

Review

Spatiotemporal Feature Enhancement for Lip-Reading: A Survey

Yinuo Ma ^{1,*} and Xiao Sun ² ¹ College of Computing, City University of Hong Kong, Hong Kong SAR 999077, China² School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China; sunx@hfut.edu.cn

* Correspondence: yinuoma2-c@my.cityu.edu.hk

Abstract: Lip-reading, a crucial technique to recognize human lip movement patterns for semantic output, has gained increasing attention due to its broad applications in public safety, healthcare, the military, and entertainment. Spatiotemporal feature enhancement techniques have played a significant role in advancing lip-reading research in deep learning. This paper presents a comprehensive review of the latest advancements in methods for lip-reading by exploring key properties of diversity enhancement techniques, involving spatial features, spatiotemporal convolution, attention mechanisms, pulse features, audio-visual features, and so on. Furthermore, according to different network structures, the six spatiotemporal feature enhancement method for lip-reading is offered. And each spatiotemporal feature enhancement method was divided into different subclasses based on the differences in the architecture structure, feature attributes, and application types. Ultimately, this is followed by an in-depth discussion of state-of-the-art spatiotemporal feature enhancement methods, accompanied by an analysis of the challenges and limitations faced, and a discussion of future research directions. From different views, this comprehensive review reveals the limitations and intrinsic disparities among these techniques in different categories for scholars to embark on innovative paths in the advancement of lip-reading.

Keywords: lip-reading; spatiotemporal features; feature enhancement; deep learning



Academic Editor: Andrea Prati

Received: 12 February 2025

Revised: 18 March 2025

Accepted: 24 March 2025

Published: 9 April 2025

Citation: Ma, Y.; Sun, X. Spatiotemporal Feature Enhancement for Lip-Reading: A Survey. *Appl. Sci.* **2025**, *15*, 4142. <https://doi.org/10.3390/app15084142>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Lip-reading, as a specialized visual language recognition system, attempts to interpret the semantic information of spoken content by analyzing the motion patterns of a speaker's lip movement captured in image sequences. Lip-reading relies solely on the sequences of lip movements from silent visual data different from acoustic signals in speech recognition and audio-visual speech recognition, which challenges the recognition precision of diverse deep learning models in intricate tasks. Recent research has highlighted the pivotal role of spatiotemporal feature enhancement in improving recognition performance [1,2], such that significant progress has been driven in lip-reading methods and model architectures. Figure 1 illustrates the fundamental venation and development orientation of spatiotemporal feature enhancement techniques in the field of lip-reading.

In recent years, deep learning-based approaches to lip-reading have garnered significant attention, leading to a surge of review papers that synthesize existing work. These reviews typically focus on aspects such as applications [3], datasets [3–5], front-end feature extraction networks [4,6], back end classifiers [4], model architectures [5,6], and audio-visual analysis [7,8]. However, there is a notable gap in the literature regarding a comprehensive survey on spatiotemporal feature enhancement. To address this gap, this paper offers a thorough review of recent advancements in lip-reading research and outlines the classification

of lip-reading methods, with a particular emphasis on spatiotemporal feature enhancement techniques. Additionally, we analyze the challenges faced in enhancing spatiotemporal features and discuss potential future directions for research.

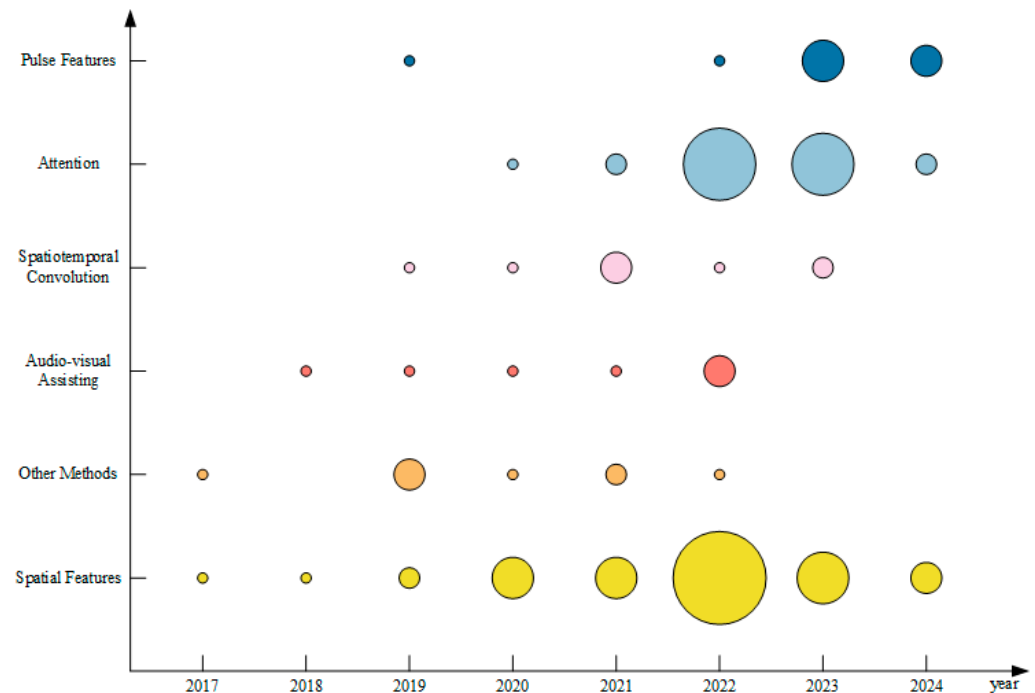


Figure 1. Development of spatiotemporal feature enhancement techniques for lip-reading. The color of the circle indicates the type of technology, and the size of the circle represents the quantity of literature.

In this survey, we conduct a comprehensive survey of more than 110 research papers focused on lip-reading. Our objective is to facilitate researchers in gaining a rapid and systematic understanding of the pertinent methods and technologies in this domain. The information from the survey is presented in Figure 1. To collect the published materials related to the research area, the following steps are followed to find corresponding keywords: First, search for literature with terms such as “lip reading”, “lip read”, “audio-visual speech recognition”, “speech recognition”, “visual speech analysis”, “lip extraction” in their titles or keywords. Next, search for alternative spellings of all critical terms. Then use Boolean operators AND and OR to connect different main clauses and spellings. It is noted that the literature originates from IEEE, Elsevier, Google Scholar, MDPI, Springer, and others to guarantee that the above search terms cover the most relevant work.

The subsequent sections are organized as follows: In Section 2, we classify various lip-reading methods and provide an in-depth analysis of four distinct lip-reading frameworks. Section 3 offers a review of the spatiotemporal feature enhancement techniques employed in lip-reading. In Section 4, we analyze the challenges associated with spatiotemporal feature enhancement. Section 5 explores potential future directions for the development of spatiotemporal feature enhancement techniques. Finally, Section 6 presents the conclusions drawn from this study. A taxonomy of spatiotemporal feature enhancement methods is illustrated in Figure 2.

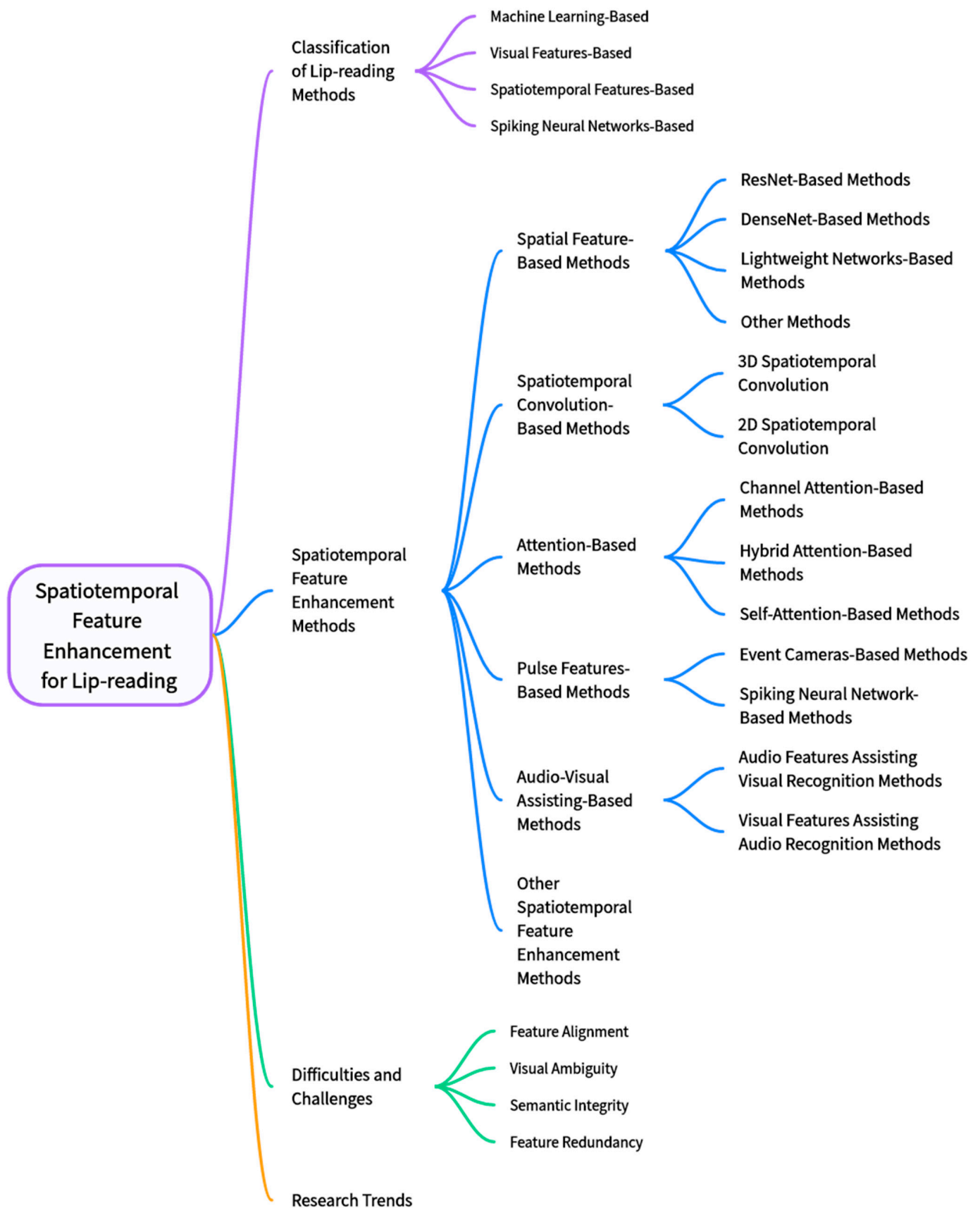


Figure 2. Taxonomy of spatiotemporal feature enhancement methods.

2. Classification of Lip-Reading Methods

Lip-reading methods are typically classified into two categories based on how features are processed: machine learning-based methods and deep learning-based methods, as illustrated in Figure 3.

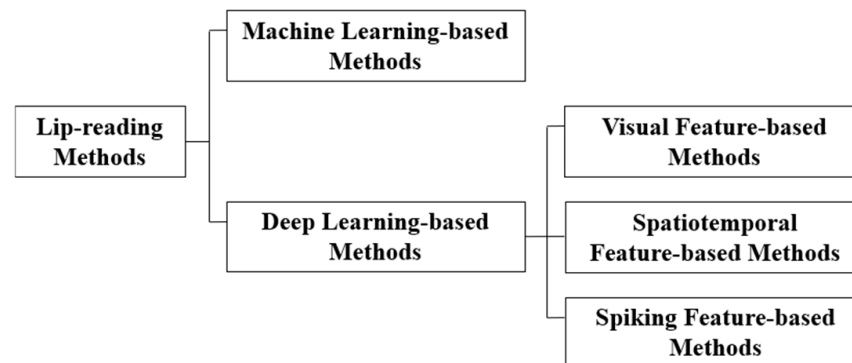


Figure 3. Classification of lip-reading methods.

Machine learning-based lip-reading methods employ traditional feature extraction algorithms to capture visual features such as spatial, textural, color, and other visual features from the Region of Interest (ROI) of the lips [5,6]. In contrast, deep learning-based lip-reading methods build feature extraction networks to extract features encompassing visual, temporal, and semantic attributes.

Based on the type of features, deep learning-based lip-reading methods are further subdivided into those based on visual features, spatiotemporal features, and pulse features, as shown in Figure 2. The key distinction between visual features and spatiotemporal features lies in their temporal and spatial information. In particular, visual features predominantly emphasize the extraction of spatial elements, i.e., visual and semantic features of the lips [4,7], which are primarily spatial features [9,10] with a small amount of temporal information, whereas spatiotemporal features concentrate on deeper and comprehensive features, intricately weaving temporal, spatial, and semantic features of the lips [11,12], which are characterized by a rich integration of spatial and temporal attributes to establish a richer, more holistic representation of the features. For example, based-pulse feature methods, as one of the based-spatial feature methods, leverage spiking neural networks to extract pulse features [13,14], which contain significant rich temporal and spatial information.

2.1. Lip-Reading Methods Based on Machine Learning

Lip-reading methods based on machine learning include two main stages: feature extraction and classification. As illustrated in Figure 4a, the process begins with the application of various transformations, such as the Discrete Cosine Transform (DCT) [15], Wavelet Transform (WT) [16], Principal Component Analysis (PCA) [17], Fourier Transform (FT) [18], and Linear Discriminant Analysis (LDA) [19,20], to extract feature descriptors from each region of interest (ROI) in video frames [21]. These transformations may also be used to extract motion features, such as optical flow [22,23], or to derive geometric features such as height, width, area, lip shape, histograms [24,25], color (texture) features [26], or model-based features (e.g., Active Appearance Model (AAM) [27], Active Shape Model (ASM) [28], Active Contour Model (ACM) [29]).

After feature extraction, these descriptors are concatenated and fed into classifiers such as Template Matching [30], Hidden Markov Models (HMM) [17,31,32], Gaussian Mixture Models (GMMs), Dynamic Time Warping (DTW) [33,34], K-Nearest Neighbors (KNN) [35], Random Forest Method (RFM) [36], Support Vector Machines (SVM) [35], and Naive Bayes (NB) [35] for classification.

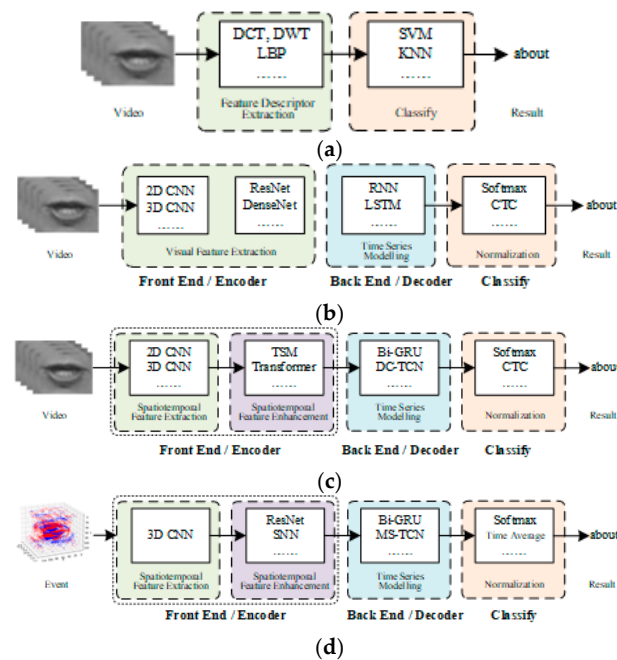


Figure 4. The illustration of the principles for lip-reading methods. (a) is the principle of machine learning-based methods; (b) is the principle of visual feature-based methods; (c) is the principle of spatiotemporal feature-based methods; (d) is the principle of pulse feature-based methods. Arrows represent data flow.

Feature extraction methods in machine learning, as illustrated in Table 1, primarily concentrate on extracting pixel features, transformation features, optical flow features, geometric features, texture features, and model-based features from lip images.

Table 1. Main feature extraction methods used in machine learning-based lip-reading.

Classification	Type of Features	Characteristic
pixel feature	multistage linear transformation, local pixel feature	The method uses lip-centered scanning lines as feature vectors, but it is sensitive to lighting variations and underperforms in high-complexity computations.
image transformation feature	Discrete Cosine Transform (DCT), Wavelet Transform (WT), Principal Component Analysis (PCA), Fourier Transform (FT), and LDA	The method uses the transformation results of all pixels as feature vectors and extracts high-frequency components for detailed information.
optical flow feature	optical flow field	The method extracts lip motion parameters and analyzes motion patterns but relies on precise positioning during preprocessing.
color feature	color space	The method focuses on color images.
model-based feature	active appearance model (AAM), active shape model (ASM), active contour model (ACM)	The method adjusts model parameters to approach the target but is prone to local minima and sensitive to the initial position.

Due to the effective capability of capturing visual features of lip images, machine learning-based feature extraction methods contribute to the long-term development of lip-reading. Nevertheless, the process of constructing these features is heavily complex and dependent on the structure of the model. This reliance on manual construction limits the scalability and broader application of these methods. Notably, these features are primarily spatial and fail to adequately reflect the temporal dynamics of lip movements, resulting in

limited feature expressiveness and low model accuracy. Consequently, the development of simple and effective feature extraction methods is becoming increasingly crucial.

As such, the exploration of deep learning frameworks for extracting visual features from lip movements has been invigorated, typically delineated into two main components: the front end, which is responsible for extracting key features from lip images, and the back end, dedicated to handling the temporal modeling of these extracted features. Alternatively, these structures can also be understood as consisting of encoder and decoder components, each contributing significantly to the overall process of analyzing lip movements. As illustrated in Figure 4b–d, deep learning-based lip-reading methods can be subdivided into three primary categories based on the type of features: visual features, spatiotemporal features, and pulse features. The following sections provide an overview of each of these approaches.

Lip-reading methods based on visual features are pivotal in capturing the morphological changes in lip contours and textures that are indicative of speech-related movements. In contrast, lip-reading methods based on spatiotemporal features are adept at capturing the dynamic interplay between spatial and temporal dimensions, which is crucial for accurate lip-reading. Lip-reading methods based on pulse features, generally draw upon pulse features, i.e., the temporal signatures of lip dynamics to artfully harness the rhythmic patterns of lip movements through time, thereby illuminating the nuances of silent speech.

2.2. Lip-Reading Methods Based on Visual Features

Lip-reading methods based on visual features, as illustrated in Figure 4b, extract visual features from lip image sequences (videos) in deep learning, paying more attention to morphological changes in lip contours and textures, which are indicative of speech-related movements. The recognition process consists of three main stages: visual feature extraction, temporal feature modeling, and classification. First, salient visual features are extracted from the front end or autoencoders by encoding lip-reading images in 2D CNNs [37], 3D CNNs [38] through these architectures like ResNet [39], DenseNet [40], EfficientNet [41], MobileNets [42], and ShuffleNet [43]. Temporal modeling of the visual feature is performed from the back end of deep models such as RNNs [44], LSTMs [45], and GRUs [46]. Finally, the classification task is carried out by using Softmax or CTC [47].

In the early stages of deep learning, researchers explored popular technologies such as ResNet18 and RNN for lip-reading, achieving notable improvements over traditional machine learning methods. Unfortunately, these methods primarily focus on extracting visual features, overlooking temporal information. Although the back end of these models employs temporal methods like RNN and LSTM to capture the sequential features of lip movements, significant temporal information is often lost due to the front-end networks' insufficient attention to temporal dynamics. As a result, the modeling of temporal features might be suboptimal. Consequently, there has been a gradual shift in focus toward extracting spatiotemporal features at the front ends of these models.

2.3. Lip-Reading Methods Based on Spatiotemporal Features

The essence of spatiotemporal features based on deep learning, as illustrated in Figure 4c, lies in extracting spatiotemporal features from lip movements, particularly capturing the dynamic interplay between spatial and temporal dimensions, which is essential for accurate lip-reading.

By replacing ResNet networks with more complex models, richer temporal information can be captured. Such a recognition process consists of four main stages: spatiotemporal feature extraction, spatiotemporal feature enhancement, temporal feature modeling, and classification. Initially, 2D CNNs act as an extractor to preserve spatial features, whereas

3D CNNs capture spatiotemporal features. Next, these models, such as TSM [48] and Transformer [49], which process both temporal and spatial features, are employed to enhance spatiotemporal features. Temporal feature modeling is then performed by using GRU, Bi-GRU [50,51], LSTM, Bi-LSTM [52], Transformer, and DC-TCN [53]. Finally, classification is carried out using SoftMax and CTC [54].

The concept of spatiotemporal feature enhancement aims to strengthen the comprehensive representation of the visual, semantic, and temporal information of the lips, capturing both the temporal and spatial information of the lip image sequence. This enhancement plays a crucial role in subsequent temporal feature modeling since its performance directly impacts the recognition accuracy, generalization capability, and robustness of the models.

2.4. Lip-Reading Methods Based on Spiking Neural Networks

Spiking Neural Networks (SNNs) exhibit common characteristics of event-based processing and the emulation of biological neural systems [55]. Closely mimicking the communication mechanisms of biological neurons, SNNs employ discrete spikes to transmit information between units. SNNs leverage pulse features, i.e., the temporal signatures of lip dynamics to enhance the recognition process for lip-reading, which are particularly effective in capturing the rhythmic patterns of lip movements associated with speech. This computational paradigm enables SNNs to effectively process spatiotemporal data [56], which is essential for lip-reading, as it facilitates the capture of key features such as articulation positions and speech rhythms. This is important for distinguishing between words that sound similar but differ in pronunciation, lip shapes, and duration, underscoring the importance of spatiotemporal information. Therefore, SNNs are well suited for lip-reading due to their strong spatiotemporal processing capabilities [13], as depicted in Figure 4d.

However, in SNNs, the reliance on event cameras, and irrelevant events caused by camera jitter and subtle changes in ambient lighting, pose new challenges to lip-reading [13]. Furthermore, while event cameras are capable of accurately capturing motion information, their ability to represent spatial information is limited. As a comparison of four lip-reading methods depicted in Table 2, recognition methods based on spatiotemporal features and those based on pulse features primarily focus on extracting and enhancing spatiotemporal features, which are the central focus of current research.

Table 2. Comparison of lip-reading methods.

Lip-Reading Methods	Characteristic Form	Front End/Encoder		Back End/Decoder
		Feature Extraction	Feature Enhancement	
Lip-reading Methods Based on Machine Learning	Handcrafted features (shape, color, texture, optical flow, models, etc.)	DCT, DWT, PCA, FT, AAM, ACM		SVM, KNN, NB, RFM, HMM
Lip-reading Methods Based on Visual Features	Spatial features (such as shape, color, texture, etc.) and temporal features	2D/3D CNN + ResNet		RNN, LSTM, GRU
Lip-reading Methods Based on Spatiotemporal Features	Spatial features (e.g., shape, color, texture), temporal features, and semantic features	3D CNN 2D CNN	CNNs, TSM, Transformer, Lightweight Network, CBAM, GCN,	GRU, Bi-GRU, LSTM, Bi-LSTM, Transformer, MS-TCN, DC-TCN
Lip-reading Methods Based on Spiking Neural Networks (SNNs)	Spatial features (e.g., shape), temporal features (e.g., pulses, sequences), and motion features.	3D CNN	ResNet, SNN	MS-TCN, Bi-GRU, Bi-LSTM, Transformer

In summary, lip language recognition based on spatiotemporal features is currently the best-performing method, while recognition based on pulse features is an emerging method. In this paper, we undertake a comprehensive review of these methodologies

for enhancing spatiotemporal features, aiming to clarify the technological progression of spatiotemporal feature extraction and enhancement.

3. Spatiotemporal Feature Enhancement Methods

Typically, such models for lip-reading that are grounded in deep learning leverage the power of 3D CNNs, 2D CNNs, or a hybrid of both to distill spatiotemporal features from sequences of lip movements. These mechanisms often struggle to fully capture and represent the temporal and spatial characteristics of video sequences. To address this issue, various spatiotemporal feature enhancement techniques have been proposed, including spatial feature enhancement, temporal convolution enhancement, attention-based enhancement, pulse feature enhancement, audio-visual auxiliary enhancement, and so on.

3.1. Spatiotemporal Feature Enhancement Based on Spatial Features

Spatiotemporal feature enhancement method based on spatial features is widely applied in the earliest deep learning methods for lip-reading by typically employing the sole mode and integration of 2D CNNs and 3D CNNs to extract spatiotemporal features, which are then enhanced in networks such as ResNet, DenseNet, EfficientNet, and ShuffleNet.

Temporal modeling is subsequently performed using RNNs, GRUs, LSTMs, etc. According to the kind of networks for feature extraction, spatial feature enhancement methods are based on ResNet, DenseNet, lightweight network, and other methods.

3.1.1. Spatial Feature Enhancement Based on ResNet

ResNet, known for its simple structure and effective recognition performance, is one of the earliest convolutional neural networks used in lip-reading and has been extensively applied for feature enhancement [57–62]. Enhancement based on ResNet is typically implemented using either 2D convolution, 3D convolution, or a combination of both convolutional approaches. Figure 5 shows the schematic diagram of the spatial feature enhancement method based on ResNet.

(1) ResNet with 2D convolution

The 2D convolution-based ResNet architecture proves particularly adept at processing spatial information. Notably, ResNet-18 and ResNet-34 have been employed in the early stages of research to extract and enhance spatiotemporal features in lip-reading.

Mudaliar et al. [63] utilized ResNet-18 to refine features derived from 3D convolutional layers, integrating a GRU-based backend for temporal feature modeling, which resulted in superior recognition performance compared to conventional methods on the LRW dataset. Similarly, concurrent studies [64,65] have typically adopted ResNet-18 as the front end to bolster spatiotemporal feature enhancement. However, a significant limitation of these models is their inability to effectively model temporal features.

Building upon these advancements, Luo et al. sought to enhance the temporal modeling capability by utilizing ResNet-18 to refine features extracted from 3D CNNs [57]. In this approach, they output a 512-dimensional vector for each time step, which was subsequently added to the corresponding position encoding and used as input for subsequent self-attention blocks. This integration of a temporal vector improved the model's ability to express features across time steps, thereby enhancing the alignment of spatiotemporal features in lip-reading videos.

To improve the recognition of fine-grained lip movements, Chen et al. [61] proposed HP-ResNet18, thereby enhancing the model's ability to describe visual actions. Specifically, HP-ResNet18 replaces the 2D convolutions of ResNet-18 with a Hierarchical Pyramid Convolution (HPCnv), which employs a set of convolutional kernels of varying sizes.

This modification enables the network to process input information at multiple spatial resolutions, facilitating the capture of both local and global contextual information and demonstrating strong generalization performance on the LRW and LRW-1000 datasets.

Thus, both approaches—HP-ResNet18's focus on spatial resolution and Luo et al.'s emphasis on temporal modeling [57]—demonstrate complementary strategies for advancing the effectiveness of lip movement recognition.

Primarily, ResNet's convolution mainly uses sliding windows in small regions for feature extraction, and is less efficient in capturing global features for lip-reading in ResNet-18. To address this, Jiang et al. [66] proposed a lip-reading framework that incorporated solid dilated convolution by replacing ResNet-18 with the Residual Global Context Network (ResGNet) to overcome the limitations of traditional convolution in capturing global features.

As an extension of ResNet-18, WideResNet-18 modified the layer arrangement of ResNet-18 and replaced the ReLU activation function with SiLU [67]. The recognition performance was evaluated on four datasets, including LRW, OuluVS, CUAVE, and SSSD, and demonstrated that the spatial feature extraction method based on ResNet-18 exhibits strong feature representation capabilities.

Compared to ResNet-18, ResNet-34 offers greater depth and improved performance, making it particularly suitable for processing highly dynamic and variable lip images [62]. Furthermore, ResNet-34 employs max-pooling layers to progressively reduce the spatial dimensions of features until the output transforms into a one-dimensional tensor for each time step. This architectural enhancement has been verified to yield a significant improvement in word-level visual speech recognition performance.

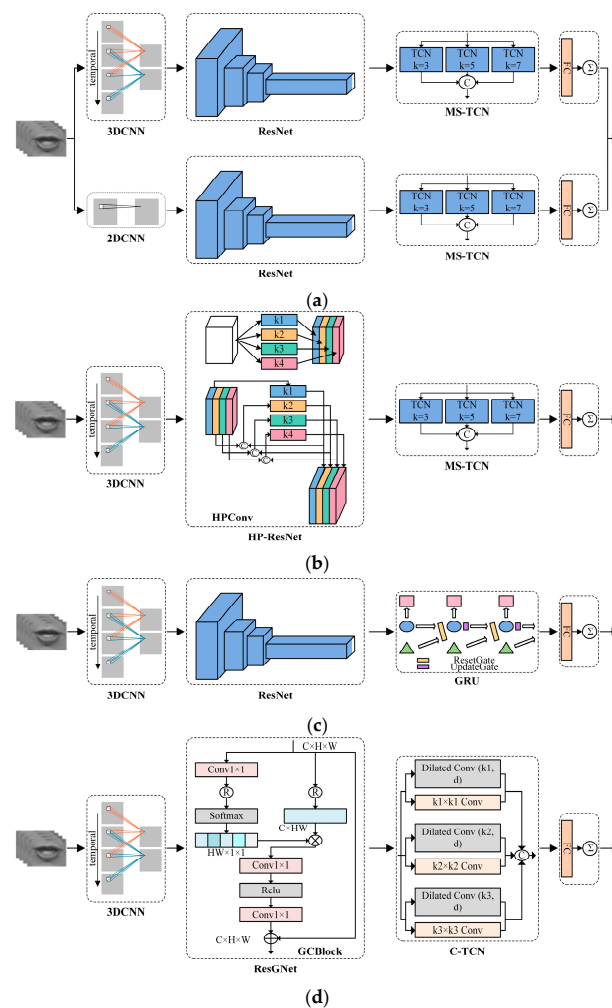


Figure 5. Cont.

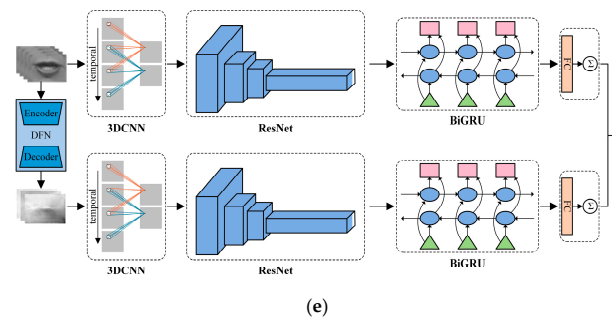


Figure 5. Spatial feature enhancement methods based on ResNet. (a) The architecture of Ref. [58]. (b) The architecture of Ref. [61]. (c) The architecture of Ref. [63]. (d) The architecture of Ref. [66]. (e) The architecture of Ref. [68].

Similarly, the spatiotemporal convolution preceding the ResNet-34 architecture incorporates ReLU activation and 2D CNN techniques, which also contribute to the reduction of the spatial dimension at each time step into a one-dimensional tensor. Consequently, this approach effectively captures lip movements that evolve [69]. Thus, the combination of ResNet-34's depth and the spatiotemporal convolution techniques underscores a comprehensive strategy for enhancing lip movement recognition in dynamic contexts.

(2) Multi-branch ResNet with 2D convolution

It is noted that the feature enhancement capabilities of single-branch ResNet are inherently limited, restricting significant improvements in lip-reading performance. To address this, multi-branch enhancement methods have been proposed to integrate both global and local features, to improve the overall effectiveness of visual feature enhancement.

Xiao et al. [68] introduced a dual-branch system that harnesses the capabilities of ResNet-18 to seamlessly integrate inputs from grayscale video streams and deformation flows, the latter generated through an encoder–decoder process. In this setup, mutual learning performed by knowledge distillation loss is designed to facilitate joint training of two branches of visual features, significantly promoting the overall system performance.

In a similar vein, Zhang et al. [58] 2022 proposed a multi-branch lip-reading method that integrates spatiotemporal, spatial, and heatmap features. ResNet-18 is responsible for extracting features from both spatiotemporal views and spatial views, combining appearance and shape features to yield more discriminative visual representations. Notably, an adaptive spatial graph model (ASGM) was introduced in the heatmap feature branch to automatically learn the spatial topology and dynamic variations of the lips, demonstrating that the combination of ResNet-18 and graph convolutional networks significantly underscores higher discriminative recognition performance.

The introduction of ResNet-18 for visual feature extraction marked a pivotal moment in capturing complementary information, which was subsequently enriched by a dual-stream fusion strategy [59]. This innovative approach promotes mutual learning between the global and local feature streams, ultimately enhancing the overall visual features.

In summary, ResNet, celebrated for its robust spatial feature modeling capabilities, has made significant strides in articulating spatiotemporal features through 2D convolution. However, despite these advancements, its ability to effectively capture and describe temporal features remains noticeably limited. This interplay between enhanced spatial representation and the challenges of temporal modeling highlights the ongoing quest for more comprehensive methods in lip movement recognition.

(3) ResNet with 3D convolution

To enhance the modeling ability of temporal features, researchers have attempted to incorporate 3D convolution into the ResNet architecture in that 3D convolution, which processes both temporal and spatial information simultaneously, is especially efficient for

capturing the dynamics of video sequences. A straightforward method involves replacing certain layers of ResNet with 3D convolutional layers.

In 2021, Zeng et al. [70] tracked the issue of the modeling ability of temporal features by transiting the first convolutional and pooling layers of ResNet from 2D to 3D, in which a compact 3D convolutional unit, Stingy Residual 3D (StiRes3D), was also introduced. Heterogeneous convolutional kernels are imposed in this unit across different input channels and then incorporated into both channel-wise and pointwise convolutions within the compact convolutional block. The experiment demonstrated that StiRes3D successfully mitigated the information loss caused by using separable convolution as an approximation for standard 3D convolution. This provides a plug-and-play solution for spatiotemporal feature extraction.

For extracting spatiotemporal information from input images consisting of multiple consecutive frames, 64-core 3D convolutional layers were embedded in ResNet by utilizing multiple convolutional neural networks and spatial attention modules; subtle variations in mouth shapes for phonetically similar words were detected, significantly enhancing the spatiotemporal feature extraction capability [54].

Further, to enhance spatiotemporal features [71], fine-grained global collaborative features (FGS) were introduced based on 3D convolution, in which the global feature learning module is for capturing the overall characteristics of the lips, whereas the local feature learning module was leveraged to learn both coarse-grained and fine-grained correlations among the features. Additionally, diffusion and fusion techniques were applied to improve the collaboration between the local and global features, strengthening the model's ability to capture critical visual information, and allowing global features to address the fine-grained requirements of lip-reading tasks more effectively.

Generally, deep learning-based lip-reading methods adopt a spatiotemporal sequence connection approach, where spatial features are first extracted, followed by global temporal modeling. Notably, these methods often fail to capture sufficient spatial information. For this reason, Huang et al. [72] proposed the Dual-Stream Spatiotemporal Separation Network (DSSN) based on ResNet with 3D convolution to extract features and retain richer spatiotemporal representations from videos. This end-to-end dual-tower architecture separately models temporal and spatial information, and integrates spatial features from the front-end network with temporal features from the back end network through collaborative learning, overcoming, to an extent, the limitations of existing lip-reading methods in fully extracting and utilizing spatial information.

3.1.2. Spatial Feature Enhancement Based on DenseNet

DenseNet, characterized by its dense connectivity pattern, facilitates the construction of feature extraction and enhancement networks, preserving information flow and gradient propagation, which simplifies the training process. With a deeper network structure compared to ResNet, DenseNet excels at extracting more complex spatial features. Figure 6 shows the schematic diagram of the spatial feature enhancement method based on DenseNet.

In virtue of this better characteristic of DenseNet, many efforts have been made to construct novel architectural structure for simulating lip movement, such as LCA-Net [73], MouthNet [74], and the work proposed by Bi et al. [75], and Jeon et al. [76]

LCA-Net [73] integrates 3D convolution with fast networks in the front-end design, forcing seamless information flow across layers. Their fast network, equipped with transformation gates t and carry gates $1-t$, allows the deep neural network to directly propagate partial input information to the output layer. By adaptively combining local features detected by individual CNN filters, the fast network encodes richer semantic features,

preserving signal strength and mitigating the vanishing gradient problem, although some redundancy in spatial convolutions remains.

The work proposed by Bi et al., aimed at effectively enhancing the model's ability to extract spatiotemporal features, utilized multiple DenseNet embedded by 3D convolution to capture both spatial and temporal features [75]. This approach encourages DenseNet to process not only the spatial characteristics of individual frames but also dynamic temporal patterns, thereby strengthening the temporal feature representation capabilities at the front end.

To extract more refined image features based on DenseNet, MouthNet is constructed by stacking DenseBlocks and TransBlocks, where each output in a DenseBlock is concatenated with the outputs of all preceding layers, performing feature reuse and dense convolution [74]. Each DenseBlock is paired with a TransBlock for sampling and dimensionality reduction, thereby achieving dense feature connections and efficient reuse. The results indicated the superior performance on the Oulu-VS2, GRID, and proprietary datasets.

In the work proposed by Jeon et al. [76], 3D convolution was employed to construct Transition layers and DenseBlocks, stacking them multiple times to form a feature enhancement network. In addition, to construct diverse spatiotemporal information, they aggregated features from multiple CNNs at different scales and depths. Specifically, they utilized the idea of Dropout to concatenate the outputs of DenseNet through Dropout free, Dropout, and Spatial Dropout, respectively, to generate richer visual features containing multi-scale and multi-depth spatiotemporal information. The comparative experiment demonstrated robust generalization performance across eight different noisy environments.

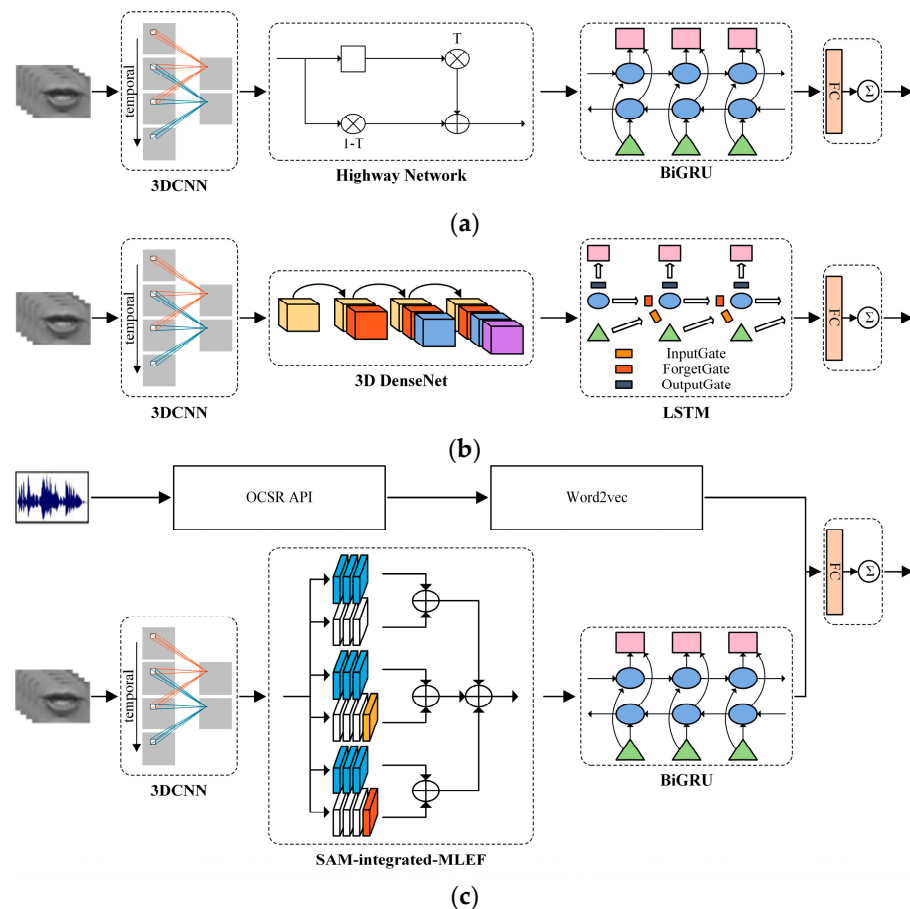


Figure 6. Spatial feature enhancement methods based on DenseNet. (a) The architecture of Ref. [73]. (b) The architecture of Ref. [75]. (c) The architecture of Ref. [76].

3.1.3. Spatial Feature Enhancement Based on Lightweight Networks

Generally, the high computational demands and stringent performance requirements of lip-reading models have severely constrained their deployment on terminal and mobile devices, thus limiting the practical applicability of lip-reading methods. As a result, reducing the parameter count and computational complexity of recognition models has become a focal point of research. In this regard, feature enhancement techniques based on lightweight neural networks, such as MobileNet, EfficientNet, GhostNet, and ShuffleNet, have gained considerable attention as effective solutions for addressing these challenges. Figure 7 shows the schematic diagram of the spatial feature enhancement method based on lightweight networks.

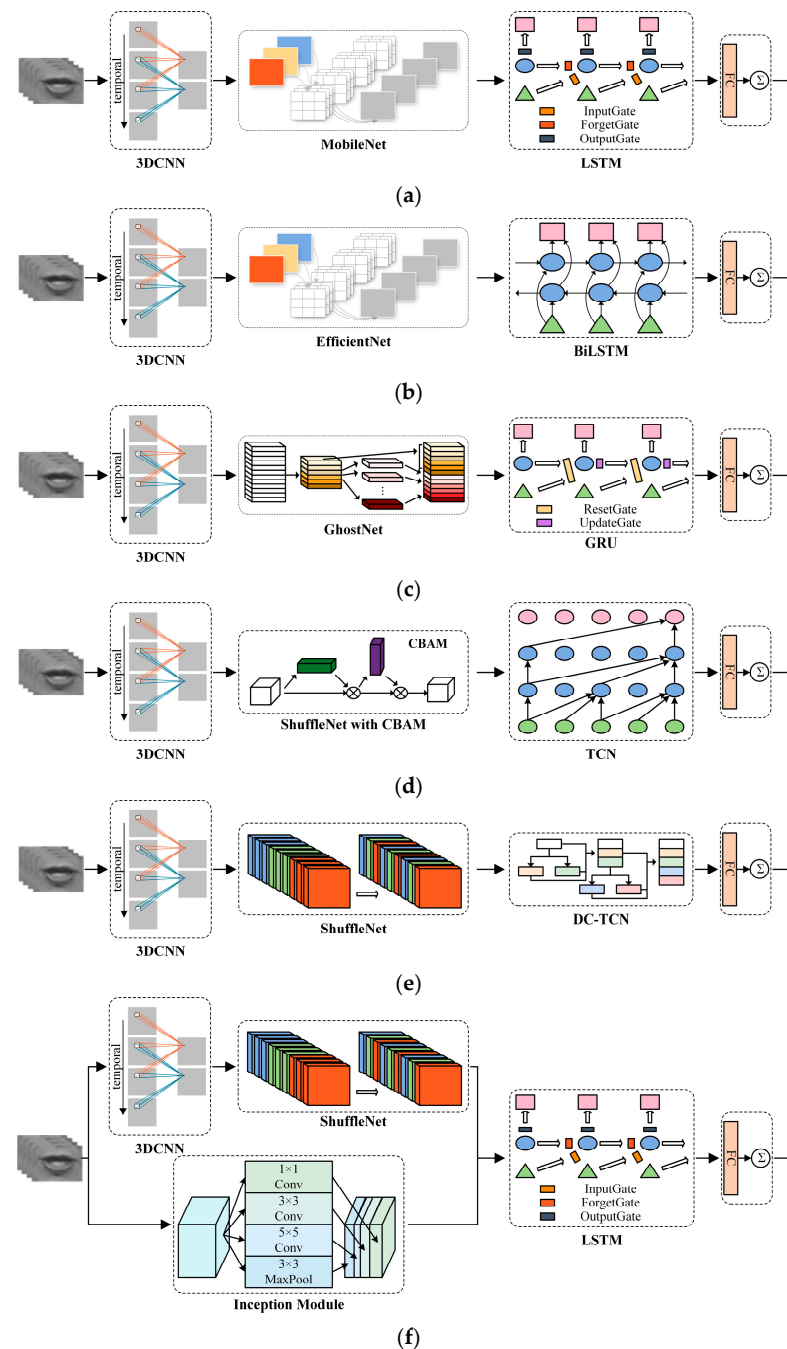


Figure 7. Spatial feature enhancement methods based on lightweight networks. (a) The architecture of Ref. [77]. (b) The architecture of Ref. [78]. (c) The architecture of Ref. [79]. (d) The architecture of Ref. [80]. (e) The architecture of Ref. [81]. (f) The architecture of Ref. [82].

(1) MobileNet

MobileNet employs depthwise separable convolution, which decomposes standard convolution into depthwise and pointwise convolutions. The method effectively reduces the parameter count and computational complexity of the network, thereby lowering the hardware requirements.

In 2019, Wen et al. [77] developed a lightweight lip-reading network using MobileNet, which extracts spatiotemporal features from each image sequence and then applies LSTM for temporal modeling. The network's simplicity enabled lip-reading on a Raspberry Pi, with its success largely attributed to MobileNet's use of depthwise separable convolutions, which enhance computational efficiency and reduce model complexity.

(2) EfficientNet

EfficientNet, as a lightweight network model, incorporates depthwise separable convolutions and Squeeze-and-Excitation (SE) blocks. The role of depthwise separable convolution is to improve computational efficiency by reducing both computational complexity and parameter count. The SE block learns inter-channel relationships and adaptively adjusts channel weights, enabling the network to emphasize important features and thereby enhancing overall performance.

In 2022, MobiLipNet, as a visual speech recognition model based on EfficientNetV2 is proposed, in which the first layer was replaced with 3D convolution. MobiLipNet uses channel width and depth multipliers for scaling to obtain multiple models with different performances. Meanwhile, inverted Residual Bottleneck (MBConv) and Fused-MBConv blocks were introduced to jointly scale the model's depth, width, and resolution. MobiLipNet provides multiple EfficientNetV2 models with varying levels of accuracy and efficiency [83].

LipSyncNet is also a lightweight lip-reading model that employs 3D CNN and EfficientNetB0 to extract temporal and spatial features, enhancing the balance across the network's width, depth, and resolution [78]. The model introduces a composite scale factor to uniformly adjust the network's dimensions, optimizing image classification performance and efficiency with disproportionately increasing computational demands.

(3) GhostNet

GhostNet is a lightweight deep learning framework based on Ghost modules, which generate phantom features through simple operations such as 1×1 convolution. These features retain high information content while being computationally efficient, achieving a balance between model performance and computational cost. Additionally, GhostNet incorporates group convolution and depthwise separable convolution to further reduce computational complexity. By maintaining high accuracy and reducing FLOPs, GhostNet enables efficient real-time inference on mobile devices.

In 2023, Zhang et al. [79] proposed Efficient-GhostNet, an optimized version of GhostNet for lip-reading. Efficient-GhostNet extracts spatial features from image sequences, reducing the model's parameters and computational load without the need for dimensionality reduction, by employing a local cross-channel interaction strategy.

(4) ShuffleNet

Unlike the above Lightweight networks, ShuffleNet primarily consists of pointwise group convolution and channel shuffle. It transforms pointwise convolution, typically used for dimensionality reduction, into pointwise group convolution. Additionally, before the 3×3 depthwise convolution, the input feature map is shuffled along the channel dimension to prevent the solidifying of channel information caused by two Group convolutions.

Based on ShuffleNet, Fu et al. [80] designed a lightweight Chinese lip-reading model by replacing the residual network. To help the network focus on the most informative features, a Convolutional Block Attention Module (CBAM) was incorporated into ShuffleNet. In

validation experiments, the method outperformed the baseline models of MobileNet v2 and ShuffleNet v2.

By combining ShuffleNet with DCTCN, CFS-DCTCN, a lip-reading model for virtual scenes, was designed to model both visual and temporal features while integrating a context understanding mechanism [81]. Experimental findings demonstrate that CFS-DCTCN extracts visual features that are strongly aligned with the training data, achieving a lip-reading accuracy of 98%.

Several works [83,84] have demonstrated that lightweight networks, such as ShuffleNet, significantly reduce the parameters and computational load of lip-reading models while maintaining high recognition accuracy. These advancements position lightweight networks as a promising direction for research and application in lip-reading technology.

3.1.4. Other Spatial Feature Enhancement Methods

Beyond the above spatiotemporal feature enhancement methods previously mentioned, additional research works have exhibited remarkable qualities in their capacity to further amplify these features. For example, Sarhan et al. [84] proposed HLR-Net, a hybrid convolutional neural network that integrates various convolutional structures, including Inception modules, gradient protection layers, and bidirectional GRU layers, to enhance spatiotemporal features. The three-layer Inception module efficiently extracts and refines spatial features, while the gradient protection layer utilizes two residual network blocks to mitigate the vanishing gradient problem. The bidirectional GRU layer further strengthens temporal features by facilitating information flow in both the forward and backward directions. Figure 8 shows the schematic diagram of the other spatial feature enhancement methods.

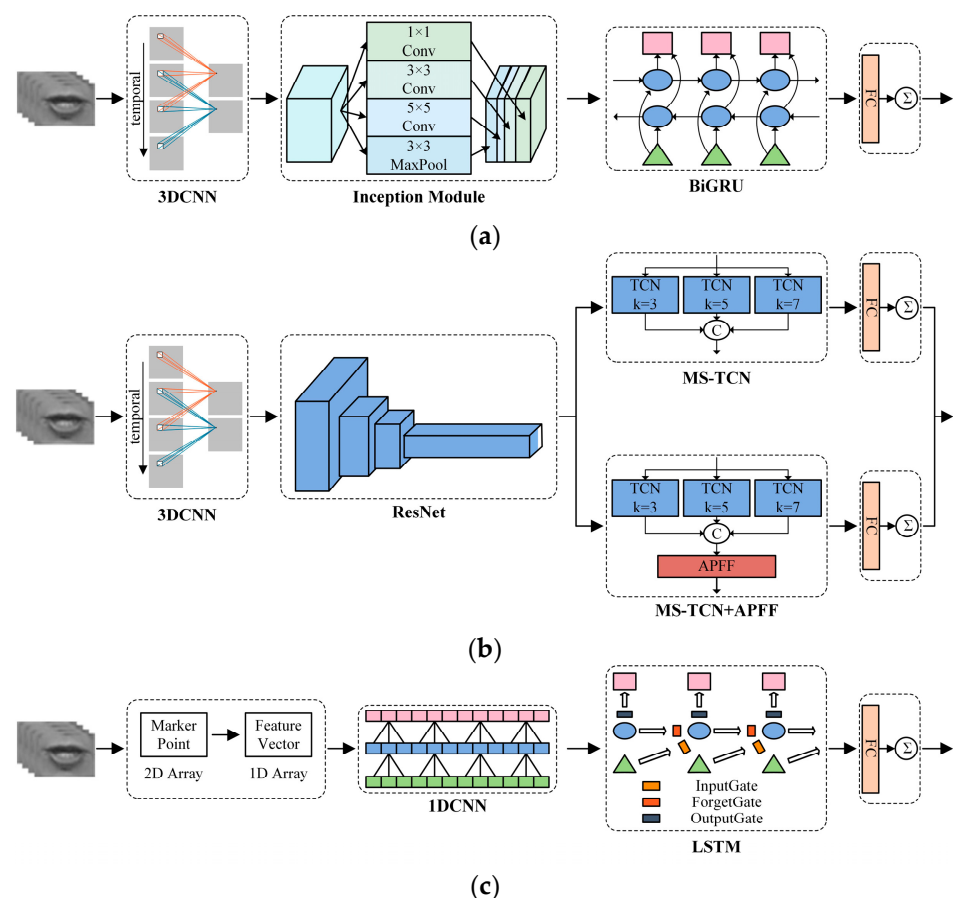


Figure 8. Other spatial feature enhancement methods. (a) The architecture of Ref. [84]. (b) The architecture of Ref. [85]. (c) The architecture of Ref. [86].

The collaboration between global spatial information and local fine-grained spatial information plays a crucial role in enhancing spatial features. For this reason, Tian et al. [85] combined global and local spatial information to enhance feature extraction, which contains two branches, dealing with global features and local features, respectively. The local feature branch divides the whole feature into three parts according to the real spatial distribution of the lip: the left corner, right corner, and the middle of the lip, and these three parts are successively entered into the local branch. These branches collaborate through a joint loss function, effectively leveraging both global and fine-grained spatial information.

Tung et al. [86] proposed a pure visual lip-reading method that extracts spatial features using the distances between 20 key points on the lips. The symmetric Euclidean distance (SED), central Euclidean distance (CED), and three-point angle (TPA) are computed to form a feature vector, which is then processed by a CNN for spatial feature extraction. Due to limited temporal feature capability, only 26 characters and 10 digits can be recognized.

To address the limitation of conventional 3D convolution in fully extracting spatial information, Peng et al. [87] employed deformable 3D convolution, which adaptively adjusts sampling positions based on lip shape structure, thereby making more effective use of spatial information.

It is worth noting that the efforts of spatial features are significant in the enhancement of spatial characteristics, but this ability is very limited in some cases, even lacking specific capability to describe temporal characteristics.

3.2. Spatiotemporal Feature Enhancement Based on Spatiotemporal Convolution

Temporal features are the key discriminative elements in lip-reading and contribute to effectively extracting temporal information from lip sequences. Spatiotemporal convolution, which effectively captures both temporal and spatial features, is widely used in lip-reading. Figure 9 shows the schematic diagram of the enhancement based on spatiotemporal convolution. These methods typically apply spatiotemporal convolution to enhance both temporal and spatial features, followed by temporal feature modeling using architectures such as MS-TCN, Bi-LSTM, and Bi-GRU. Spatiotemporal convolution-based enhancement is primarily categorized into 3D spatiotemporal convolution approaches and 2D spatiotemporal convolution approaches.

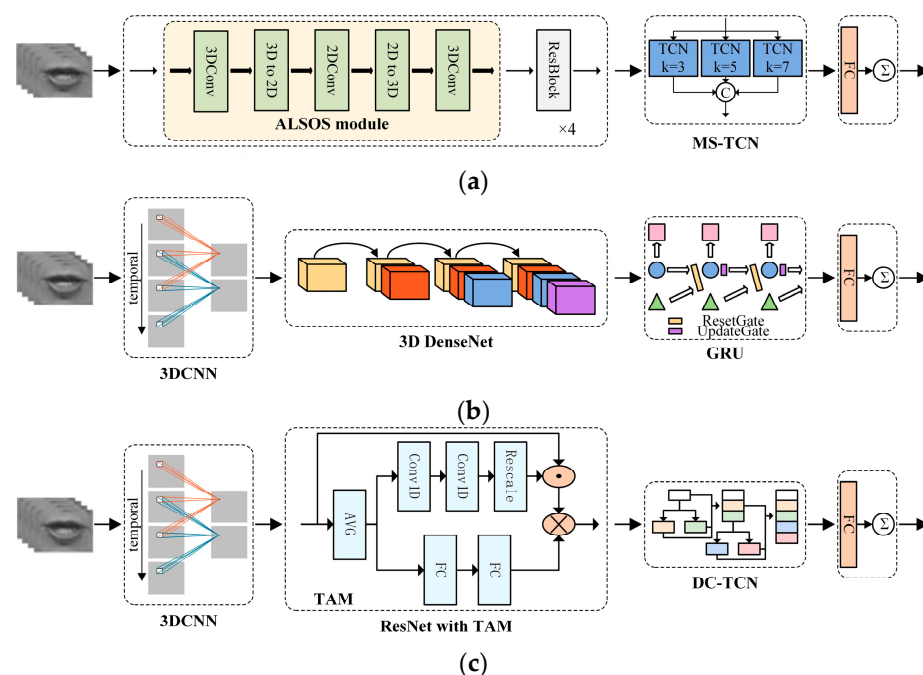


Figure 9. Cont.

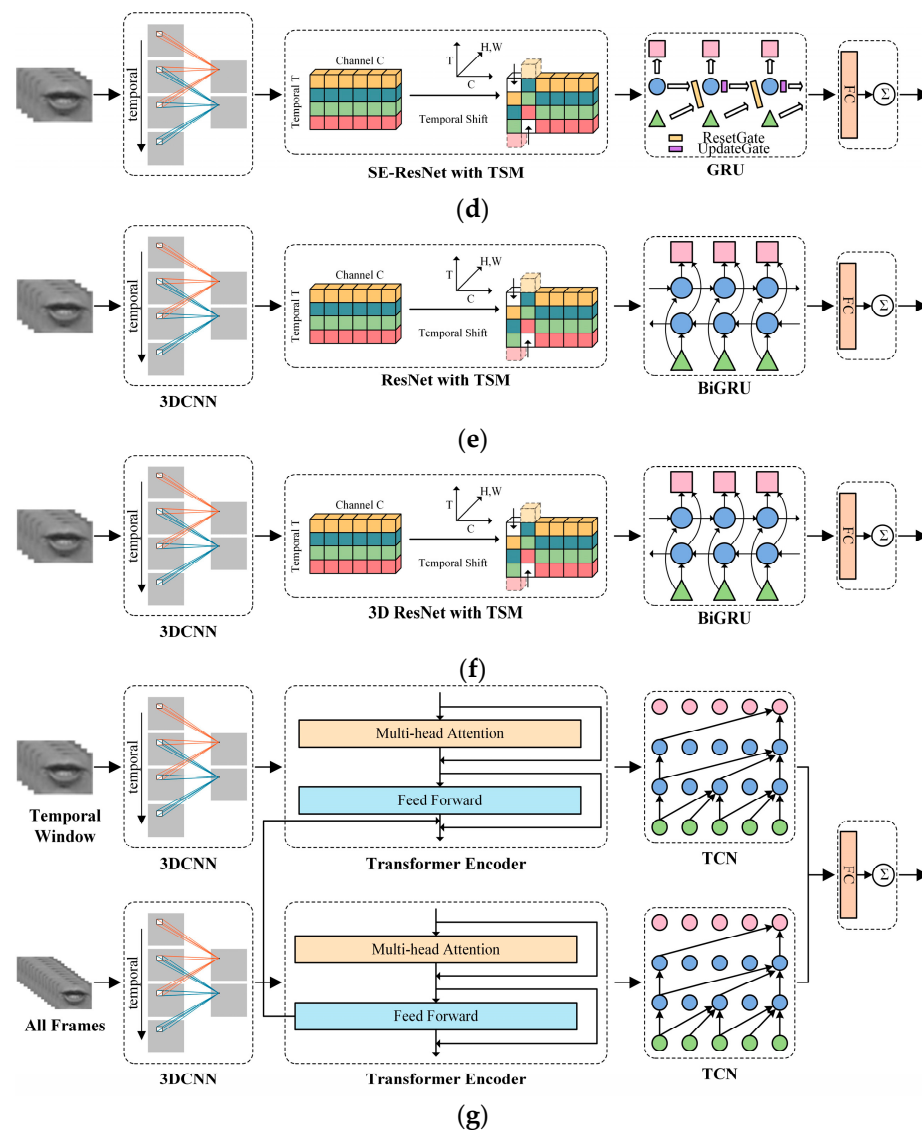


Figure 9. The typical spatiotemporal feature enhancement methods based on spatiotemporal convolution. (a) The architecture of Ref. [88]. (b) The architecture of Ref. [89]. (c) The architecture of Ref. [90]. (d) The architecture of Ref. [91]. (e) The architecture of Ref. [92]. (f) The architecture of Ref. [93]. (g) The architecture of Ref. [94].

3.2.1. Three-Dimensional Spatiotemporal Convolution Enhancement

Three-dimensional spatiotemporal convolutions proficiently extract features from video frames, skillfully capturing the spatiotemporal nuances embedded within the sequences [95,96]. This remarkable ability to simultaneously model spatial and temporal features render them exceptionally well suited for revealing the distinct characteristics of lip images.

In 2021, Dimitrios et al. [88] proposed the alternating spatiotemporal and spatial convolution (ALOS) module, which transforms sequences by preserving their length while adjusting the dimensions of sequence elements (e.g., height, width, and channels). The ALOS module consists of 3D convolutions for spatiotemporal feature extraction, 2D convolutions for spatial feature extraction, and components for converting sequences to feature maps. Nestled within the architecture of ResNet blocks, the ALOS framework introduces an innovative front end that significantly enhances the extraction of spatiotemporal features. While this integration bolsters recognition performance, the incorporation of multiple ALOS modules inevitably leads to an increase in both parameters and computational complexity.

In the next year, another front-end spatiotemporal feature enhancement network emerged by stacking 3D densely connected CNN blocks with 3D Transition blocks [89]. By introducing a temporal dimension to the densely connected convolutional kernels and pooling layers, spatiotemporal features can be well preserved. Noteworthy is that diverse spatiotemporal features can be recognized through different CNN structures and depths, and such spatiotemporal features can be further strengthened by virtue of multi-layer feature fusion of 3D CNNs in the front-end network [97].

To address the limited temporal feature extraction capabilities of standard 3D convolutions, Huang et al. [90] proposed a Temporal Adaptive Module (TAM). TAM consists of a local branch that provides position-sensitive information and a global branch associated with long-term temporal dependencies. This integration contributes to capturing complex temporal structures and enhancing robust temporal modeling ability. Furthermore, to improve the temporal feature representation capabilities, TAM can be integrated into classic network structures.

The improved feature representation capabilities of multi-granularity features were validated by a dual-branch 3D convolutional structure [11]. Specifically, by employing a 3D convolutional ResNet-52 to extract medium-granularity spatiotemporal features as a spatiotemporal enhancement branch, dynamic changes in video sequences can be effectively captured. To extract fine-granularity spatial features, a ResNet-34 was used to construct a spatial enhancement branch to capture subtle lip shape details. Additionally, unlike the direct concatenation of 2D and 3D convolutions, it incorporated a coarse-granularity spatiotemporal feature fusion module in the back end, and an attention mechanism was designed to integrate fine-grained and mid-level features into a unified representation, offering global latent patterns learning within the sequences. The comparative analysis reveals the heightened sensitivity to subtle variations in the shape of lip movement, excelling in effective distinguishing of visually similar phonemes, such as “b” and “p”.

3.2.2. Two-Dimensional Spatiotemporal Convolution Enhancement

Three-dimensional CNNs have demonstrated effectiveness in extracting temporal sequence information [69]; however, their substantial hardware requirements and computational demands limit the practical application of lip-reading models. As such, there is a rising focus on the integration of temporal feature extraction modules within 2D CNN architectures for video action recognition.

In 2019, Lin et al. [48] unveiled the Temporal Shift Module (TSM), a novel mechanism that displaces specific channels along the temporal axis to amplify the interaction between contiguous frames. Integrating the Temporal Shift Module (TSM) into the 2D CNN framework facilitates effectively performing temporal modeling without increasing the computational complexity associated with the original 2D CNN architecture [91]. Inevitably, the absence of explicit temporal action modeling in TSM constrains its capacity to thoroughly capture detailed lip motion information.

In 2021, Hao et al. [92] introduced a novel and efficient spatiotemporal feature extractor that amalgamates temporal convolution with 2D CNN architecture. This innovative approach not only emulates the temporal feature extraction capabilities of 3D models, adeptly capturing the details of lip motion, but also preserves a parameter volume and computational expense akin to that of 2D models. In the same year, Li et al. [93] further refined the Temporal Shift Module by incorporating multiple channel attention mechanisms, thereby empowering 2D convolution to glean intricate spatiotemporal features. This enhancement also alleviates the potential ramifications of channel dependencies inherent in time-shift modules.

Drawing inspiration from the SlowFast action recognition network, Wiriyathamabhum [94] unveiled the SpotFast network, which employs a temporal window as the spot path while utilizing all frames as the fast path. Notably, spatiotemporal features can be enriched through the synthesis of dual temporal convolutions in SpotFast derived from both the time window trajectory and the comprehensive frame pathway.

In essence, spatiotemporal convolution enhancement techniques extract temporal features via meticulously crafted convolutional architectures, effectively capturing both temporal and spatial information. Nonetheless, these methods encounter challenges in discerning the importance of diversity features and frequently fall short in their ability to emphasize the most discriminative elements.

3.3. Spatiotemporal Feature Enhancement Based on Attention

Attention mechanisms serve a crucial function in forcing models to hone in on prominent features but suppress irrelevant ones. In spatiotemporal feature enhancement for lip-reading, attention-based methods include channel attention, hybrid attention, self-attention, and multi-attention combinations. Figure 10 shows the schematic diagram of spatiotemporal feature enhancement based on attention.

Channel attention is essential for achieving recognition performance, as channels possess diverse discriminative strengths. By bestowing different weights upon each channel, the model can prioritize those of greater significance while elegantly suppressing the inconsequential, thereby elevating overall efficacy.

In contrast, hybrid attention deftly intertwines channel and spatial focus, allowing the model to engage with multiple features in concert. This approach, often manifested through the Convolutional Block Attention Module (CBAM), proves particularly effective in the realm of lip-reading.

Meanwhile, the emergence of self-attention and the Transformer architecture has transformed the landscape of lip-reading, empowering models to navigate and emphasize pertinent features across various dimensions.

The fusion of multiple attention mechanisms breaks free from the limitations of a single perspective, broadening the scope of feature extraction and significantly enhancing the effectiveness of recognition methods.

As depicted in Figure 10, this enhancement leverages various attention architectures, such as Squeeze-and-Excitation (SE), the Convolutional Block Attention Module (CBAM), and the Transformer, to illuminate vital spatiotemporal features. This spotlighting is further complemented by the modeling of temporal sequences, employing frameworks like DC-TCN, MS-TCN, and Transformer. This approach creates a rich fabric of insights, deepening our understanding of the data.

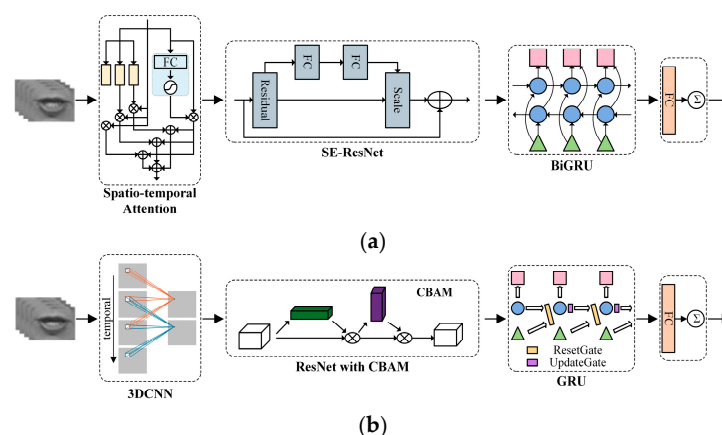


Figure 10. Cont.

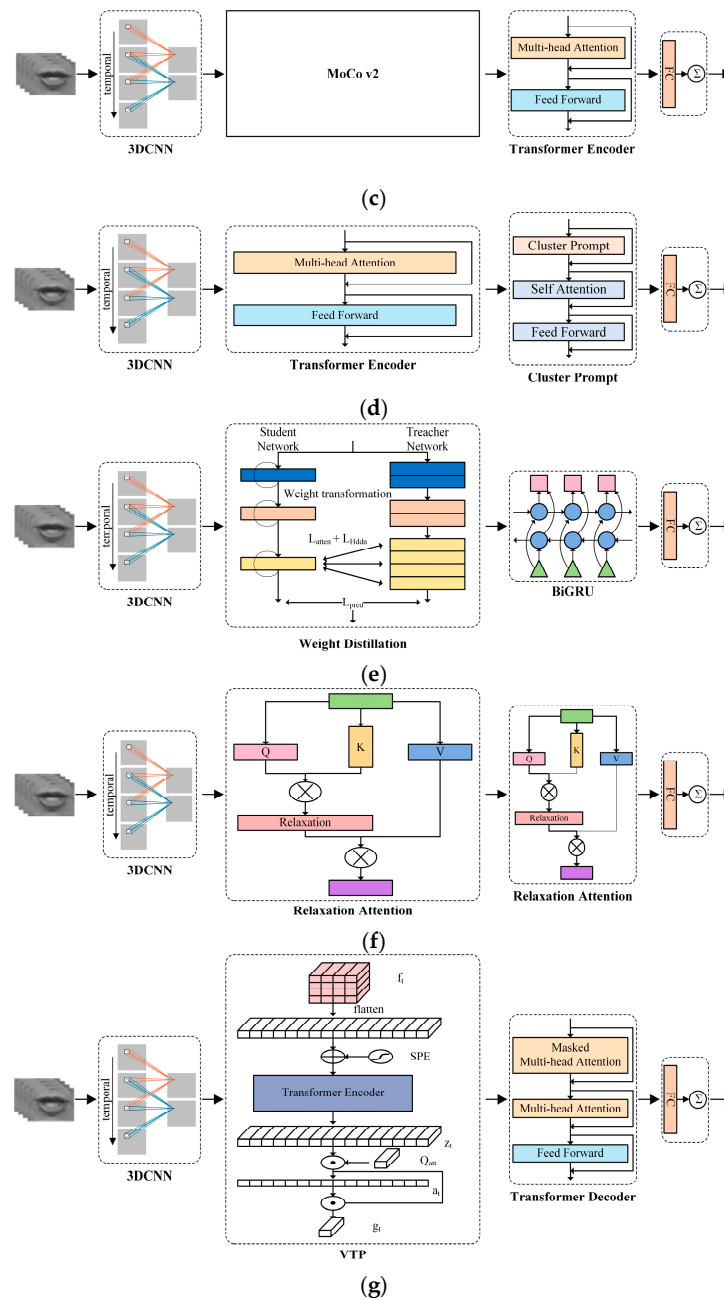


Figure 10. The typical spatiotemporal feature enhancement methods based on attention. (a) The architecture of Ref. [98]. (b) The architecture of Ref. [99]. (c) The architecture of Ref. [100]. (d) The architecture of Ref. [101]. (e) The architecture of Ref. [102]. (f) The architecture of Ref. [103]. (g) The architecture of Ref. [104].

3.3.1. Spatiotemporal Feature Enhancement Based on Channel Attention

Channel attention plays a crucial role in lip-reading systems by modulating the influence of channel features on performance. As different channels possess varying degrees of discriminative power, their impact on recognition accuracy can be substantial. By applying channel attention to assign dynamic weights to each channel, the model can prioritize the most relevant channels while suppressing irrelevant or redundant ones, thereby improving overall performance. SE blocks are frequently employed in lip-reading to draw out the rich tapestry of channel attention features.

Based on the SE module, Elashmawy et al. [105] designed a novel method called SE-ResNet by imposing Squeeze-and-Excitation (SE) blocks on ResNet in the front-end architecture aimed at harnessing attention mechanisms to capture channel dependencies. This

method enriches the spatiotemporal features of video frames by assigning greater weights to important areas within the input, allowing the model to concentrate on the most informative channels and effectively tackle the visual ambiguity present in phoneme recognition.

Differently from SE-ResNet, Haq et al. [98] harnessed the power of SE-ResNet18 as the front-end network to enrich spatiotemporal features in speech recognition. It is significant that this approach adeptly reduced noise interference, achieving superior performance on both the LRW dataset and a new dataset with sentence-level Mandarin recognition. These findings underscore the significance of channel attention in improving recognition accuracy, particularly in challenging noisy environments.

3.3.2. Spatiotemporal Feature Enhancement Based on Hybrid Attention

The attention mechanism mixed Channel-spatial features, like CBAM, artfully integrates the strengths of channel and spatial attention, demonstrating remarkable capability in lip-reading by adeptly capturing a wide range of features.

Based on CBAM attention, Lu et al. [99] and Fu et al. [80] made efforts to fortify the representation of spatiotemporal features respectively. The former embedded CBAM in ResNet50, offering a refinement that profoundly heightened the network's sensitivity to the nuanced distinctions between acoustically similar words in the Mandarin database. The latter seamlessly integrated CBAM into ShuffleNet, empowering the front-end network to concentrate on both the channel and spatial characteristics of lip movements, adeptly navigating the complexities of spatiotemporal features within video sequences.

Differently from the above two methods, Chen et al. [1] introduced a fused attention mechanism that intricately weaves together temporal, spatial, and channel dimensions, assigning weights to input data through convolution to filter out vital information before merging these insights. Furthermore, they designed the TCSAM module, which was embedded in ResNet-18 to adeptly extract lip features from images, thereby enhancing the model's capacity to emphasize both channel and spatial characteristics of lip shape and significantly improving the representational power of these two features.

3.3.3. Spatiotemporal Feature Enhancement Based on Self-Attention

With the development of self-attention and Transformer, multi-head attention mechanisms have gained prominence as a vital component in the realm of lip-reading.

(1) Spatiotemporal feature enhancement based on multi-head attention

In 2020, Luo et al. [57] integrated self-attention modules into the front end, allowing the model to weight features across different temporal steps through a series of self-attention blocks. On the back end, self-attention, multi-head attention, and Vanilla Attention Blocks were employed, enabling the model to maintain sensitivity to the entire input sequence while capturing correlations between different parts of the sequence. This approach achieved optimal recognition performance in multilingual lip-reading.

In the context of a large vocabulary system under noisy conditions, a feature-enhanced network constructed with multi-head self-attention blocks effectively harnesses auxiliary information streams, making it suitable for a large vocabulary audio-visual lip-reading across diverse scenarios. Experimental validations demonstrate that this approach exhibits robust performance in large vocabulary recognition, even in noisy environments [106].

In 2023, to recognize the complexity of a single lip shape corresponding to multiple words, Li et al. [107] expanded upon the framework of multi-head attention blocks to design a multi-head K-V memory network, adept at modeling the intricate one-to-many mapping relationships between lip movements, thereby enhancing lip-reading performance even in challenging noisy environments.

(2) Spatiotemporal feature enhancement based on Transformer

The application of Transformer models in the enhancement of audio and visual features has increasingly centered on the domain of multimodal feature enhancement, yielding significant advancements in recent years that further enrich the interplay between disparate sensory modalities.

For example, in 2022, to harness the capabilities of multi-head self-attention mechanisms, a 3D Visual Transformer (3DCvT) is introduced to adeptly learn the latent distribution of the input modality, facilitating a more profound extraction of spatiotemporal features across successive frames [108].

In the realm of audio-visual multimodal lip-reading, in 2022, Pan et al. [100] utilized the Transformer framework for self-supervised learning, where the front end employs six-layer Transformers to separately extract visual and audio features, succeeded by another six-layer Transformer that refines the fused audio-visual representations; the back end similarly incorporates six-layer Transformers for decoding, implementing CTC loss to facilitate the self-supervised learning process.

For facilitating multimodal alignment within the phoneme space, in 2023, Ref. [101] harnessed Transformer for feature extraction and enhancement; they proposed an open-modal speech recognition (OpenSR) system that permits models trained on a singular modality, such as pure audio, to be effectively applied across multiple modalities, including both pure visual and audio-visual inputs. This innovative system not only enables modal transfer between modalities but also achieves exceptionally competitive zero-shot performance, outpacing conventional few-shot and full-shot lip-reading methods.

A comparative analysis of three Transformer-based lip-reading models, Transformer-CTC, Av-HuBERT, and Moco-word2vec, revealed that the Moco framework, which elegantly intertwines convolutional networks with Transformers, consistently outperformed its counterparts across video, audio, and audio-visual recognition tasks [109]. This finding underscores the notion that the synergy of convolutional and Transformer networks substantially enriches feature expressiveness, leading to remarkable advancements in model performance.

(3) Spatiotemporal feature enhancement based on Transformer + CNN

The Transformer architecture excels at capturing temporal features within lengthy sequences, showcasing its remarkable proficiency in modeling long-range dependencies. In contrast, convolutional networks (CNNs) are particularly effective at extracting local and spatial features, finely tuned for the retrieval of nuanced details. Thus, the harmonious integration of Transformer and CNN models facilitates the extraction of a richer tapestry of both temporal and spatial attributes, thereby significantly enhancing the performance of lip-reading.

In 2022, the integration of 3D convolution into visual Transformers marked a notable advancement in the extraction of spatiotemporal features from sequential video frames. This innovative hybrid architecture exploits the temporal dynamics rendered by 3D convolution, while simultaneously harnessing the global feature extraction prowess of the Transformer. As a result, this synergy substantially enhances the model's capacity to decipher intricate patterns of dynamic visual information, i.e., offering a deeper understanding of the fluidity and complexity inherent in video content.

Adeptly tackling the intricate challenges associated with visual-spatial feature extraction, temporal feature modeling, and the pursuit of lightweight design, in 2024, a pioneering application of the Transformer architecture emerged. This innovative approach promoted the capabilities of the Transformer to extract both local and global features from successive images, thereby enriching the understanding of visual information [102]. To further streamline the model, weight transformation and weight distillation techniques were deftly integrated within the convolutional and Transformer frameworks. These strategic measures

not only facilitated the compression of the model but also effectively addressed the multifaceted issues at the intersection of spatial and temporal feature extraction, culminating in a robust and efficient solution for contemporary visual tasks.

The advent of regularized Dropout represented a notable improvement in achieving alignment between the training and reducing stages, fostering a more cohesive learning strategy [103]. In a parallel vein, a relaxed attention Transformer is designed to alleviate the overfitting problem [110]. By incorporating regularization terms into the self-attention layers of the encoder and implementing a relaxed decoding approach that constrained the internal language model within the decoder, their method effectively mitigated the risks of overfitting, thereby promoting a more robust and generalizable performance.

Certainly, the Visual Transformer Pooling Block (VTP) also can be merged into Transformer framework as an innovative constructor to effectively aggregate prior visual representations [104]. The key concept of this method is to optimize the flow of information, which facilitates a more refined comprehension of visual data by emphasizing relevant features. Experimental results implied that by incorporating this visual feature enhancement module, the VTP exhibited exceptional performance in sub-word-level lip-reading, reflecting the enhanced capacity of attention mechanisms to distill and prioritize critical visual cues in the pursuit of accurate and discriminating recognition.

The advent of HVSEM, a hybrid approach that intricately combines visual sub-word and end-to-end modeling, marked a significant evolution in the field. This method employs cascaded Transformers enhanced by cross-layer attention mechanisms, effectively refining visual features and bolstering the exchange of information between the cascaded structures. By doing so, it not only enriches the feature representation but also minimizes error accumulation, a critical challenge in complex modeling scenarios. The efficacy of HVSEM has been demonstrated through its successful application in both visual sub-word collaboration and end-to-end word-level lip-reading, showcasing the potent synergy derived from advanced attention mechanisms in navigating intricate recognition tasks.

(4) Spatiotemporal feature enhancement based on multi-attention combination

A single attention mechanism may confine its focus to limited features, whereas the integration of multiple attention mechanisms significantly broadens the scope and depth of a single attention, thereby enhancing feature representation and facilitating improved model generalization. Capitalizing on this principle, the Conformer emerges as a powerful architecture that harmoniously fuses the merits of Transformer and convolutional structures, adept at capturing both long-range dependencies and local features.

By leveraging the architecture's capacity to seamlessly integrate robust feature extraction from both Transformer and convolutional layers, the model designed by Ma et al. adeptly processes and interprets complex spatiotemporal data across diverse languages. This innovative strategy represents a significant advancement in lip-reading technology, illustrating the profound efficiency and potential of combined attention mechanisms in unraveling the intricacies of visual different language interpretation [111].

3.4. Spatiotemporal Feature Enhancement Based on Pulse Features

Spatiotemporal feature enhancement based on pulse features captures lip motion and spatial information by processing binary pulse signals, primarily through event cameras and spiking neural network enhancement methods.

The intention of enhancement based on pulse features is to process binary pulse signals and preserve latent pulse feature information in lip motion and spatial information. By harnessing these innovative approaches inherent in event camera technology and advanced spiking neural network methods, the system can effectively discern intricate dynamics

of lip movement, thus enriching the representation of spatial information pertinent to visual communication.

3.4.1. Spatiotemporal Feature Enhancement Based on Event Cameras

Event cameras distinguish themselves from conventional frame-based cameras by their ability to continuously capture data, generating asynchronous binary pulses that encode critical information regarding time, position, and intensity. This unique capability results in event images that provide superior recognition accuracy compared to their frame-based counterparts, particularly well-suited for the extraction and enhancement of lip spatiotemporal features. From the principles of spatiotemporal feature enhancement methods based on event cameras illustrated in Figure 11, it can be revealed that these methods resemble spatial feature-based enhancement since it is fundamentally characterized by the use of an event stream as its input, thereby enabling a more dynamic and responsive analysis of lip movements in real-time scenarios.

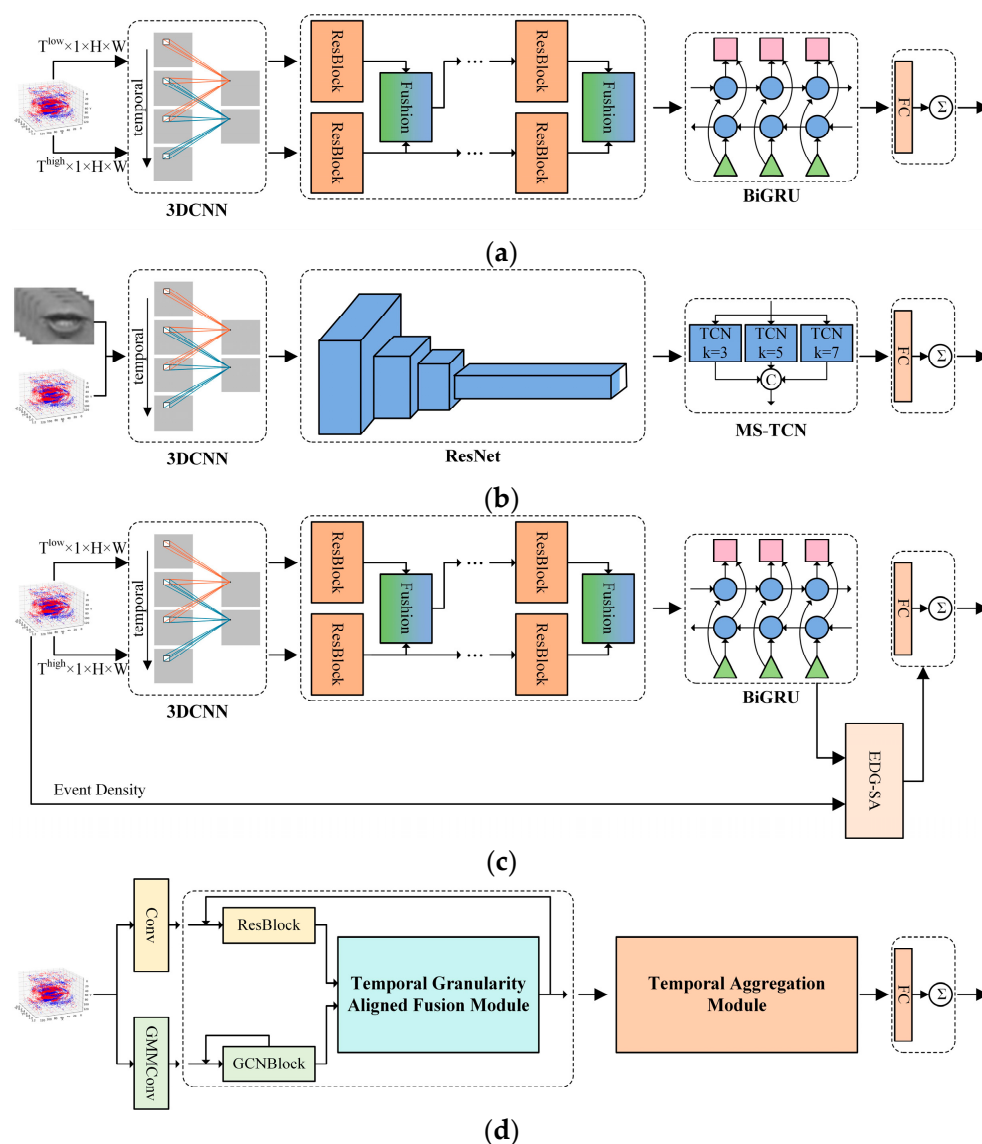


Figure 11. The typical spatiotemporal feature enhancement methods based on event cameras. (a) The architecture of Ref. [112]. (b) The architecture of Ref. [113]. (c) The architecture of Ref. [114]. (d) The architecture of Ref. [115].

In 2022, Tan et al. [112] introduced the multi-scale temporal-spatial feature perception network (MSTP), which employs a multi-branch architecture consisting of low-speed and high-speed event processing streams. This design allows the model to concurrently capture comprehensive spatial features through the low-speed branch and intricate temporal characteristics via the high-speed branch. Additionally, the authors developed a multi-flow module (MFM) that utilizes attention-weighted maps to effectively incorporate the multi-scale temporal-spatial features from both event streams, thereby elevating perceptual capability to temporal-spatial features.

To effectively address the challenge of recognizing isolated phonemes, in 2023, a novel lip-reading technique was introduced by combining frame and event cameras [113]. By employing ResNet-18 for the enhancement of both types of visual data, their experiments revealed that these multi-modalities strategies yielded superior recognition accuracy over the utilization of a single modality alone.

In 2024, researchers proposed a multi-granularity spatiotemporal feature learning framework designed to extract intricate features from events characterized by microsecond-level temporal resolution [114]. This comprehensive framework includes two perceptual branches. One focuses on low-frame-rate features with complete spatial data but coarse temporal information. The other is on high-frame-rate features with fine temporal resolution at the cost of spatial detail. A temporal aggregation sub-network is incorporated to consolidate features from these branches, thereby fully capturing temporal correlations within event streams.

Branches operating at different frame rates can capture spatiotemporal features with varying degrees of granularity. However, the aggregation of events into event frames often results in the loss of intricate temporal details. In response to this challenge, Zhang et al. [115] proposed a novel event representation method that establishes connections between key local voxels and a graph list, facilitating the alignment and fusion of global spatial features derived from event frames with local spatiotemporal characteristics represented in the voxel graph. Additionally, they developed a time aggregation module incorporating positional encoding to effectively capture both local absolute spatial information and global temporal dynamics. Experimental results revealed that this innovative approach surpassed the performance of both event-based and video-based lip-reading techniques, further building on the previously introduced multi-granularity framework and emphasizing the crucial role of fine-grained temporal information in enhancing recognition accuracy.

3.4.2. Spatiotemporal Feature Enhancement Based on Spiking Neural Networks

Spiking neural networks (SNNs) are distinguished by their event-driven processing paradigm and emulation of biological neural architectures, whereby information is conveyed through discrete spikes, mirroring the functionality of biological neurons [55]. This inherent feature renders SNNs particularly adept at processing spatiotemporal data, which is indispensable in the domain of lip-reading. Such data allow for the extraction of critical phonetic characteristics, including articulation positions and rhythmic patterns of speech. The capacity of SNNs to discern phonetically analogous lexical items—differentiated by variations in lip morphology, enunciation processes, and temporal dynamics—underscores the significance of spatiotemporal information in auditory–visual integration tasks, highlighting their utility in lip-reading applications.

As illustrated in Figure 12, the enhancement derived from the spiking neural network is intricately tied to its reliance on event-based input processing. Initially, the SNN engages with raw event data, meticulously extracting and amplifying the underlying spatiotemporal features. Once this enhancement is achieved, the temporal dynamics are subsequently modeled through advanced architectures, including Gated Recurrent Unit (GRU) and

Bidirectional Gated Recurrent Unit (Bi-GRU), which facilitate a nuanced understanding of the data's temporal progression.

The integration of two asynchronous modalities, dynamic vision sensor (DVS) and dynamic audio sensor (DAS), within a spiking neural network provides a harmonious compensation of spatiotemporal feature information [116]. The lip-reading framework could perform training effectively in each modality independently, as well as in a joint training configuration. Empirical results demonstrate that the model with a joint training strategy achieves a 23% higher recognition accuracy than the DVS model.

In 2023, a novel solution emerged to tackle the challenge of erroneous spikes generated by identical timestamps in event-driven algorithms [117]. This efficient event-driven learning approach not only addresses this issue but also reinforces the temporal feature processing capabilities inherent in event-driven neural networks. Importantly, this mechanism demonstrates compatibility with a range of spike-based plasticity mechanisms, enhancing its versatility and applicability in diverse neural network architectures.

In the same year, event-based sensors were utilized to construct the multi-grained spatiotemporal feature perceived (MSTP) network, designed to capture the events generated by lip movements [118]. This innovative model employs an SNN architecture to classify short event sequences into distinct word categories. Leveraging neuromorphic computing, this approach unfolds the benefits of high efficiency and low latency, making the model particularly well suited for real-time embedded applications.

Though SNNs demonstrate a commendable ability to extract spatiotemporal features, particularly temporal characteristics related to lip movements, they still face limitations in distinguishing between visually similar phonetic elements. In 2023, Liu et al. introduced the spatial-temporal attention block (STAB), an event-based lip-reading model that enhances the differentiation of visually similar words by employing spatial and temporal attention branches that dynamically emphasize the most relevant features reflecting lip movements [13]. Through a fusion mechanism, the model establishes a comprehensive and focused representation of lip activity that effectively differentiates between visually similar words, surpassing existing SNN architectures in terms of both recognition accuracy and energy efficiency.

The sparse activation and event-driven nature of SNNs render them exceptionally suited for low-power, low-latency edge applications within end-to-end neuromorphic hardware pipelines. Regrettably, their accuracy is often hindered by a range of intrinsic limitations and external influences. For this reason, Dampfhofer et al. [14] proposed a novel SNN model that modifies the ResNet18 block with spiking neurons and employs a dual-stream input approach using both positive and negative event samples. To enhance the sparsity of the SNN, an additional spike loss function was introduced alongside the standard accuracy loss, resulting in significant improvements in model performance and enabling precise, low-power end-to-end lip-reading.

Liquid state machines (LSMs) have gained recognition within SNNs due to their low training costs and exceptional suitability for processing spatiotemporal sequences in event streams. Capitalizing on the cost-effective nature of LSMs, Yu et al. [119] introduced a soft fusion method for lip-reading, designed to extract the spatiotemporal features of lip movement event streams. This soft fusion approach effectively employs the principles of the attention mechanism, utilizing a series of masks to modulate the weights of features derived from both visual and audio channels. By selectively accentuating the most pertinent information from each modality, this method artfully integrates visual representations with auditory cues, culminating in an enriched feature representation that significantly enhances recognition performance.

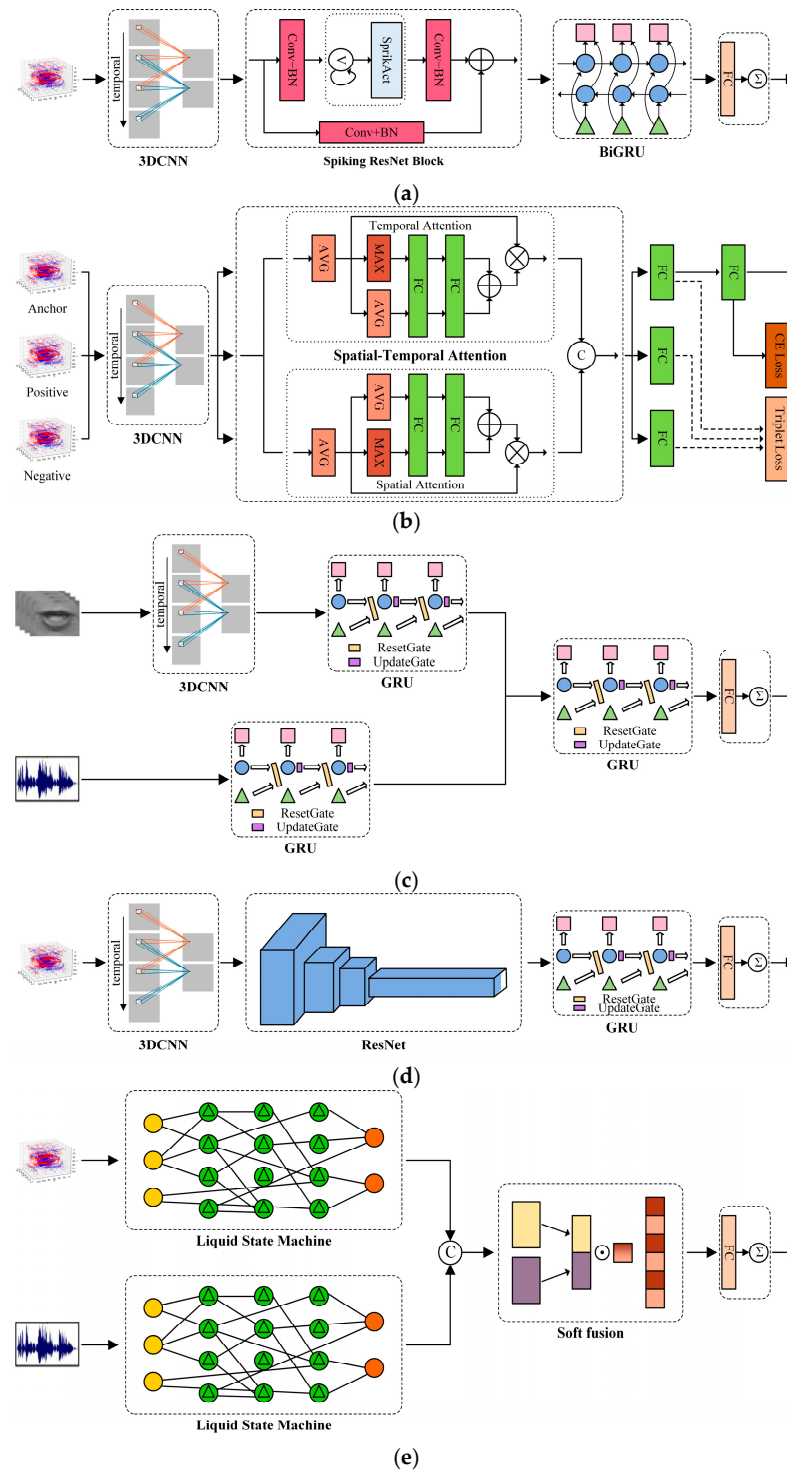


Figure 12. The typical spatiotemporal feature enhancement methods based on spiking neural networks. (a) The architecture of Ref. [14]. (b) The architecture of Ref. [55]. (c) The architecture of Ref. [116]. (d) The architecture of Ref. [118]. (e) The architecture of Ref. [119].

3.5. Spatiotemporal Feature Enhancement Based on Audio-Visual Assisting

Isolated visual lip-reading and audio recognition models often grapple with challenges of limited recognition performance and diminished robustness in the face of interference. To address these shortcomings, researchers have delved into a variety of multimodal feature enhancement strategies that effectively harness the complementary strengths of both visual and auditory cues. These audio-visual approaches emphasize two primary dimensions:

leveraging visual features to enhance the accuracy of audio recognition, and employing audio features to bolster the effectiveness of visual recognition.

Figure 13 illustrates the foundational principles and representative models of these audio-visual assisted enhancement methods, showcasing their potential to foster deeper understanding and improved performance in recognition tasks.

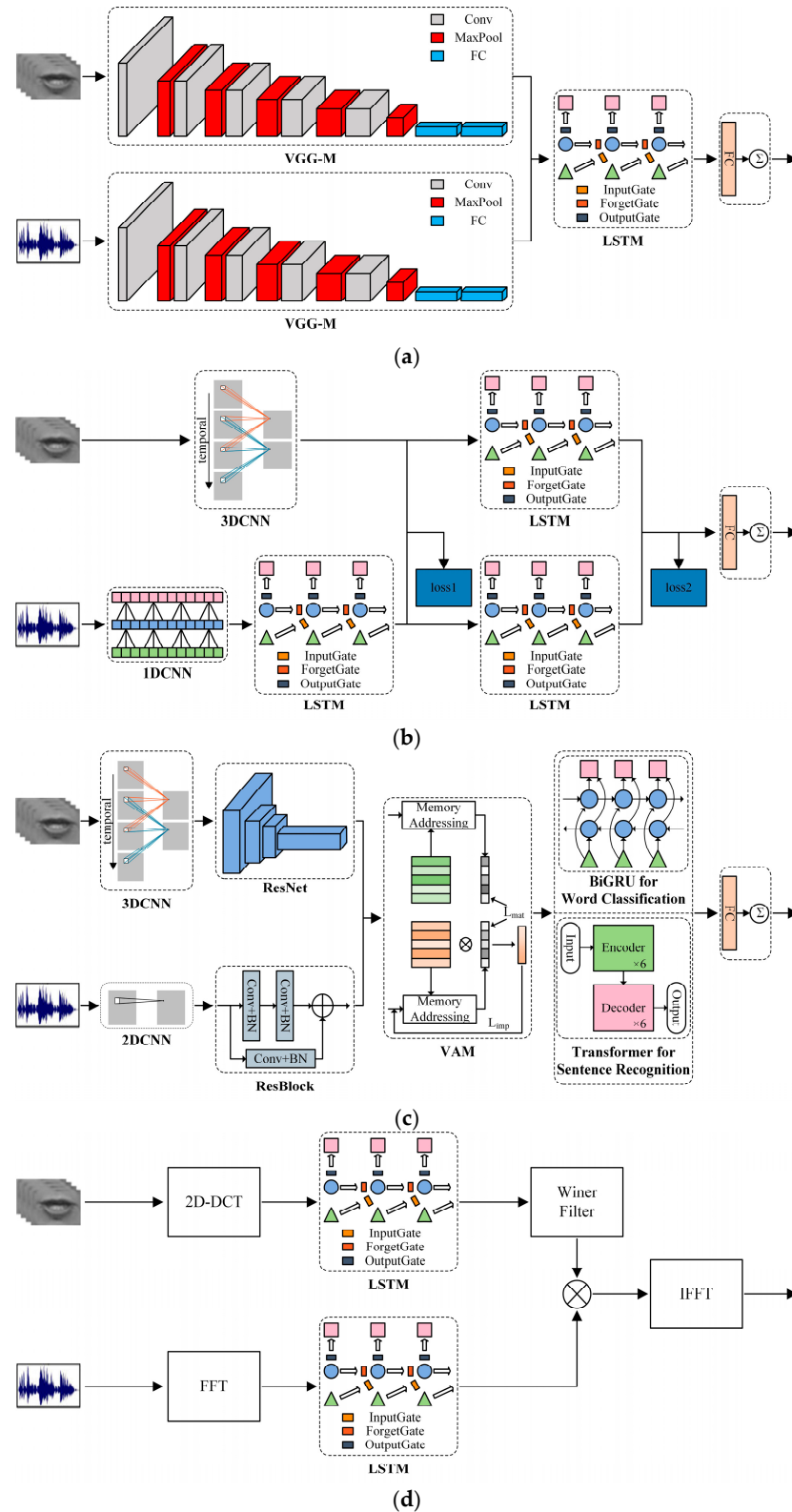


Figure 13. Cont.

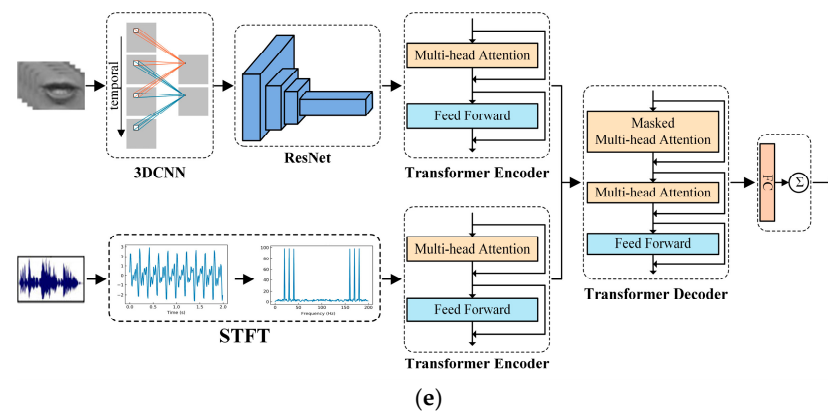


Figure 13. The typical spatiotemporal features enhancement methods based on audio-visual assisting. (a) The architecture of Ref. [120]. (b) The architecture of Ref. [121]. (c) The architecture of Ref. [122]. (d) The architecture of Ref. [123]. (e) The architecture of Ref. [124].

3.5.1. Audio Features Assisting Visual Recognition

In the realm of lip-reading, the incorporation of auditory features to strengthen visual recognition has emerged as a pivotal strategy for overcoming the shortcomings of cross-modal enhancement methods. Extensive research has been dedicated to multimodal feature enhancement techniques that capitalize on the synergistic potential of auditory and visual signals to elevate recognition performance. Among these, cross-modal fusion and audio-visual memory augmentation stand out as prominent methodologies that facilitate audio-assisted visual recognition, effectively bridging the gap between the two modalities and enriching the overall comprehension of spoken communication.

(1) Cross-modal fusion enhancement

Cross-modal fusion is established through the careful design of feature enhancement networks meticulously crafted for the distinct characteristics of auditory and visual modalities. By harmonizing the strengths of both types of data, this intricate architecture empowers auditory features to augment and amplify visual cues, cultivating a more comprehensive and robust recognition system, and simultaneously paving the way for more intricate interpretations of multimodal information.

In 2018, Chung et al. [120] unveiled SyncNet, a dual-stream convolutional neural network that employs cross-modal self-supervised learning to effectively capture the joint embedding of auditory signals and lip movements from unlabeled data, demonstrating superior performance compared to existing methodologies across prominent datasets and excelling in tasks such as audio-visual synchronization, active speaker detection, and multi-frame temporal lip-reading. By deftly weaving together disparate data sources, SyncNet has not only propelled the field forward but also redefined the possibilities of multimodal recognition.

In 2020, Handa et al. [121] presented an innovative approach by introducing a 128-dimensional subspace designed to encapsulate the feature vectors of speech signals alongside their corresponding lip movements. They conceptualized lip-reading as an unconstrained processing of natural speech signals within video sequences. To achieve this, features were meticulously extracted from both audio and visual channels: the audio channel employed 1D convolution to derive temporal features, while the visual channel utilized 3D convolution to generate rich visual feature vectors. These temporal and visual features were subsequently transformed into long-term audio and visual representations using Long Short-Term Memory (LSTM) networks, culminating in feature alignment and model training and enhancing the robustness and efficacy of lip-reading.

These enhancement methods for cross-modal fusion hinge upon the effective alignment of features from both modalities, necessitating a high quality of speech signal input. This critical requirement imposes a significant influence on the overall performance of the model, as the integrity of the sound data serves as the foundation for accurate and robust multimodal processing.

(2) Audio-visual memory enhancement

The challenges posed by articulatory movements in lip-reading can be attributed primarily to two interrelated factors: the inherent lack of information conveyed by lip movements and the presence of homophonic words that appear similar on the lips but differ in meaning. These issues represent significant roadblocks to achieving high accuracy in lip-reading. In response to these obstacles, researchers have sought to address the limitations by exploring dual-modal approaches that enhance audiovisual memory, thereby striving to improve the accuracy and effectiveness of lip-reading.

To address the challenge of insufficient visual information related to lip movements, and the insufficiency of visual information, a cross-modal memory enhancement method known as Visual-Audio Memory (VAM) was proposed [122]. VAM is based on a key-value storage structure, functioning as a cross-modal memory and involving two main components, i.e., the lip-video key memory, which stores visual cues related to lip shapes, and the audio value memory, which captures the associated audio characteristics. This architecture enables the audio value memory to annotate audio features while simultaneously guiding the lip-video key memory to retain the locations of these annotated audio signals. Consequently, this innovative approach alleviates the limitations posed by the inherent lack of visual data to an extent. Leveraging audio information to supplement and enhance the visual features of lip movements enriches the lip-reading process and enhancing overall recognition accuracy.

Addressing the challenges posed by insufficient visual information and the presence of homophones, Kim [125] introduced multi-head visual-audio memory (MVM) by modeling the interrelationships between paired audiovisual representations for the effective retention of audio features. The architecture comprises multi-head key memories that safeguard visual features alongside a single value memory dedicated to audio knowledge, allowing MVM to draw forth potential candidate audio features during the recognition phase. During inference, visual inputs can retrieve and store audio representations from memory by examining the learned interrelationships. This synergy empowers the recognition model to enrich visual data with these audio representations, thereby significantly improving its proficiency in discerning homophones.

Enhancing audiovisual memory necessitates supplementary storage for both visual and auditory features. This involves querying the storage system for data matching that generates corresponding auditory features for the stored visual characteristics, requiring additional storage capacity from the system. Furthermore, the efficacy of this data matching is contingent upon the quantity of stored features, which significantly influences the model's overall recognition performance.

3.5.2. Visual Features Assisting Audio Recognition

In 2019, Adeel et al. cultivated the complementary strengths of deep learning and acoustic modeling through filtering techniques to craft a deep learning regression model for speech enhancement. Specifically, they employed an enhanced visual Wiener filter (EVWF) to estimate the clean audio power spectrum by utilizing an inverse filter-bank transformation that facilitates speech recognition through visual features while effectively eliminating the need for voice activity detection (VAD) and noise estimation. This innovative approach

mitigated the detrimental impact of noise on speech recognition, paving the way for clearer auditory understanding [123].

Multi-modality information from visual and auditory cues demonstrated a remarkable improvement in recognition accuracy. These multi-modalities, grounded on the Transformer self-attention mechanism, employed Connectionist Temporal Classification (CTC) loss and the other utilizing sequence-to-sequence loss to explore how lip-reading enhances audio-based speech recognition. Their investigations into speech recognition under noisy conditions revealed that visual information significantly bolsters recognition performance, particularly when audio is compromised by noise [124].

A Mel spectrogram, as a tool in the fields of speech and music research, represents the frequency spectrum of an audio signal on a Mel scale based on human auditory perception, and represents time with color intensity reflecting the energy levels at corresponding frequencies as visual representation. Inspired by Mel spectrograms, in 2022, a novel approach is designed by synthesizing sound from lip movements by encoding visual features into hidden states for identifying speakers by capturing appropriate lip movements in an unconstrained manner [82]. Subsequently, it employed an attention mechanism to autoregressively generate frames of the Mel spectrogram, effectively integrating the speaker's visual characteristics into the synthesized speech. This innovative framework underscores the potential of leveraging visual information to enhance speaker identification, paving the way for more robust multimodal recognition systems.

These studies demonstrate that visual information can significantly aid in speech recognition, particularly in noisy environments. By leveraging visual cues, the effectiveness of speech recognition systems is greatly enhanced, which further provides a robust solution to the difficulties encountered in acoustic interference.

3.6. Other Spatiotemporal Feature Enhancement Methods

Building upon these advancements in integrating visual information for improved speech recognition, researchers have also incorporated various techniques such as restricted Boltzmann machines, moment features, optical flow, graph convolution, and variational temporal masking into convolutional neural networks (CNNs). These elements work together to construct hybrid front-end modules that significantly enhance spatiotemporal feature extraction capabilities. By blending these diverse methodologies, the aim is to refine the analysis of intricate data patterns, thereby advancing the effectiveness of applications in both speech and visual recognition.

Figure 14 shows the schematic diagram of other spatiotemporal feature enhancement methods.

A restricted Boltzmann machine (RBM) is a class of stochastic neural networks designed to model and learn the underlying probability distribution of its input data for tasks such as feature extraction, dimensionality reduction, and collaborative filtering. To jointly learn visual features of the lip region from multiple perspectives, Petridis et al. embraced the multi-branch approach to learn the visual features of lips from diverse perspectives. This architecture enabled the dynamic modeling of the temporal characteristics inherent in each perspective, ensuring that the subtleties of lip movements were accurately captured, resulting in a comprehensive representation that significantly enriches the understanding of lip dynamics in the context of speech recognition [126].

Orthogonal Hahn moments, as a feature-extracting tool of lip movement, is merged in the first layer of the CNN architecture to construct the Hahn Convolutional Neural Network (HCNN) [127]. This method utilizes discrete orthogonal Hahn moments to compute the moments of the input images, resulting in a moment matrix whose dimensions are determined by the order of the moments, providing an optimized representation of the

images, significantly reducing the dimensionality of the processed data while preserving crucial features essential for subsequent analysis.

Meanwhile, efficient strategies such as the enhanced visually derived Wiener filter (EVWF) and a Variational Time Masking (VTM) module are effective in enhancing spatiotemporal feature extraction capabilities in a backend structure. Specially, EVWF is used to estimate the clean audio power spectrum and eliminate voice.

Activity detection (VAD) and noise estimation for visual-driven speech recognition [123]. The VTM module automatically analyzes the frame-level feature importance, aiming to restrict the flow of irrelevant or noisy visual features to the back end network without compromising prediction accuracy, which contributes to enhancing the interpretability and generalization of lip-reading models [128].

Optical flow, as a computer vision technique that estimates the motion of objects between consecutive frames in a video for accurate tracking motion-based recognition tasks, has demonstrated satisfactory performance levels comparable to video data. Building on this insight, optical flow is integrated as an additional input alongside grayscale video to construct a dual-stream fusion method for extracting a more comprehensive extraction of spatiotemporal features [129]. Moreover, at the front end of this methodology, a 3D Inception network is employed to learn these features from both the grayscale video and optical flow data. The significant classification accuracy on large-scale image and video datasets underscores the effectiveness of multiple data modalities in the advancement of lip-reading.

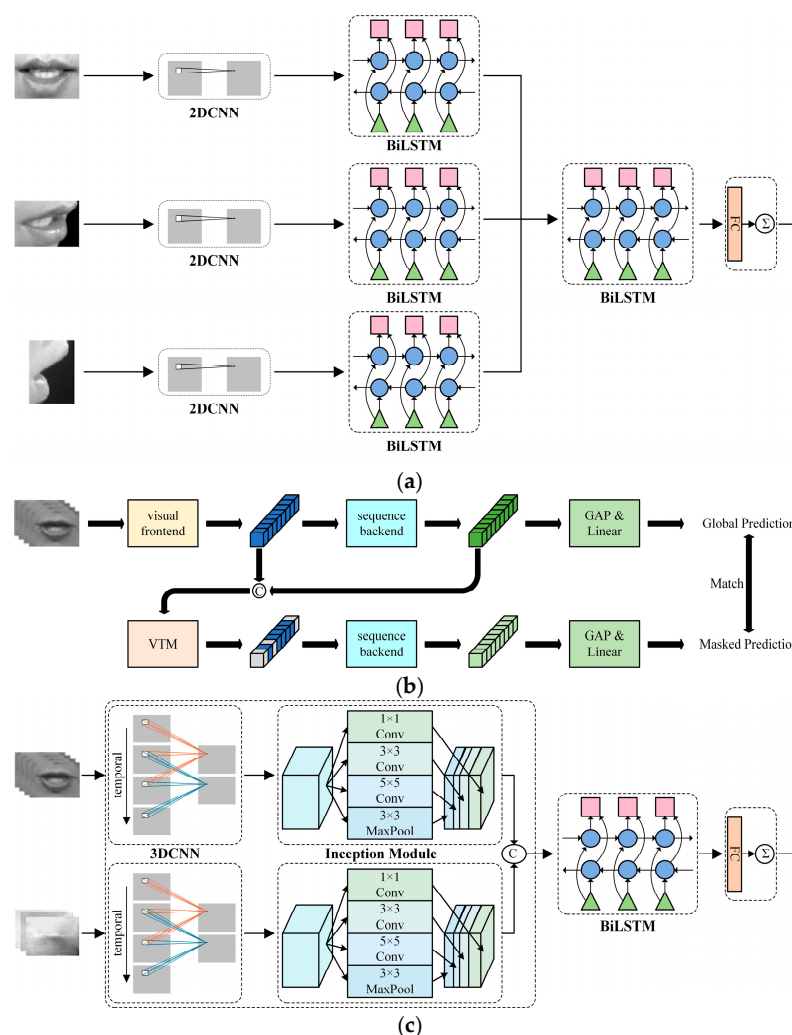


Figure 14. Cont.

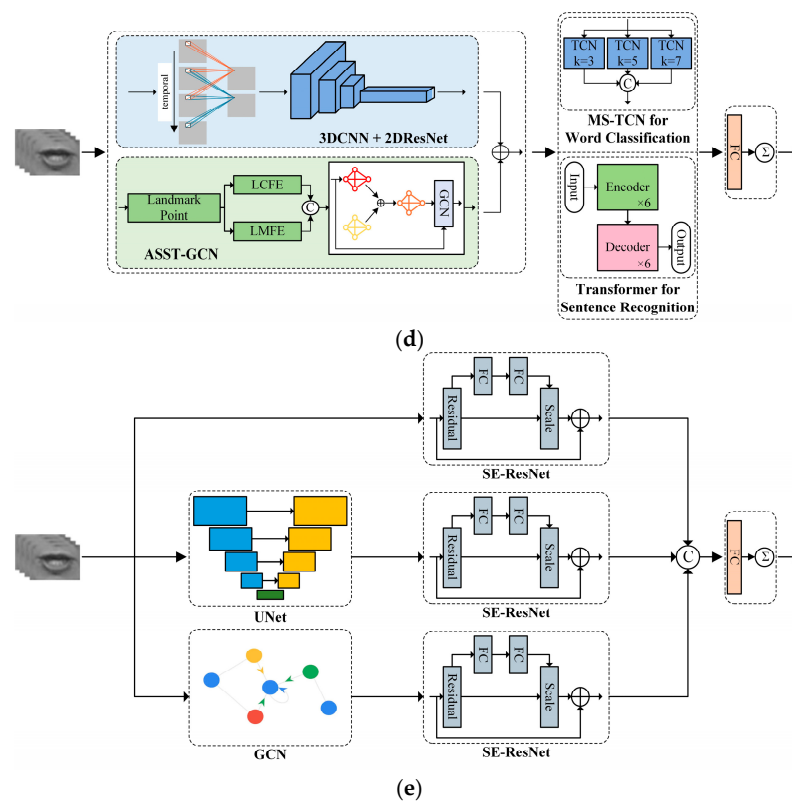


Figure 14. The other typical spatiotemporal feature enhancement methods. (a) The architecture of Ref. [126]. (b) The architecture of Ref. [128]. (c) The architecture of Ref. [129]. (d) The architecture of Ref. [130]. (e) The architecture of Ref. [131].

Graph convolution is a neural network operation consisting of the learning of features by aggregating information, the relational structure, and connectivity inherent in graph-structured data. Due to its superiority in handling irregular data over CNN, graph convolution is used to learn more discriminative shape-based visual features from a graph adjacency matrix defined by the manifold distance between nodes between lip key points and their interrelationships [132]. In fact, in this model, such predefined lip connections, which are meticulously engineered rather than naturally occurring, greatly restrict the representational capacity of the complex variations and dynamics of lip movements, undermining its effectiveness in tasks such as lip-reading or facial expression analysis.

Further advancements in graph convolution were made by Sheng et al. [130], where a three-branch feature enhancement method, consisting of global flow, local flow, and audio flow, was proposed to extract and fuse global spatial features from video sequences, and temporal-spatial and speech features from the lip graph sequence. For the dynamic lip shape modeling task, a graph convolution network is designed by embedding an adaptive semantic spatiotemporal graph convolution (ASST-GCN) module for processing eight subgraphs. By combining the multi-modal information, the model improved the accuracy in identifying dynamic lip movement.

In the same year, studies by Zhang et al. [131] showcased the effectiveness of the graph structure and lip shape segmentation network across adjacent frames for extracting content-independent features. This innovative approach utilized a segmentation network to isolate the region of interest (ROI) corresponding to the lip shapes. They constructed a local feature extractor based on U-Net, alongside a graph-based extractor for neighboring features, effectively merging the outputs of these two distinct feature extraction mechanisms to establish enhanced features. Moreover, they devised a three-tier feature fusion strategy comprising feature-level, score-level, and decision-level integration. Experimental results

demonstrated that this method possesses content-agnostic feature extraction capabilities, showcasing its robustness in diverse applications.

Table 3 presents a comprehensive comparative analysis of various spatiotemporal feature enhancement methods, illuminating their unique characteristics and performance metrics to enhance understanding of their contributions to diverse fields. In the table, “Structural Complexity” refers to the model structural complexity of spatiotemporal feature extraction methods, which is used to measure the depth, width, and number of sub-modules of the model, and is usually related to the parameter count of the model. Models with complex structures usually have more parameters, while models with simple structures usually have fewer parameters.

Table 3. Comparison of spatiotemporal feature enhancement methods.

Methods	Typical Networks	Elements	Enhancement Spatial Feature	Ability Temporal Feature	Recognition Rate	Structure Complexity	Parameter
spatial features [63–87,133]	ResNet, DenseNet, ShuffleNet	Convolution block, residual block	Weak	Weak	Low	Low	Small
temporal convolution [88–97]	ACNet, TSM, GSF, TBC	Temporal convolution block	Weak	Strong	High	High	Small
attention [98–111]	CBAM, BAM, Transformer	Attention block	Strong	Strong	High	High	Large
event cameras [112–115]	Event streams	High-speed event stream, Low-speed event stream	Strong	Relatively Strong	High	High	Large
spiking neural network [13,14,55,116–119]	Spiking Neural Networks (SNNs)	Event streams, Pulse signal	Relatively Strong	Strong	High	Low	Fewer
audio-visual assisting [82,120–125]	Audio-visual complementation	Visual processing unit, audio processing unit	Strong	Strong	High	High	Huge
others [126–132]	graph, moment, filter, variational time masking	Graph convolution, moment feature, filter, variational encoder	Relatively Strong	Relatively Strong	High	-	Large

4. Difficulties and Challenges

Despite significant advancements in spatiotemporal feature enhancement within lip-reading systems, a multitude of challenges, such as feature alignment, visual ambiguity, semantic integrity, and feature redundancy, remain, necessitating ongoing efforts to refine these techniques.

4.1. Feature Alignment

Spatiotemporal feature alignment: A foundational challenge in achieving effective lip-reading lies in the alignment of spatiotemporal features. The interdependence and integration of temporal and spatial information are essential for accurate feature representation. That means that disruptions in this relationship readily lead to significant misalignments, resulting in inadequate representation of the lip movements and thus diminishing the model’s efficacy. Therefore, aligning spatiotemporal features during both extraction and enhancement is crucial for the success of lip-reading.

Phoneme-to-viseme mappings: In the realm of audiovisual lip-reading, since phonemes and visemes are often processed separately, a substantial lack of coherence arises between the features associated with these two modes, which significantly and negatively impacts recognition performance. Consequently, the precise alignment of phonetic elements and visual elements is also of paramount importance for lip-reading as well as aligning spatiotemporal features, especially for recognition of subtle lip movements.

Speaker alignment: This is particularly challenging given the considerable variability among speakers in facial attributes, such as skin tone, head and mouth morphology, inter-organ distances, and posture, as well as differences in speech characteristics like pronunciation tendencies and speech rate.

4.2. Visual Ambiguity

Once feature alignment is established, the next challenge that emerges is visual ambiguity, particularly due to the prevalence of homophones—words that produce identical lip shapes but convey distinct meanings. This ambiguity arises from two main factors: the identical lip shapes of words like “their” and “there”, and the similarity in lip shapes across various phonemes, despite underlying differences in tongue position, such as with “c”, “d”, and “e”. This existence of these ambiguities complicates the recognition process, as accurate lip-reading hinges on differentiating between visually similar yet semantically diverse expressions. Thus, resolving visual ambiguity is fundamental for promoting the accuracy and reliability of spatiotemporal feature recognition.

4.3. Semantic Integrity

Closely related to the challenge of visual ambiguity is the issue of semantic integrity in the recognition process. The dimensions of lip-reading images are typically limited, which creates obstacles in constructing complex feature enhancement networks capable of capturing deeper semantic meanings.

While it is relatively straightforward to extract shallow spatial features, capturing comprehensive semantic features remains elusive. This limitation further exacerbates the challenges posed by visual ambiguities and misaligned features, as a lack of robust semantic representation can lead to inadequate distinctions between similar lip shapes. Therefore, developing strategies for the accurate and effective extraction of deep semantic features is another vital aspect of enhancing the integrity of spatiotemporal characteristics.

4.4. Feature Redundancy

Finally, alongside the aforementioned challenges, feature redundancy poses an additional significant hurdle in spatiotemporal feature enhancement. To improve feature augmentation, these networks often rely on vast numbers of model parameters and process enormous amounts of spatiotemporal information, resulting in a proliferation of enhanced features. However, only a small fraction of these features typically contains critical discriminative information, with the remainder being redundant. This redundancy not only hampers the efficiency of the model but also complicates the extraction and alignment processes. Therefore, addressing the challenge of eliminating redundant features and reducing the overall model parameters becomes an essential concern in the pursuit of effective spatiotemporal feature enhancement.

In summary, while advancements in spatiotemporal feature enhancement have been made, challenges in feature alignment, visual ambiguity, semantic integrity, and feature redundancy persist. These aspects are all interrelated, and their resolution is essential for refining lip-reading systems and achieving more accurate and robust performance. Ongoing research and innovation are necessary to navigate these complexities.

5. Research Trends

From the above analysis of the myriad challenges faced by spatiotemporal feature enhancement methods in lip-reading, we can envision several promising research directions that may pave the way for future advancements in this field.

5.1. Development of Time Convolution

The advancement of time convolution techniques is pivotal for precise extraction of temporal features, especially essential for capturing the dynamics of lip movements. Techniques such as ACNet, Temporal Shift Module (TSM), Grouped Spatial Filtering (GSF), and Tied Block Convolution have emerged as promising methods for temporal feature enhancement. However, challenges remain in optimizing these techniques for real-time processing. The complexity of accurately modeling temporal dependencies within lip movements necessitates significant computational resources, which can hinder performance in practical applications. Therefore, future research must focus on refining these algorithms to balance accuracy and efficiency, ensuring that they can operate effectively in dynamic environments where lip movements occur rapidly and with varying intensity.

5.2. Integration of Event Cameras and SNN

The integration of event cameras and SNN offers a novel concept for processing event-driven data for enhancing spatiotemporal feature extraction. Event cameras capture changes in scenes at high temporal resolutions, providing a wealth of information about lip movements. However, the challenge lies in effectively processing this event-driven data with SNNs, which operate on different principles compared to traditional neural networks. Researchers must develop methods to bridge the gap between the data generated by event cameras and the computational requirements of SNNs. This includes addressing issues related to data representation, noise reduction, and the design of SNN architectures that can efficiently learn from event-based inputs. As this integration unfolds, it promises to unlock new levels of responsiveness and accuracy in lip-reading systems.

5.3. Multimodal Approaches

Leveraging multimodal data—such as images, text, video, audio, and event streams—is a critical research direction aimed at developing recognition models that exploit the complementary advantages of different modalities. The challenge here lies in effectively fusing these diverse data types to enhance recognition accuracy. Techniques such as phoneme-to-viseme transformation and phoneme-visual fusion must be refined to ensure that the integration of modalities does not introduce noise or ambiguity. In addition, by combining visual and textual modalities through methods such as phoneme voxel conversion and phoneme voxel fusion, the ambiguity problem of homophones can be better resolved. Furthermore, establishing robust frameworks for synchronizing data from various sources is essential. Future research should focus on developing sophisticated algorithms that can dynamically adjust to the strengths of each modality, thereby creating a cohesive recognition system that capitalizes on the rich information provided by multimodal inputs.

5.4. Large Models

The development and fine-tuning of large models for lip-reading aim to improve performance by leveraging extensive datasets. While large models have shown promise in enhancing recognition capabilities, they also present challenges related to training complexity and resource consumption. The sheer size of these models can lead to overfitting, especially when the available training data are limited or not sufficiently diverse. Future research should explore strategies for regularization and transfer learning, encouraging large models to generalize better across different contexts and speaker variations.

5.5. Lightweight Models

In contrast to large models, the development of efficient, lightweight lip-reading models is essential for deployment on edge devices, where computational resources are

limited. The primary challenge is to ensure real-time performance while maintaining high accuracy with fewer parameters. Researchers must innovate in model architecture, exploring techniques such as model pruning, quantization, and knowledge distillation to reduce the size and complexity of the models. Striking a balance between efficiency and performance is critical, as lightweight models must still capture the intricate details of lip movements without sacrificing recognition quality. This area of research is particularly important as lip-reading technology becomes increasingly integrated into mobile devices and real-time applications.

5.6. Multi-Language Models

Expanding lip-reading to support multiple languages addresses the needs of diverse global communication scenarios. However, this endeavor presents significant challenges, particularly in capturing the linguistic nuances and phonetic variations inherent in different languages. Future research must focus on developing models that generalize for diverse languages, constructing multilingual datasets that encompass a wide range of accents and dialects. By addressing these challenges, researchers can enhance the applicability of lip-reading technology in multilingual contexts, fostering more inclusive communication solutions.

In conclusion, the evolving landscape of lip-reading technology highlights the need for innovative feature enhancement strategies to tackle the complexities of real-world applications. By addressing the challenges associated with the development of time convolution, the integration of event cameras and SNNs, multimodal approaches, large and lightweight models, and multi-language capabilities, researchers can significantly advance this field. Each of these directions not only holds the potential for improved recognition performance but also contributes to the broader goal of making lip-reading technology more accessible and effective in diverse communication scenarios.

6. Conclusions

Spatiotemporal feature enhancement is a key technology driving improvements in lip-reading performance. In this paper, we provide a comprehensive overview of the latest advancements in spatiotemporal feature enhancement for lip-reading. Notably, our analyses reveal a promising trend: the shift from enhancing visual features alone to a more integrated approach that includes spatial, temporal, and motion features. This evolution is especially evident in the context of increasingly complex tasks, significantly advancing lip-reading technology.

Moreover, the shift from manual to automated enhancement techniques marks a significant milestone in spatiotemporal feature enhancement, broadening research perspectives and offering key insights for selecting appropriate methods. The categorization of lip-reading methods by feature types provides a comprehensive analysis of the technology's evolution and establishes a solid foundation for future research in the field.

Author Contributions: Y.M. and X.S. conceived the idea. Y.M. searched and analyzed the literature. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Special Project of the National Natural Science Foundation of China (62441614), Anhui Province Key R&D Program (202304a05020068), and General Programmer of the National Natural Science Foundation of China (62376084).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

- Chen, H.; Li, W.; Cheng, Z.; Liang, X.; Zhang, Q. TCS-LipNet: Temporal & Channel & Spatial Attention-Based Lip Reading Network. In Proceedings of the International Conference on Artificial Neural Networks, Heraklion, Crete, Greece, 26–29 September 2023; pp. 413–424.
- Ma, P.; Wang, Y.; Petridis, S.; Shen, J.; Pantic, M. Training Strategies for Improved Lip-Reading. In Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 8472–8476.
- Preethi, S.J. Analyzing lower half facial gestures for lip reading applications: Survey on vision techniques. *Comput. Vis. Image Underst.* **2023**, *233*, 103738.
- Fenghour, S.; Chen, D.; Guo, K.; Li, B.; Xiao, A.P. Deep Learning-Based Automated Lip-Reading: A Survey. *IEEE Access* **2021**, *9*, 3107946.
- Pu, G.; Wang, H. Review on research progress of machine lip reading. *Vis. Comput.* **2023**, *39*, 3041–3057.
- Lu, Y.; Yan, J.; Gu, K. Review on Automatic Lip Reading Techniques. *Int. J. Pattern Recognit. Artificial Intell.* **2018**, *32*, 1856007.
- Sheng, C.; Kuang, G.; Bai, L.; Hou, C.; Guo, Y.; Xu, X.; Pietikäinen, M.; Liu, L. Deep Learning for Visual Speech Analysis: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 6001–6022.
- Prabhavalkar, R.; Hori, T.; Sainath, T.N.; Schlüter, R.; Watanabe, S. End-to-End Speech Recognition: A Survey. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 325–351.
- Santos, C.; Cunha, A.; Coelho, P. A Review on Deep Learning-Based Automatic Lipreading. In Proceedings of the International Conference on Wireless Mobile Communication and Healthcare, Vila Real, Portugal, 29–30 November 2023; pp. 180–195.
- Oghbaie, M.; Sabaghi, A.; Hashemifard, K.; Akbari, M. Advances and Challenges in Deep Lip Reading. *arXiv* **2021**, arXiv:2110.07879.
- Wang, C. Multi-grained spatio-temporal modeling for lip-reading. *arXiv* **2019**, arXiv:1908.11618.
- Zhang, X.; Cheng, F.; Shilin, W. Spatio-Temporal Fusion Based Convolutional Sequence Learning for Lip Reading. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 713–722.
- Liu, Q.; Ge, M.; Li, H. Intelligent event-based lip reading word classification with spiking neural networks using spatio-temporal attention features and triplet loss. *Inf. Sci.* **2024**, *675*, 120660.
- Dampfhofer, M.; Mesquida, T. Neuromorphic Lip-reading with signed spiking gated recurrent units. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–18 June 2024; pp. 2141–2151.
- Morade, S.S.; Patnaik, S. Visual Lip Reading using 3D-DCT and 3D-DWT and LSDA. *Int. J. Comput. Appl.* **2016**, *136*, 7–15.
- Morade, S.S.; Patnaik, S. Lip reading using DWT and LSDA. In Proceedings of the 2014 IEEE International Advance Computing Conference (IACC), Gurgaon, India, 21–22 February 2014; pp. 1013–1018.
- Potamianos, G.; Graf, H.P.; Cosatto, E. An image transform approach for HMM based automatic lipreading. In Proceedings of the 1998 International Conference on Image Processing, Chicago, IL, USA, 7 October 1998; pp. 173–177.
- Yu, K.; Jiang, X.; Bunke, H. Lipreading using Fourier transform over time. In Proceedings of the Computer Analysis of Images and Patterns, Kiel, Germany, 10–12 September 1997; pp. 472–479.
- Lan, Y.; Harvey, R.; Theobald, B.-J. Insights into machine lip reading. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 4825–4828.
- Deypir, M.; Alizadeh, S.; Zoughi, T.; Boostani, R. Boosting a multi-linear classifier with application to visual lip reading. *Expert Syst. Appl.* **2011**, *38*, 941–948. [\[CrossRef\]](#)
- Lin, B.-S.; Yao, Y.-H.; Liu, C.-F.; Lien, C.-F.; Lin, B.-S. Development of Novel Lip-Reading Recognition Algorithm. *IEEE Access* **2017**, *5*, 794–801. [\[CrossRef\]](#)
- Mase, K.; Pentland, A. Automatic Lip reading by optical flow analysis. *Syst. Comput. Jpn.* **1991**, *22*, 67–76. [\[CrossRef\]](#)
- Tamura, S.; Iwano, K.; Furui, S. Multi-Modal Speech Recognition Using Optical-Flow Analysis for Lip Images. *J. VLSI Signal Process. Syst. Signal Image Video Technol.* **2004**, *36*, 117–124. [\[CrossRef\]](#)
- Ma, X.; Yan, L.; Zhong, Q. Lip Feature Extraction Based on Improved Jumping-Snake Model. In Proceedings of the 2016 35th Chinese Control Conference (CCC), Chengdu, China, 27–29 July 2016; pp. 6928–6933.
- Wu, W.; Kuruoglu, E.E.; Wang, S.; Li, S.; Li, J. Automatic Lip Contour Extraction Using Both Pixel-Based and Parametric Models. *Chin. J. Electron.* **2013**, *22*, 76–82.
- Chen, J.; Tiddeman, B.; Zhao, G. Real-Time Lip Contour Extraction and Tracking Using an Improved Active Contour Model. In Proceedings of the 4th International Symposium (ISVC), Las Vegas, NV, USA, 1–3 December 2008; pp. 236–245.

27. Cootes, T.F.; Edwards, G.J.; Taylor, C.J. Active Appearance Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 681–685. [\[CrossRef\]](#)
28. Cootes, T.F.; Taylor, C.J.; Cooper, D.H.; Graham, J. Active Shape Models-Their Training and Application. *Comput. Vis. Image Underst.* **1995**, *61*, 38–59. [\[CrossRef\]](#)
29. Haque, S.; Togneri, R.; Bennamoun, M.; Sui, C. A lip extraction algorithm using region-based ACM with automatic contour initialization. In Proceedings of the 2013 IEEE Workshop on Applications of Computer Vision (WACV), Clearwater Beach, FL, USA, 15–17 January 2013; pp. 275–280.
30. Petajan, E.D. Automatic Lipreading to Enhance Speech Recognition (Speech Reading). Ph.D. Thesis, Electrical Engineering, University of Illinois at Urbana, Champaign, IL, USA, 1984; p. 261.
31. Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **1989**, *77*, 257–286. [\[CrossRef\]](#)
32. Matthews, I.; Cootes, T.F.; Bangham, J.A.; SCox, R.H. Extraction of visual features for lipreading. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 198–213. [\[CrossRef\]](#)
33. Ibrahim, M.Z.; Mulvaney, D.J. Geometrical-based lip-reading using template probabilistic multi-dimension dynamic time warping. *J. Vis. Commun. Image Represent.* **2015**, *30*, 219–233. [\[CrossRef\]](#)
34. Bregler, C.; Hild, H.; Manke, S.; Waibe, A. Improving connected letter recognition by lipreading. In Proceedings of the 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, Minneapolis, MN, USA, 27–30 April 1993; pp. 557–560.
35. Morade, S.S.; Patnaik, S. Comparison of classifiers for lip reading with CUAVE and TULIPS database. *Optik* **2015**, *126*, 5753–5761.
36. Pei, Y.; Kim, T.-K.; Zha, H. Unsupervised Random Forest Manifold Alignment for Lipreading. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 129–136.
37. Rathod, S.B.; Mahajan, R.A.; Agrawal, P.; Patil, R.R.; Verma, D.A. Enhancing Lip Reading: A Deep Learning Approach with CNN and RNN Integration. *J. Electr. Syst.* **2024**, *20*, 463–471.
38. Wang, H.; Pu, G.; Chen, T. A lip reading method based on 3D convolutional vision transformer. *IEEE Access* **2022**, *10*, 77205–77212.
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
40. Huang, G.; Liu, Z.; Maaten LV, D.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
41. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 1–11.
42. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
43. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
44. Schmidt, R.M. Recurrent Neural Networks (RNNs): A gentle Introduction and Overview. *arXiv* **2019**, arXiv:1912.05911.
45. Hochreiter, S.; Schmidhuber, J. Long Short-Term memory. *Neural Comput.* **1997**, *9*, 1735–1780.
46. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:1412.3555.
47. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In Proceedings of the 23rd International Conference on Machine Learning (ICML), Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376.
48. Lin, J.; Gan, C.; Han, S. Tsm: Temporal shift module for efficient video understanding. In Proceedings of the IEEE/CVF international conference on computer vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7083–7093.
49. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
50. Ni, R.; Jiang, H.; Zhou, L.; Lu, Y. Lip Recognition Based on Bi-GRU with Multi-Head Self-Attention. In Proceedings of the International Conference on Artificial Intelligence Applications and Innovations, Corfu, Greece, 27–30 June 2024; pp. 99–110.
51. Miled, M.; Messaoud, M.a.B.; Bouzid, A. Lip reading of words with lip segmentation and deep learning. *Multimed. Tools Appl.* **2022**, *82*, 551–571.
52. Atila, Ü.; Sabaz, F. Turkish lip-reading using Bi-LSTM and deep learning models. *Eng. Sci. Technol. Int. J.* **2022**, *35*, 101206.
53. Ma, P.; Wang, Y.; Shen, J.; Petridis, S.; Pantic, M. Lip-reading with Densely Connected Temporal Convolutional Networks. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 2856–2865.
54. Jeon, S.; Kim, M.S. End-to-End Sentence-Level Multi-View Lipreading Architecture with Spatial Attention Module Integrated Multiple CNNs and Cascaded Local Self-Attention-CTC. *Sensors* **2022**, *9*, 3597.

55. Liu, Q.; Xing, D.; Tang, H.; Ma, D.; Pan, G. Event-based action recognition using motion information and spiking neural networks. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, Montreal-themed Virtual Reality, Montreal, QC, Canada, 19–26 August 2021; pp. 1743–1749.
56. Kasabov, N.; Capecci, E. Spiking neural network methodology for modelling, classification and understanding of EEG spatio-temporal data measuring cognitive processes. *Inf. Sci.* **2015**, *294*, 565–575. [\[CrossRef\]](#)
57. Luo, M.; Yang, S.; Chen, X.; Liu, Z.; Shan, S. Synchronous bidirectional learning for multilingual lip reading. *arXiv* **2020**, arXiv:2005.03846.
58. Zhang, X.; Zhang, C.; Sui, J.; Sheng, C.; Deng, W.; Liu, L. Boosting lip reading with a multi-view fusion network. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 18–22 July 2022; pp. 1–6.
59. Xiao, Y.; Teng, L.; Liu, X.; Zhu, A. Exploring complementarity of global and local information for effective lip reading. *J. Electron. Imaging* **2023**, *32*, 023001. [\[CrossRef\]](#)
60. Chen, H.; Wang, Q.; Du, J.; Wan, G.-S.; Xiong, S.-F.; Yin, B.-C.; Pan, J.; Lee, C.-H. Collaborative Viseme Subword and End-to-end Modeling for Word-level Lip Reading. *IEEE Trans. Multimed.* **2024**, *26*, 9358–9371. [\[CrossRef\]](#)
61. Chen, H.; Du, J.; Hu, Y.; Dai, L.-R.; Lee, C.-H.; Yin, B.-C. Lip-reading with hierarchical pyramidal convolution and self-attention. *arXiv* **2020**, arXiv:2012.14360.
62. Stafylakis, T.; Tzimiropoulos, G. Combining residual networks with LSTMs for lipreading. *arXiv* **2017**, arXiv:1703.04105.
63. Mudaliar, N.K.; Hegde, K.; Ramesh, A.; Patil, V. Visual speech recognition: A deep learning approach. In Proceedings of the 2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 10–12 June 2020; pp. 1218–1221.
64. Fenghour, S.; Chen, D.; Guo, K.; Xiao, P. Lip reading sentences using deep learning with only visual cues. *IEEE Access* **2020**, *8*, 215516–215530. [\[CrossRef\]](#)
65. Wu, Z.; Chen, W.; Xu, J.; Wang, Y. Lip Reading Based on 3D Face Modeling and Spatial Transformation Learning. In Proceedings of the 2021 7th International Conference on Computer and Communications (ICCC), Chengdu, China, 10–13 December 2021; pp. 965–969.
66. Jiang, J.; Zhao, Z.; Yang, Y.; Tian, W. GSLip: A Global Lip-Reading Framework with Solid Dilated Convolutions. In Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan, 30 June–5 July 2024; pp. 1–8.
67. Arakane, T.; Saitoh, T. Efficient DNN Model for Word Lip-Reading. *Algorithms* **2023**, *16*, 269. [\[CrossRef\]](#)
68. Xiao, J.; Yang, S.; Zhang, Y.; Shan, S.; Chen, X. Deformation flow based two-stream network for lip reading. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 16–20 November 2020; pp. 364–370.
69. El-Bialy, R.; Chen, D.; Fenghour, S.; Hussein, W.; Xiao, P.; Karam, O.H.; Li, B. Developing phoneme-based lip-reading sentences system for silent speech recognition. *CAAI Trans. Intell. Technol.* **2023**, *8*, 129–138. [\[CrossRef\]](#)
70. Zeng, Q.; Du, J.; Wang, Z. HMM-based Lip Reading with Stingy Residual 3D Convolution. In Proceedings of the 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, 14–17 December 2021; pp. 1438–1443.
71. Sun, B.; Xie, D.; Luo, D.; Yin, X. A Lipreading Model Based on Fine-Grained Global Synergy of Lip Movement. In Proceedings of the 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI), Macao, China, 31 October–2 November 2022; pp. 848–854.
72. Huang, A.; Zhang, X. Dual-flow Spatio-temporal Separation Network for Lip Reading. *J. Phys. Conf. Ser.* **2022**, *2400*, 012028. [\[CrossRef\]](#)
73. Xu, K.; Li, D.; Cassimatis, N.; Wang, X. LCANet: End-to-end lipreading with cascaded attention-CTC. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 548–555.
74. He, L.; Ding, B.; Wang, H.; Zhang, T. An optimal 3D convolutional neural network based lipreading method. *IET Image Process.* **2022**, *16*, 113–122. [\[CrossRef\]](#)
75. Bi, C.; Zhang, D.; Yang, L.; Chen, P. An lipreading Modle with DenseNet and E3D-LSTM. In Proceedings of the 2019 6th International Conference on Systems and Informatics (ICSAI), Shanghai, China, 2–4 November 2019; pp. 511–515.
76. Jeon, S.; Kim, M.S. End-to-end lip-reading open cloud-based speech architecture. *Sensors* **2022**, *22*, 2938. [\[CrossRef\]](#) [\[PubMed\]](#)
77. Wen, J.; Lu, Y. Automatic lip reading system based on a fusion lightweight neural network with Raspberry Pi. *Appl. Sci.* **2019**, *9*, 5432. [\[CrossRef\]](#)
78. Jeevakumari, S.A.; Dey, K. LipSyncNet: A Novel Deep Learning Approach for Visual Speech Recognition in Audio-Challenged Situations. *IEEE Access* **2024**, *12*, 110891–110904. [\[CrossRef\]](#)
79. Zhang, G.; Lu, Y. Research on a Lip Reading Algorithm Based on Efficient-GhostNet. *Electronics* **2023**, *12*, 1151. [\[CrossRef\]](#)
80. Fu, Y.; Lu, Y.; Ni, R. Chinese lip-reading research based on ShuffleNet and CBAM. *Appl. Sci.* **2023**, *13*, 1106. [\[CrossRef\]](#)

81. Li, Y.; Hashim, A.S.; Lin, Y.; Nohuddin, P.N.; Venkatachalam, K.; Ahmadian, A. AI-based visual speech recognition towards realistic avatars and lip-reading applications in the metaverse. *Appl. Soft Comput.* **2024**, *164*, 111906. [\[CrossRef\]](#)
82. Sumanth, S.; Jyosthana, K.; Reddy, J.K.; Geetha, G. Computer Vision Lip Reading(CV). In Proceedings of the 2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC), Bhubaneswar, India, 19–20 November 2022; pp. 1–6.
83. Koumparoulis, A.; Potamianos, G. Accurate and resource-efficient lipreading with efficientnetv2 and transformers. In Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 8467–8471.
84. Sarhan, A.M.; Elshennawy, N.M.; Ibrahim, D.M. HLR-net: A hybrid lip-reading model based on deep convolutional neural networks. *Comput. Mater. Contin.* **2021**, *68*, 1531–1549. [\[CrossRef\]](#)
85. Tian, W.; Zhang, H.; Peng, C.; Zhao, Z.-Q. Lipreading model based on whole-part collaborative learning. In Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 2425–2429.
86. Tung, H.; Tekin, R. New Feature Extraction Approaches Based on Spatial Points for Visual-Only Lip-Reading. *Trait. Signal* **2022**, *39*, 659–668. [\[CrossRef\]](#)
87. Peng, C.; Li, J.; Chai, J.; Zhao, Z.; Zhang, H.; Tian, W. Lip Reading Using Deformable 3D Convolution and Channel-Temporal Attention. In Proceedings of the International Conference on Artificial Neural Networks, Bristol, UK, 6–9 September 2022; pp. 707–718.
88. Tsourounis, D.; Kastaniotis, D.; Fotopoulos, S. Lip reading by alternating between spatiotemporal and spatial convolutions. *J. Imaging* **2021**, *7*, 91. [\[CrossRef\]](#) [\[PubMed\]](#)
89. Jeon, S.; Elsharkawy, A.; Kim, M.S. Lipreading Architecture Based on Multiple Convolutional Neural Networks for Sentence-Level Visual Speech Recognition. *Sensors* **2022**, *22*, 72. [\[CrossRef\]](#) [\[PubMed\]](#)
90. Huang, J.; Teng, L.; Xiao, Y.; Zhu, A.; Liu, X. Lip Reading Using Temporal Adaptive Module. In Proceedings of the International Conference on Neural Information Processing, Changsha, China, 20–23 November 2023; pp. 347–356.
91. Sun, B.; Xie, D.; Shi, H. MALip: Modal Amplification Lipreading based on reconstructed audio features. *Signal Process. Image Commun.* **2023**, *117*, 117002. [\[CrossRef\]](#)
92. Hao, M.; Mamut, M.; Yadikar, N.; Aysa, A.; Ubul, K. How to use time information effectively? Combining with time shift module for lipreading. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 7988–7992.
93. Li, H.; Mamut, M.; Yadikar, N.; Zhu, Y.; Ubul, K. Channel Enhanced Temporal-Shift Module for Efficient Lipreading. In Proceedings of the Chinese Conference on Biometric Recognition, Shanghai, China, 10–12 September 2021; pp. 474–482.
94. Wiriyathamabhum, P. SpotFast Networks with Memory Augmented Lateral Transformers for Lipreading. In Proceedings of the International Conference on Neural Information Processing, Bangkok, Thailand, 23–27 November 2020; pp. 554–561.
95. Assael, Y.M.; Shillingford, B.; Whiteson, S.; Freitas, N.D. Lipnet: End-to-end sentence-level lipreading. *arXiv* **2016**, arXiv:1611.01599.
96. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 221–231. [\[CrossRef\]](#)
97. Zhang, P.; Wang, D.; Lu, H.; Wang, H.; Ruan, X. Amulet: Aggregating multi-level convolutional features for salient object detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 202–211.
98. Haq, M.A.; Ruan, S.-J.; Cai, W.-J.; Li, L.P.-H. Using lip reading recognition to predict daily Mandarin conversation. *IEEE Access* **2022**, *10*, 53481–53489. [\[CrossRef\]](#)
99. Lu, Y.; Xiao, Q.; Jiang, H. A Chinese Lip-Reading System Based on Convolutional Block Attention Module. *Math. Probl. Eng.* **2021**, *2021*, 6250879. [\[CrossRef\]](#)
100. Pan, X.; Chen, P.; Gong, Y.; Zhou, H.; Wang, X.; Lin, Z. Leveraging unimodal self-supervised learning for multimodal audio-visual speech recognition. *arXiv* **2022**, arXiv:2203.07996.
101. Cheng, X.; Jin, T.; Li, L.; Lin, W.; Duan, X.; Zhao, Z. Opensr: Open-modality speech recognition via maintaining multi-modality alignment. *arXiv* **2023**, arXiv:2306.06410.
102. Wang, H.; Cui, B.; Yuan, Q.; Pu, G.; Liu, X.; Zhu, J. Mini-3DCvT: A lightweight lip-reading method based on 3D convolution visual transformer. *Vis. Comput.* **2025**, *41*, 1957–1969. [\[CrossRef\]](#)
103. Li, Z.; Lohrenz, T.; Dunkelberg, M.; Fingscheidt, T. Transformer-Based Lip-Reading with Regularized Dropout and Relaxed Attention. In Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, 9–12 January 2023; pp. 723–730.
104. Prajwal, K.R.; Afouras, T.; Zisserman, A. Sub-word level lip reading with visual attention. In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5162–5172.

105. Elashmawy, S.; Ramsis, M.; Eraqi, H.M.; Eldeshnawy, F.; Mabrouk, H.; Abugabal, O.; Sakr, N. Spatio-temporal attention mechanism and knowledge distillation for lip reading. *arXiv* **2021**, arXiv:2108.03543.
106. Yu, W.; Zeiler, S.; Kolossa, D. Reliability-based large-vocabulary audio-visual speech recognition. *Sensors* **2022**, *22*, 5501. [[CrossRef](#)] [[PubMed](#)]
107. Li, D.; Gao, Y.; Zhu, C.; Wang, Q.; Wang, R. Improving speech recognition performance in noisy environments by enhancing lip reading accuracy. *Sensors* **2023**, *23*, 2053. [[CrossRef](#)]
108. Varshney, M.; Yadav, R.; Namboodiri, V.P.; Hegde, R.M. Learning speaker-specific lip-to-speech generation. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 491–498.
109. Yang, W.; Li, P.; Yang, W.; Liu, Y.; He, Y.; Petrosian, O.; Davydenko, A. Research on robust audio-visual speech recognition algorithms. *Mathematics* **2023**, *11*, 1733. [[CrossRef](#)]
110. Lohrenz, T.; Möller, B.; Li, Z.; Fingscheidt, T. Relaxed attention for transformer models. In Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, QLD, Australia, 18–23 June 2023; pp. 1–10.
111. Ma, P.; Petridis, S.; Pantic, M. Visual speech recognition for multiple languages in the wild. *Nat. Mach. Intell.* **2022**, *4*, 930–939.
112. Tan, G.; Wang, Y.; Han, H.; Cao, Y.; Wu, F.; Zha, Z.-J. Multi-grained spatio-temporal features perceived network for event-based lip-reading. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20094–20103.
113. Kanamaru, T.; Arakane, T.; Saitoh, T. Isolated single sound lip-reading using a frame-based camera and event-based camera. *Front. Artif. Intell.* **2023**, *5*, 1070964. [[CrossRef](#)]
114. Tan, G.; Wan, Z.; Wang, Y.; Cao, Y.; Zha, Z.-J. Tackling Event-Based Lip-Reading by Exploring Multigrained Spatiotemporal Clues. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, 1–13.
115. Zhang, W.; Wang, J.; Luo, Y.; Yu, L.; Yu, W.; He, Z. MTGA: Multi-view Temporal Granularity aligned Aggregation for Event-based Lip-reading. *arXiv* **2024**, arXiv:2404.11979.
116. Li, X.; Neil, D.; Delbruck, T.; Liu, S.-C. Lip reading deep network exploiting multi-modal spiking visual and auditory sensors. In Proceedings of the 2019 IEEE International Symposium on Circuits and Systems (ISCAS), Sapporo, Japan, 26–29 May 2019; pp. 1–5.
117. Ning, L.; Dong, J.; Xiao, R.; Tan, K.C.; Tang, H. Event-driven spiking neural networks with spike-based learning. *Memetic Comput.* **2023**, *15*, 205–217.
118. Bulzomi, H.; Schweiker, M.; Gruel, A.E.; Martinet, J. End-to-end Neuromorphic Lip Reading. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023.
119. Yu, X.; Wang, L.; Chen, C.; Tie, J.; Guo, S. Multimodal Learning of Audio-Visual Speech Recognition with Liquid State Machine. In Proceedings of the International Conference on Neural Information Processing, Changsha, China, 20–23 November 2023; pp. 552–563.
120. Chung, J.S.; Zisserman, A. Learning to lip read words by watching videos. *Comput. Vis. Image Underst.* **2018**, *173*, 76–85.
121. Handa, A.; Agarwal, R.; Kohli, N. A multimodal keyword spotting system based on lip movement and speech features. *Multimed. Tools Appl.* **2020**, *79*, 20461–20481.
122. Kim, M.; Hong, J.; Park, S.J.; Ro, Y.M. CroMM-VSR: Cross-modal memory augmented visual speech recognition. *IEEE Trans. Multimed.* **2021**, *24*, 4342–4355.
123. Adeel, A.; Gogate, M.; Hussain, A.; Whitmer, W.M. Lip-reading driven deep learning approach for speech enhancement. *IEEE Trans. Emerg. Top. Comput. Intell.* **2019**, *5*, 481–490.
124. Afouras, T.; Chung, J.S.; Senior, A.; Vinyals, O.; Zisserman, A. Deep Audio-Visual Speech Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 8717–8727.
125. Kim, M.; Yeo, J.H.; Ro, Y.M. Distinguishing Homophones Using Multi-Head Visual-Audio Memory for Lip Reading. In Proceedings of the The Thirty-Sixth AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022.
126. Petridis, S.; Wang, Y.; Li, Z.; Pantic, M. End-to-End Multi-View Lipreading. In Proceedings of the British Machine Vision Conference (BMVC), London, UK, 4–7 September 2017; p. 161.
127. Mesbah, A.; Berrahou, A.; Hammouchi, H.; Berbia, H.; Qjidaa, H.; Daoudi, M. Lip reading with Hahn convolutional neural networks. *Image Vis. Comput.* **2019**, *88*, 76–83.
128. Sheng, C.; Liu, L.; Deng, W.; Bai, L.; Liu, Z.; Lao, S.; Kuang, G.; Pietikäinen, M. Importance-aware information bottleneck learning paradigm for lip reading. *IEEE Trans. Multimed.* **2022**, *25*, 6563–6574.
129. Weng, X.; Kitani, K. Learning Spatio-Temporal Features with Two-Stream Deep 3D CNNs for Lipreading. In Proceedings of the British Machine Vision Conference (BMVC), Cardiff, UK, 9–12 September 2019; pp. 1–13.
130. Sheng, C.; Zhu, X.; Xu, H.; Pietikäinen, M.; Liu, L. Adaptive semantic-spatio-temporal graph convolutional network for lip reading. *IEEE Trans. Multimed.* **2021**, *24*, 3545–3557. [[CrossRef](#)]

131. Zhang, C.; Zhao, H. Lip Reading using Local-Adjacent Feature Extractor and Multi-Level Feature Fusion. In Proceedings of the 2021 2nd International Conference on Computer Information and Big Data Applications, Wuhan, China, 26–28 March 2021; p. 012083.
132. Liu, H.; Chen, Z.; Yang, B. Lip Graph Assisted Audio-Visual Speech Recognition Using Bidirectional Synchronous Fusion. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 3520–3524.
133. Ma, P.; Martinez, B.; Petridis, S.; Pantic, M. Towards Practical Lipreading with Distilled and Efficient Models. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 7608–7612.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.