtual avatar modeling is a crucial technology in the field of Metaverse. With the rapid development of Metaverse technology, virtual characters-oriented VSA technologies also came into being. Considering that existing VSA methods mostly focus on realistic speakers, virtual characters-oriented VSA research is a potential direction in the future.

Security & robustness for VSA. Security and robustness are important requirements in the public safety field of VSA technology. Recent research has demonstrated that deep learning-based AI systems are vulnerable to different types of attacks, such as adversarial attacks , and spoofing attacks. This raises serious concerns in the field of security. However, security and robustness are not taken seriously in existing VSA approaches.

Privacy preserving for VSA. As VSA involves face-related private information, it is hard to construct a public large-scale audio-visual dataset, which also hinders the development of VSA. To address this issue, available privacy-preserving techniques such as Federated Learning, Homomorphic Encryption, and Secure Multi-Party Computation can be helpful. However, to the best of our knowledge, privacy preserving VSA research has not yet started.

References

[1] T. Chen, "Audiovisual speech processing," *IEEE signal processing magazine*, vol. 18, no. 1, pp. 9–21, 2001.

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

[3] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *NeruIPS*, vol. 31, 2018.

[4] X. Ji, H. Zhou, K. Wang, W. Wu, C. C. Loy, X. Cao, and F. Xu, "Audio-driven emotional video portraits," in *CVPR*, 2021, pp. 14 080–14 089.

[5] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM TOG*, vol. 36, no. 4, pp. 1–12, 2017.

[6] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE TPAMI*, 2018.

[7] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," *arXiv:1804.04121*, 2018.

[8] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE TMM*, vol. 2, no. 3, pp. 141–151, 2000.

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

[10] N. Tye-Murray, M. S. Sommers, and B. Spehar, "Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing," *Ear and hearing*, vol. 28, no. 5, pp. 656–668, 2007.

[11] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *CVPR*, 2021, pp. 5039–5049.

[12] Z. Akhtar, C. Micheloni, and G. L. Foresti, "Biometric liveness detection: Challenges and research opportunities," *IEEE Security & Privacy*, vol. 13, no. 5, pp. 63–72, 2015.

[13] A. Rekik, A. Ben-Hamadou, and W. Mahdi, "Human machine interaction via visual speech spotting," in *ACIVS*, 2015, pp. 566–574.

[14] K. Sun, C. Yu, W. Shi, L. Liu, and Y. Shi, "Lip-interact: Improving mobile device interaction with silent speech commands," in *ACM UIST*, 2018, pp. 581–593.

[15] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "Makelttalk: speaker-aware talking-head animation," *ACM TOG*, vol. 39, no. 6, pp. 1–15, 2020.

[16] P. Garrido, L. Valgaerts, H. Sarmadi, I. Steiner, K. Varanasi, P. Perez, and C. Theobalt, "Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track," in *Computer graphics forum*, vol. 34, no. 2. Wiley Online Library, 2015, pp. 193–204.

[17] L. Cappelletta and N. Harte, "Viseme definitions comparison for visual-only speech recognition," in *EUSIPCO*, 2011, pp. 2109–2113.

[18] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[19] K. Kirchho, "Robust speech recognition using articulatory information," Ph.D. dissertation, Citeseer, 1999.

[20] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for hmm based automatic lipreading," in *ICIP*, 1998, pp. 173–177.

[21] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE TPAMI*, vol. 24, no. 2, pp. 198–213, 2002.

[22] S. Deena, S. Hou, and A. Galata, "Visual speech synthesis by modelling coarticulation dynamics using a non-parametric switching state-space model," in *ICMI Workshop*, 2010, pp. 1–8.

[23] R. Anderson, B. Stenger, V. Wan, and R. Cipolla, "Expressive visual text-to-speech using active appearance models," in *CVPR*, 2013, pp. 3382–3389.

[24] T. Kim, Y. Yue, S. Taylor, and I. Matthews, "A decision tree framework for spatiotemporal sequence prediction," in *ACM SIGKDD KDD*, 2015, pp. 577–586.

[25] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *NeurIPS*, 2012, pp. 1097–1105.

[26] B. Fan, L. Wang, F. K. Soong, and L. Xie, "Photo-real talking head with deep bidirectional lstm," in *ICASSP*, 2015, pp. 4884–4888.

[27] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *ACCV*, 2016, pp. 87–103.

[28] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *CVPR*, 2017, pp. 3444–3453.

[29] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv:1806.05622*, 2018.

[30] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv:1706.08612*, 2017.

[31] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen, "Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *FG*, 2019, pp. 1–8.

[32] W. Chen, X. Tan, Y. Xia, T. Qin, Y. Wang, and T.-Y. Liu, "Duallip: A system for joint lip reading and generation," in *ACM MM*, 2020, pp. 1985–1993.

[33] S. Ren, Y. Du, J. Lv, G. Han, and S. He, "Learning from the master: Distilling cross-modal advanced knowledge for lip reading," in *CVPR*, 2021, pp. 13 325–13 333.

[34] C.-C. Yang, W.-C. Fan, C.-F. Yang, and Y.-C. F. Wang, "Cross-modal mutual learning for audio-visual speech recognition and manipulation," in *AAAI*, 2022.

[35] Y. Zhao, R. Xu, X. Wang, P. Hou, H. Tang, and M. Song, "Hearing lips: Improving lip reading by distilling speech recognizers," in *AAAI*, vol. 34, no. 04, 2020, pp. 6917–6924.

[36] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T.-Y. Liu, and W.-Y. Ma, "Dual learning for machine translation," *NeurIPS*, vol. 29, 2016.

[37] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *NeurIPS*, vol. 27, 2014.

[38] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *ACM MM*, 2020, pp. 484–492.

[39] Y. Song, J. Zhu, D. Li, A. Wang, and H. Qi, "Talking face generation by conditional recurrent adversarial network," in *IJCAI*, 2019, pp. 919–925.

[40] K. Vougioukas, S. Petridis, and M. Pantic, "End-to-end speech-driven facial animation with temporal gans," *arXiv:1805.09313*, 2018.

[41] S. Wang, L. Li, Y. Ding, and X. Yu, "One-shot talking face generation from single-speaker audio-visual correlation learning," *arXiv:2112.02749*, 2021.

[42] L. Chen, G. Cui, Z. Kou, H. Zheng, and C. Xu, "What comprises a good talking-head video generation?: A survey and benchmark," *arXiv:2005.03201*, 2020.

[43] S. Fenghour, D. Chen, K. Guo, B. Li, and P. Xiao, "Deep learning-based automated lip-reading: A survey," *IEEE Access*, 2021.

[44] A. Fernandez-Lopez and F. M. Sukno, "Survey on automatic lip-reading in the era of deep learning," *Image and Vision Computing*, vol. 78, pp. 53–72, 2018.

[45] W. Mattheyses and W. Verhelst, "Audiovisual speech synthesis: An overview of the state-of-the-art," *Speech Communication*, vol. 66, pp. 182–217, 2015.

[46] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, "A review of recent advances in visual speech decoding," *Image and vision computing*, vol. 32, no. 9, pp. 590–605, 2014.

[47] R. Arandjelovic and A. Zisserman, "Objects that sound," in *ECCV*, 2018, pp. 435–451.

[48] S.-W. Chung, J. S. Chung, and H.-G. Kang, "Perfect match: Improved cross-modal embeddings for audio-visual synchronisation," in *ICASSP*, 2019, pp. 3965–3969.

[49] S.-W. Chung, H. G. Kang, and J. S. Chung, "Seeing voices and hearing voices: learning discriminative embeddings using cross-modal self-supervision," *arXiv:2004.14326*, 2020.

[50] C. Sheng, M. Pietikäinen, Q. Tian, and L. Liu, "Cross-modal self-supervised learning for lip reading: When contrastive learning meets adversarial training," in *ACM MM*, 2021, pp. 2456–2464.

[51] T. Afouras, J. S. Chung, and A. Zisserman, "Asr is all you need: Cross-modal distillation for lip reading," in *ICASSP*, 2020, pp. 2143–2147.

[52] P. Ma, B. Martinez, S. Petridis, and M. Pantic, "Towards practical lipreading with distilled and efficient models," in *ICASSP*, 2021, pp. 7608–7612.

[53] H. Liu, Z. Chen, and B. Yang, "Lip graph assisted audio-visual speech recognition using bidirectional synchronous fusion." in *INTERSPEECH*, 2020, pp. 3520–3524.

[54] C. Sheng, X. Zhu, H. Xu, M. Pietikainen, and L. Liu, "Adaptive semantic-spatio-temporal graph convolutional network for lip reading," *IEEE TMM*, 2021.

[55] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," in *Interspeech*, 2017.

[56] Y. M. Assael, B. Shillingford, S. Whiteson, and N. De Freitas, "Lipnet: End-to-end sentence-level lipreading," *arXiv:1611.01599*, 2016.

[57] T. Afouras, A. Zisserman *et al.*, "Sub-word level lip reading with visual attention," *arXiv:2110.07603*, 2021.

[58] C. Wang, "Multi-grained spatio-temporal modeling for lip-reading," *arXiv:1908.11618*, 2019.

[59] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *ICASSP*, 2020, pp. 6319–6323.

[60] T. Afouras, J. S. Chung, and A. Zisserman, "Deep lip reading: a comparison of models and an online application," *arXiv:1806.06053*, 2018.

[61] X. Zhang, F. Cheng, and S. Wang, "Spatio-temporal fusion based convolutional sequence learning for lip reading," in *ICCV*, 2019, pp. 713–722.

[62] K. Xu, D. Li, N. Cassimatis, and X. Wang, "Lcanet: End-to-end lipreading with cascaded attention-ctc," in *FG*, 2018, pp. 548–555.

[63] W. Li, S. Wang, M. Lei, S. M. Siniscalchi, and C.-H. Lee, "Improving audio-visual speech recognition performance with cross-modal student-teacher training," in *ICASSP*, 2019, pp. 6560–6564.

[64] R. Kumar, J. Sotelo, K. Kumar, A. de Brébisson, and Y. Bengio, "Obamanet: Photo-realistic lip-sync from text," *arXiv:1801.01442*, 2017.

[65] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *CVPR*, 2019, pp. 7832–7841.

[66] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews, "A deep learning approach for generalized speech animation," *ACM TOG*, vol. 36, no. 4, pp. 1–11, 2017.

[67] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, "Neural voice puppetry: Audio-driven facial reenactment," in *ECCV*, 2020, pp. 716–731.

[68] Z. Zhang, L. Li, Y. Ding, and C. Fan, "Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset," in *CVPR*, 2021, pp. 3661–3670.

[69] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3d speaking styles," in *CVPR*, 2019, pp. 10 101–10 111.

[70] A. Lahiri, V. Kwatra, C. Frueh, J. Lewis, and C. Bregler, "Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization," in *CVPR*, 2021, pp. 2755–2764.

[71] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *AAAI*, vol. 33, no. 01, 2019, pp. 9299–9306.

[72] Y. Sun, H. Zhou, Z. Liu, and H. Koike, "Speech2talking-face: Inferring and driving a face with synchronized audio-visual representation," in *IJCAI*, 2021.

[73] Y. Guo, K. Chen, S. Liang, Y.-J. Liu, H. Bao, and J. Zhang, "Ad-nerf: Audio driven neural radiance fields for talking head synthesis," in *ICCV*, 2021, pp. 5784–5794.

[74] Z. Ye, M. Xia, R. Yi, J. Zhang, Y.-K. Lai, X. Huang, G. Zhang, and Y.-j. Liu, "Audio-driven talking face video generation with dynamic convolution kernels," *IEEE TMM*, 2022.

[75] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "Avicar: Audio-visual speech corpus in a car environment," in *Eighth International Conference on Spoken Language Processing*, 2004.

[76] http://www.isle.illinois.edu/sst/AVICAR/#information .

[77] M. Cooke, J. Barker, S. Cunningham, and 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.

[78] http://spandh.dcs.shef.ac.uk/gridcorpus/ .

[79] A. Czyzewski, B. Kostek, P. Bratoszewski, J. Kotus, and M. Szykulski, "An audio-visual corpus for multimodal automatic speech recognition," *Journal of Intelligent Information Systems*, vol. 49, no. 2, pp. 167–192, 2017.

[80] http://www.modality-corpus.org/ .

[81] I. Anina, Z. Zhou, G. Zhao, and M. Pietikäinen, "Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis," in *FG*, vol. 1, 2015, pp. 1–5.

[82] http://www.cse.oulu.fi/CMV/Downloads .

[83] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *ICASSP*, 2015, pp. 2130–2134.

[84] https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrw1.html .

[85] https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html .

[86] https://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox1.html .

[87] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM ToG*, vol. 36, no. 4, pp. 1–13, 2017.

[88] https://github.com/supasorn/synthesizing_obama_network_training .

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

[90] https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs3.html .

[91] https://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox2.html .

[92] B. Shillingford, Y. Assael, M. W. Hoffman, T. Paine, C. Hughes, U. Prabhu, H. Liao, H. Sak, K. Rao, L. Bennett *et al.*, "Large-scale visual speech recognition," *arXiv:1807.05162*, 2018.

[93] http://vipl.ict.ac.cn/view_database.php?id=14 .

[94] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *ICCV*, 2019, pp. 1–11.

[95] https://github.com/ondyari/FaceForensics .

[96] https://voca.is.tue.mpg.de/ .

[97] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "Mead: A large-scale audio-visual dataset for emotional talking-face generation," in *ECCV*, 2020, pp. 700–717.

[98] https://wywu.github.io/projects/MEAD/MEAD.html .

[99] https://github.com/MRzzm/HDTF .

[100] M. Mori, K. F. MacDorman, and N. Kageki, "The uncanny valley [from the field]," *IEEE Robotics & Automation Magazine*, vol. 19, no. 2, pp. 98–100, 2012.

[101] S.-W. Chung, J. S. Chung, and H.-G. Kang, "Perfect match: Self-supervised embeddings for cross-modal retrieval," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 568–576, 2020.

[102] P. Korshunov and S. Marcel, "Deepfakes: a new threat to face recognition? assessment and detection," *arXiv:1812.08685*, 2018.

[103] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *CVPR*, 2016, pp. 2387–2395.

[104] https://github.com/MarekKowalski/FaceSwap/ .

[105] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM TOG*, vol. 38, no. 4, pp. 1–12, 2019.

[106] A. Fernandez-Lopez, O. Martinez, and F. M. Sukno, "Towards estimating the upper bound of visual-speech recognition: The visual lip-reading feasibility database," in *FG*, 2017, pp. 208–215.

[107] https://austalk.edu.au/ .

[108] V. Verkhodanova, A. Ronzhin, I. Kipyatkova, D. Ivanko, A. Karpov, and M. Železnỳ, "Havrus corpus: high-speed recordings of audio-visual russian speech," in *International Conference on Speech and Computer*, 2016, pp. 338–345.

[109] E. S. Ristad and P. N. Yianilos, "Learning string-edit distance," *IEEE TPAMI*, vol. 20, no. 5, pp. 522–532, 1998.

[110] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.

[111] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "Mocogan: Decomposing motion and content for video generation," in *CVPR*, 2018, pp. 1526–1535.

[112] B. Amos, B. Ludwiczuk, M. Satyanarayanan *et al.*, "Openface: A general-purpose face recognition library with mobile applications," *CMU School of Computer Science*, vol. 6, no. 2, p. 20, 2016.

[113] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," in *ICCV*,

2019, pp. 9459–9468.

[114] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019, pp. 4690–4699.

[115] N. D. Narvekar and L. J. Karam, "A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection," in *2009 International Workshop on Quality of Multimedia Experience*, 2009, pp. 87–91.

[116] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu, "Lip movements generation at a glance," in *ECCV*, 2018, pp. 520–535.

[117] A. Koumparoulis, G. Potamianos, Y. Mroueh, and S. J. Rennie, "Exploring roi size in deep learning based lipreading." in *AVSP*, 2017, pp. 64–69.

[118] C. Zhang and H. Zhao, "Lip reading using local-adjacent feature extractor and multi-level feature fusion," in *Journal of Physics: Conference Series*, vol. 1883, no. 1, 2021, p. 012083.

[119] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *NeurIPS*, 2012, pp. 1097–1105.

[120] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE TNNLS*, 2021.

[121] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large scale image recognition," in *ICLR*, 2015.

[122] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[123] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," in *CVPR*, 2017.

[124] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *CVPR*, 2017.

[125] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: an extremely efficient convolutional neural network for mobile devices," in *CVPR*, 2018.

[126] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *ICASSP*, 2018, pp. 6548–6552.

[127] D. Feng, S. Yang, S. Shan, and X. Chen, "Learn an effective lip reading model without pains," *arXiv:2011.07557*, 2020.

[128] J. Hu, L. Shen, and G. Sun, "Squeeze and excitation networks," in *CVPR*, 2018.

[129] A. Koumparoulis and G. Potamianos, "Mobilipnet: Resource-efficient deep learning based lipreading." in *Interspeech*, 2019, pp. 2763–2767.

[130] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *CVPR*, 2014, pp. 1867–1874.

[131] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu *et al.*, "A survey on vision transformer," *IEEE TPAMI*, 2022.

[132] X. Weng and K. Kitani, "Learning spatio temporal features with two stream deep 3D CNNs for lipreading," in *BMVC*, 2019.

[133] M. Luo, S. Yang, S. Shan, and X. Chen, "Pseudo-convolutional policy gradient for sequence-to-sequence lip-reading," in *FG*, 2020, pp. 273–280.

[134] X. Zhao, S. Yang, S. Shan, and X. Chen, "Mutual information maximization for effective lip reading," in *FG*, 2020, pp. 420–427.

[135] J. Xiao, S. Yang, Y. Zhang, S. Shan, and X. Chen, "Deformation flow based two-stream network for lip reading," in *FG*, 2020, pp. 364–370.

[136] X. Chen, J. Du, and H. Zhang, "Lipreading with densenet and resbi-lstm," *Signal, Image and Video Processing*, vol. 14, no. 5, pp. 981–989, 2020.

[137] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *NeurIPS*, vol. 28, 2015.

[138] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[139] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017, pp. 6299–6308.

[140] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018.

[141] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006, pp. 369–376.

[142] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *NeurIPS*, vol. 28, 2015.

[143] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," *arXiv:1508.01211*, 2015.

[144] A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, and R. Prabhavalkar, "An analysis of incorporating an external language model into a sequence-to-sequence model," in *ICASSP*, 2018, pp. 1–5828.

[145] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *IJCV*, vol. 129, no. 6, pp. 1789–1819, 2021.

[146] J. Yu, S.-X. Zhang, J. Wu, S. Ghorbani, B. Wu, S. Kang, S. Liu, X. Liu, H. Meng, and D. Yu, "Audio-visual recognition of overlapped speech for the lrs2 dataset," in *ICASSP*, 2020, pp. 6984–6988.

[147] T. Makino, H. Liao, Y. Assael, B. Shillingford, B. Garcia, O. Braga, and O. Siohan, "Recurrent neural network transducer for audio-visual speech recognition," in *IEEE ASRU workshop*, 2019, pp. 905–912.

[148] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *ICASSP*, 2021, pp. 7613–7617.

[149] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *ACCV*, 2016, pp. 251–263.

[150] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *CVPR*, vol. 2, 2006, pp. 1735–1742.

[151] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," in *NeurIPS*, 2018, pp. 7763–7774.

[152] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *ECCV*, 2018, pp. 631–648.

[153] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. So Kweon, "Learning to localize sound source in visual scenes," in *CVPR*, 2018, pp. 4358–4366.

[154] P. Ma, R. Mira, S. Petridis, B. W. Schuller, and M. Pantic, "Lira: Learning visual speech representations from audio through self-supervision," *arXiv:2106.09171*, 2021.

[155] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *SIGGRAPH*, 1997, pp. 353–360.

[156] P. Garrido, L. Valgaerts, O. Rehmsen, T. Thormahlen, P. Perez, and C. Theobalt, "Automatic face reenactment," in *CVPR*, 2014, pp. 4217–4224.

[157] S. Fu, R. Gutierrez-Osuna, A. Esposito, P. K. Kakumanu, and O. N. Garcia, "Audio/visual mapping with cross-modal hidden markov models," *IEEE TMM*, vol. 7, no. 2, pp. 243–252, 2005.

[158] L. Xie and Z.-Q. Liu, "Realistic mouth-synching for speech-driven talking face using articulatory modelling," *IEEE TMM*, vol. 9, no. 3, pp. 500–510, 2007.

[159] A. Jamaludin, J. S. Chung, and A. Zisserman, "You said that?: Synthesising talking faces from audio," *IJCV*, vol. 127, no. 11, pp. 1767–1779, 2019.

[160] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," *IJCV*, vol. 127, no. 2, pp. 115–142, 2019.

[161] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134.

[162] S. Sinha, S. Biswas, and B. Bhowmick, "Identity-preserving realistic talking face generation," in *IJCNN*, 2020, pp. 1–10.

[163] D. Das, S. Biswas, S. Sinha, and B. Bhowmick, "Speech-driven facial animation using cascaded gans for learning of motion and texture," in *ECCV*, 2020, pp. 408–424.

[164] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv:1412.5567*, 2014.

[165] S. A. Jalalifar, H. Hasani, and H. Aghajan, "Speech-driven facial reenactment using conditional generative adversarial networks," *arXiv:1803.07461*, 2018.

[166] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv:1411.1784*, 2014.

[167] S. Wang, L. Li, Y. Ding, C. Fan, and X. Yu, "Audio2head: Audio-driven one-shot talking-head generation with natural head motion," *arXiv:2107.09293*, 2021.

[168] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," *NeurIPS*, vol. 32, 2019.

[169] Y. Lu, J. Chai, and X. Cao, "Live speech portraits: real-time photorealistic talking-head animation," *ACM TOG*, vol. 40, no. 6, pp. 1–17, 2021.

[170] Y.-A. Chung and J. Glass, "Generative pre-training for speech with autoregressive predictive coding," in *ICASSP*, 2020, pp. 3497–3501.

[171] H. X. Pham, S. Cheung, and V. Pavlovic, "Speech-driven 3d facial animation with implicit emotional awareness: a deep learning approach," in *CVPR Workshops*, 2017, pp. 80–88.

[172] H. X. Pham, Y. Wang, and V. Pavlovic, "End-to-end learning for 3d facial animation from speech," in *ICMI*, 2018, pp. 361–365.

[173] A. Hussen Abdelaziz, B.-J. Theobald, J. Binder, G. Fanelli, P. Dixon, N. Apostoloff, T. Weise, and S. Kajareker, "Speaker-independent speech-driven visual speech synthesis using domain-adapted acoustic models," in *ICMI*, 2019, pp. 220–225.

[174] R. Yi, Z. Ye, J. Zhang, H. Bao, and Y.-J. Liu, "Audio-driven talking face video generation with learning-based personalized head pose," *arXiv:2002.10137*, 2020.

[175] X. Yao, O. Fried, K. Fatahalian, and M. Agrawala, "Iterative text-based

editing of talking-heads using neural retargeting," *ACM ToG*, vol. 40, no. 3, pp. 1–14, 2021.

[176] P. Tzirakis, A. Papaioannou, A. Lattas, M. Tarasiou, B. Schuller, and S. Zafeiriou, "Synthesising 3d facial motion from "in-the-wild" speech," in *FG*, 2020, pp. 265–272.

[177] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *IEEE TVCG*, vol. 20, no. 3, pp. 413–425, 2013.

[178] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," *ACM TOG*, vol. 37, no. 4, pp. 1–14, 2018.

[179] H. Kim, M. Elgharib, M. Zollhöfer, H.-P. Seidel, T. Beeler, C. Richardt, and C. Theobalt, "Neural style-preserving visual dubbing," *ACM TOG*, vol. 38, no. 6, pp. 1–13, 2019.

[180] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set," in *CVPR Workshops*, 2019, pp. 0–0.

[181] L. Song, W. Wu, C. Qian, R. He, and C. C. Loy, "Everybody's talkin': Let me talk as you want," *IEEE TIFS*, 2022.

[182] H. Wu, J. Jia, H. Wang, Y. Dou, C. Duan, and Q. Deng, "Imitating arbitrary talking style for realistic audio-driven talking face synthesis," in *ACM MM*, 2021, pp. 1478–1486.

[183] C. Zhang, Y. Zhao, Y. Huang, M. Zeng, S. Ni, M. Budagavi, and X. Guo, "Facial: Synthesizing dynamic talking face with implicit attribute learning," in *ICCV*, 2021, pp. 3867–3876.

[184] L. Li, S. Wang, Z. Zhang, Y. Ding, Y. Zheng, X. Yu, and C. Fan, "Write-a-speaker: Text-based emotional and rhythmic talking-head generation," in *AAAI*, vol. 35, no. 3, 2021, pp. 1911–1920.

[185] K. Aberman, R. Wu, D. Lischinski, B. Chen, and D. Cohen-Or, "Learning character-agnostic motion for motion retargeting in 2d," *arXiv:1905.01680*, 2019.

[186] J. Liu, B. Hui, K. Li, Y. Liu, Y.-K. Lai, Y. Zhang, Y. Liu, and J. Yang, "Geometry-guided dense perspective network for speech-driven facial animation," *IEEE TVCG*, 2021.

[187] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "Faceformer: Speech-driven 3d facial animation with transformers," *arXiv:2112.05329*, 2021.

[188] A. Richard, M. Zollhöfer, Y. Wen, F. De la Torre, and Y. Sheikh, "Meshtalk: 3d face animation from speech using cross-modality disentanglement," in *ICCV*, 2021, pp. 1173–1182.

[189] J. S. Chung, A. Jamaludin, and A. Zisserman, "You said that?" in *BMVC*, 2017.

[190] S. Si, J. Wang, X. Qu, N. Cheng, W. Wei, X. Zhu, and J. Xiao, "Speech2video: Cross-modal distillation for speech to video generation," *arXiv:2107.04806*, 2021.

[191] N. Sadoughi and C. Busso, "Speech-driven expressive talking lips with conditional sequential generative adversarial networks," *IEEE TAC*, vol. 12, no. 4, pp. 1031–1044, 2019.

[192] S. E. Eskimez, Y. Zhang, and Z. Duan, "Speech driven talking face generation from a single image and an emotion condition," *IEEE TMM*, 2021.

[193] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu, "Pose-controllable talking face generation by implicitly modularized audio-visual representation," in *CVPR*, 2021, pp. 4176–4186.

[194] P. KR, R. Mukhopadhyay, J. Philip, A. Jha, V. Namboodiri, and C. Jawahar, "Towards automatic face-to-face translation," in *ACM MM*, 2019, pp. 1428–1436.

[195] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic speech-driven facial animation with gans," *IJCV*, vol. 128, no. 5, pp. 1398–1413, 2020.

[196] S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan, "End-to-end generation of talking faces from noisy speech," in *ICASSP*, 2020, pp. 1948–1952.

[197] D. Zeng, H. Liu, H. Lin, and S. Ge, "Talking face generation with expression-tailored generative adversarial network," in *ACM MM*, 2020, pp. 1716–1724.

[198] S. Chen, Z. Liu, J. Liu, Z. Yan, and L. Wang, "Talking head generation with audio and speech related facial action units," in *BMVC*, 2021.

[199] H. Zhu, H. Huang, Y. Li, A. Zheng, and R. He, "Arbitrary talking face generation via attentional audio-visual coherence learning," in *IJCAI*, 2021, pp. 2362–2368.

[200] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020, pp. 405–421.