Adaptive Semantic-Spatio-Temporal Graph Convolutional Network for Lip Reading

Changchong Sheng, Xinzhong Zhu, Huiying Xu, Matti Pietikäinen, *Fellow, IEEE*, and Li Liu, *Senior Member, IEEE,*

Abstract—**The goal of this work is to recognize words, phrases, and sentences being spoken by a talking face without given the audio. Current deep learning approaches for lip reading focus on exploring the appearance and optical flow information of videos. However, these methods do not fully exploit the characteristics of lip motion. In addition to appearance and optical flow, the mouth contour deformation usually conveys significant information that is complementary to others. However, the modeling of dynamic mouth contour has received little attention than that of appearance and optical flow. In this work, we propose a novel model of dynamic mouth contours called Adaptive Semantic-Spatio-Temporal Graph Convolution Network (ASST-GCN), to go beyond previous methods by automatically learning both the spatial and temporal information from videos. To combine the complementary information from appearance and mouth contour, a two-stream visual front-end network is proposed. Experimental results demonstrate that the proposed method significantly outperforms the state-of-the-art lip reading methods on several large-scale lip reading benchmarks.**

Index Terms—**Lip Reading, Semantic-Spatio-Temporal, Adaptive Graph Convolution Network, Two-Stream.**

I. Introduction

AUTOMATIC Lip Reading (ALR), also known as Visual Speech Recognition (VSR), aims to decode the content of a speech from the speaker's mouth movements. ALR has been attracting increasing attention in recent years since it plays a significant role in many applications such as audio-video speech recognition (AVSR), health care, public security, and human-computer interaction [1], [2], [3], [4], [5]. Recently, advances in deep learning and the availability of large-scale datasets have brought significant progress for ALR [1], [6], [7], [8], [9], [10], [11]. However, the low accuracy of state-of-the-art ALR approaches is still far from meeting the requirements of real-world applications. ALR has many challenges [12], such as visual ambiguities, speaker dependency, pose variations, environment illumination, video resolution, *etc*.

Current deep learning based lip reading methods consist of two main sub-networks [9]: a visual front-end network for extracting spatio-temporal visual features and a sequence back-end network for modeling the temporal dependency from the extracted features. For the visual front-end part, two types of network architectures are commonly used, *i.e.*, a shallow 3D CNN + deep 2D CNNs (*e.g.*, ResNet [13]) and 3D CNNs (*e.g.*, I3D [14]). The low dimensional visual features are extracted directly with the front-end network followed by global average pooling [10]. The back-end part further explores temporal context information from the extracted visual features. Current sequence back-end networks are designed by borrowing ideas from speech recognition or natural language processing. They can be divided into three categories: Temporal Convolutional Networks (TCN) [10], [15], Recurrent Neural Network (RNN) [7], [10], [16], and Transformer [1], [11], [17].

Despite the recent progress, current visual front-end networks have the following shortcomings. Firstly, current methods do not fully exploit the characteristics of lip dynamics. They focus on exploring the appearance and optical flow information of videos [9]. In addition to appearance and optical flow, the mouth contour deformation usually conveys significant information that is complementary to others. Secondly, different parts of the talking mouth (mouth corner, teeth, chin, *etc.*) contain rich semantic information, which is critical for ALR. However, normal CNNs can hardly capture the complex semantic relationship among these local parts.

To address these drawbacks, in this paper, we aim to go beyond previous methods by explicitly modeling the dynamic mouth contours to capture the motion of mouth contour, local subtle movements, and semantic information contained in the underlying structure. To achieve this goal, several facial landmark points, named Lip Reading related Landmark Points (LRLPs), on the lip region [18] are selected to model the mouth contour deformation. Human landmark information has been widely explored for deep learning based human-centric understanding tasks. Such as visual fashion analysis [19], [20], human parsing [21], [22], [23], *etc*.

Recently, graph convolutional networks (GCNs), which extend convolution operations for graph data, have been successfully adopted in many applications [24], [25], [26], [27], [28], [29]. Inspired by the recent great success of GCNs, we propose to introduce GCNs to model the mouth contour deformation. However, there are some difficulties in the process of graph construction. (1) Different from ordinary graph data, there is

This work was partially supported by the Academy of Finland under grant 331883, Outstanding Talents of "Ten Thousand Talents Plan" in Zhejiang Province (project no. 2018R51001), and the Natural Science Foundation of China (project no. 61976196). (*Corresponding author: Xinzhong Zhu.*)

Changchong Sheng, Matti Pietikäinen and Li Liu are with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. (Email: Changchong.Sheng@oulu.fi; Matti.Pietikainen@oulu.fi; li.liu@oulu.fi)

Xinzhong Zhu is with the College of Mathematics, Physics and Information Engineering, Zhejiang Normal University, Jinhua 321004, China, and also with the Research Institute of Ningbo Cixing Co. Ltd, Ningbo 315336,China. (Email: zxz@zjnu.edu.cn)

Huiying Xu is an associate professor with the College of Mathematics and Computer Science, Zhejiang Normal University, and also the researcher of Research Institute of Ningbo Cixing Co. Ltd, China. (Email: xhy@zjnu.edu.cn)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3102433, IEEE Transactions on Multimedia

no natural connectivity between those LRLPs. If we introduce a predefined adjacency matrix, it is not guaranteed to be optimal for the lip reading task. (2) For deep GCNs, we believe that different layers contain different levels of relations. However, the topology of the graph applied in ordinary GCN is fixed over all the layers, which lacks the flexibility and capacity to model the multilevel relations contained in different layers [30]. (3) A single type of graph structure may not be sufficient for lip reading tasks. Both the natural semantic relations and the dynamic spatio-temporal relations of those LRLPs should be considered.

Motivated by this, a novel model of dynamic mouth contours, called Adaptive Semantic-Spatio-Temporal Graph Convolution Network (ASST-GCN), is proposed in this work. It parameterizes two kinds of adaptive graphs for graph convolution. One of them is referred as the semantic graph, which is obtained by learning the shared adjacency matrix from the datasets. Another is referred as the dynamic spatio-temporal graph, whose adjacency matrix is built based on the input data. Furthermore, to combine the complementary information from appearance and mouth contour, a two-stream visual front-end network is proposed. The major contributions of this work are summarized as follows.

- We propose a novel model of dynamic mouth contours called Adaptive Semantic-Spatio-Temporal Graph Convolution Network (ASST-GCN), which goes beyond previous methods by automatically learning both the semantic and spatio-temporal information from videos.
- In order to combine the complementary information from appearance and mouth contour, a two-stream visual front-end network is proposed.
- Experiments on both word-level and sentence-level lip reading tasks clearly show that the proposed method significantly outperforms the baseline lip reading methods on several large-scale lip reading benchmarks.

The rest of paper is organized as follows. Sec. II introduces the related work of lip reading and adaptive graph convolution network. Sec. III describes the overall pipeline of our lip reading model and the proposed ASST-GCN module. The components of our proposed ASST-GCN module are introduced in detail in Sec. III. The ablation study and the comparison with the state-of-the-art methods are shown in Sec. IV. Sec. IV also provides some quantitative results and discussions. Sec. V concludes the paper.

II. Related Work

A. Lip Reading

The input video of ALR contains a large amount of redundant information (such as pose, illumination, gender, *et al.*) that is unrelated to the ALR task, while the information really related to the ALR task is lip movement. The key to ALR tasks, spatio-temporal feature extraction, is to effectively filter out redundant information while keeping lip movement information as much as possible. Before the emergence of deep learning based methods, researchers did a lot of work on ALR research which mainly focuses on the spatio-temporal feature extraction of videos. There are two main types of traditional methods: appearance-based and shape-based. The former uses the pixel value of ROI as the original feature space, then utilizes different data dimension reduction methods to obtain compact and effective feature representations. For dimension reduction methods, linear transformation methods such as Principal Component Analysis (PCA) [31], Discrete Cosine Transform (DCT) [32], Linear Discriminant Analysis (LDA) [32] and Maximum Likelihood Linear Transformation (MLLT) [33] are commonly used; Besides, optical flow, Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) [34], manifold learning and graph embedding methods such Locality Discriminant Graph (LDG) [35], Random Forest Manifold Alignment (RFMA) [36] and so on are also used for feature extraction. Shape-based methods perform feature extraction based on the shape of the ROI (lips, chin, cheeks, *et al.*). Compared to appearance-based methods, those methods have better interpretability and generalization while needing more manual annotation. Main attributes of lip contour (height, width *et al.*) or Articulatory Features (AFs) [37], [38] are mainly used to small-scale recognition tasks; Active Shape Model (ASM) [39] is one of the most commonly used shape-based methods that use facial landmarks to extract spatio-temporal features. In addition, some researchers proposed a more powerful method, the Active Appearance Model (AAM) [40], that furtherly improve the performance by combining appearance-based and shape-based methods. For classifier, Support Vector Machine (SVM), template matching, Maximum a Posteriori (MAP), and Regularized Discriminant Analysis (RDA) are mainstream classifiers for isolated recognition tasks; Hidden Markov Model (HMM) is widely used for continuous recognition tasks.

The works on deep lip reading mainly focus on the architecture design of these two sub-networks: visual front-end networks and sequence back-end networks. As for the design of visual front-end networks, a lot of works utilize deep CNNs to perform visual features extraction. For example, [10] proposes a simple variation of ResNet (changing the first 2D convolution layer to 3D convolution layer). This model consists of a shallow 3D CNN and deep 2D CNN, and it achieves 83% recognition accuracy for word-level lip reading on LRW [8] dataset. Due to the considerable performance of the model, most lip reading models [1], [11], [41] adopt it as the backbone network for visual features extraction. Besides, deep 3D CNNs are also used to extract visual features. In paper [9], the authors successfully migrate the two-stream (the raw grayscale video stream and the dense optical flow stream) I3D model to lip reading, and achieve comparable performance on word-level lip reading. However, dense optical flow calculation is very time consuming, resulting in low recognition efficiency.

For the design of sequence processing back-end networks, there are two main lip reading tasks: word-level and sentence-level. The former is to recognize isolated words from the input videos, usually trained with multi-classification cross-entropy loss. Stafylakis *et al.* have created the baseline word-level lip reading model with BiLSTM based back-end network [10]. Martinez *et al.* improved the state-of-the-art model by replacing the BiLSTM back-end with Multi-Scale TCN

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3102433, IEEE Transactions on Multimedia

(MSTCN) [42]. The latter is to do sentence-level sequence prediction, both connectionist temporal classification loss (CTC) [43] and sequence-to-sequence loss [44] can be used to train the model. LipNet [7], consisting of 3D CNNs and BiGRUs, is the first end-to-end sentence-level lipreading model that simultaneously learns spatio-temporal visual features and sequence model. Besides, Afouras *et al*. introduce transformer self-attention architecture to lip reading. They propose Transformer-CTC model and Transformer-seq2seq model [1], and further discuss the difference between the two models in detail.

B. Adaptive Graph Convolution Networks

Encouraged by the great success of CNNs in the field of computer vision, extensive works further explore how to apply convolution operations to graph related data. The principle of constructing GCNs mainly follows two streams: spatial-based [28], [45], [46] and spectral-based [47], [48], [49], [29], [50].

Spectral-based methods treat graph convolution as graph signal processing. They assume graphs to be undirected, and perform graph convolution in the frequency domain with the favor of the graph Fourier transform. In contrast, spatial-based methods define graph convolution on graph nodes based on their spatial adjacency relations, just like CNNs on image pixels. This work follows the spatial-based methods.

Normal spatial-based GCNs are only built for known and fixed graph structure. However, in some cases (e.g. classification of point cloud, skeleton-based action recognition, human-centric understanding), the fixed graph structure may lack the flexibility and capacity to model the complex graph data. In [51], Qi *et al*. propose Graph Parsing Neural Network (GPNN) for human-object interaction (HOI) recognition. The proposed GPNN offers a general framework that explicitly represents HOI structures with graphs and automatically parses the optimal graph structures in an end-to-end manner. In [21], Fan *et al*. propose a novel spatio-temporal reasoning graph network for human gaze communication in social videos from both atomic-level and event-level. The proposed model can iteratively learns gaze communication structures and node representations. Besides, some researchers try to make the network adaptively learn the expanded graph structure. To adapt GCNs to arbitrary graph structure and size, Li *et al*. [52] propose an Adaptive Graph Convolutional Network module, called SGC-LL, to learn residual graph Laplacian via learning the optimal metric and feature transform. Shi *et al*. [53] propose an adaptive graph convolution neural network (2s-AGCN) for skeleton-based action recognition. They redefine the graph architecture of the skeleton data and embed it into the network parameters to be jointly learned and updated with the model.

C. Audio-Visual Cross-Modal Learning

Cross-modal learning from vision and audio has attracted increasing interest in recent years. Based on the natural co-occurring characteristics of audio and video, audio model and visual model can be jointly trained for diverse tasks. *e.g*., visual sound separation [54], [55], [56], visual music generation [57], [58], [59], visual sound localization [60], [61], *etc*.

Audio-visual cross-modal learning is also widely used on lip reading tasks. In [15], Afouras *et al*. propose a cross-modal distillation framework to train a lip reading model by distilling knowledge from a pre-trained ASR model. Based on the framework, unlabelled video data can be leveraged to improve lip reading performance further. Based on the natural synchronization characteristics of audio and video, sounds and lip movements can be treated as mutual supervisory signals. Motivated by this, a series of works try to learn discriminative visual representations for lip reading by cross-modal self-supervised learning [62], [63], [64]. Considering that the cost of large-scale lip reading annotation can be prohibitive, cross-modal self-supervised learning for lip reading has received a growing amount of attention due to its high label efficiency.

III. Proposed Methodology

In this section, we firstly describe the overall pipeline of lip reading models. Then we illustrate the design motivation of our two-stream front-end network. Finally, we introduce the novel proposed ASST-GCN module used in the local stream in detail.

A. The Overall Pipeline

Given the aligned and mouth-centered cropped input video, the objective of the visual front-end network is to extract visual spatio-temporal features representing visual speech patterns and dynamics. The visual front-end network has a relatively small receptive field on the temporal dimension, which is insufficient for ALR tasks. The sequence back-end network focuses on further aggregating long-term temporal contextual information.

Currently, relatively more attention has been paid to the sequence back-end networks, which are borrowed from the fields of speech recognition or natural language processing. The visual front-end, which learns discriminative spatio-temporal features, plays a critical role in ALR. However, the current visual front-end networks have some drawbacks that we have mentioned in Sec. I.

To address these issues, a two-stream front-end network is proposed in this paper, as summarized in Fig. 1. It consists of a ”local stream”, focusing on capturing semantic preserved lip contour information and local motion information around the lip, and a ”global stream”, aiming at modeling the global motion information of the lip. The features extracted by the two streams are concatenated to be input to the sequence back-end network. The individual modules are described in detail in the following subsections.

Our proposed two-stream front-end combines both holistic level and contour level features. The holistic level features from the global stream capture comprehensive information on the mouth region. However, it is also sensitive to the speech-unrelated redundant information that we have mentioned above. In contrast, contour features focus on the description of overall shape, and thus are more robust compared to global features. In other words, the information generated by these two streams is complementary to each other, so the

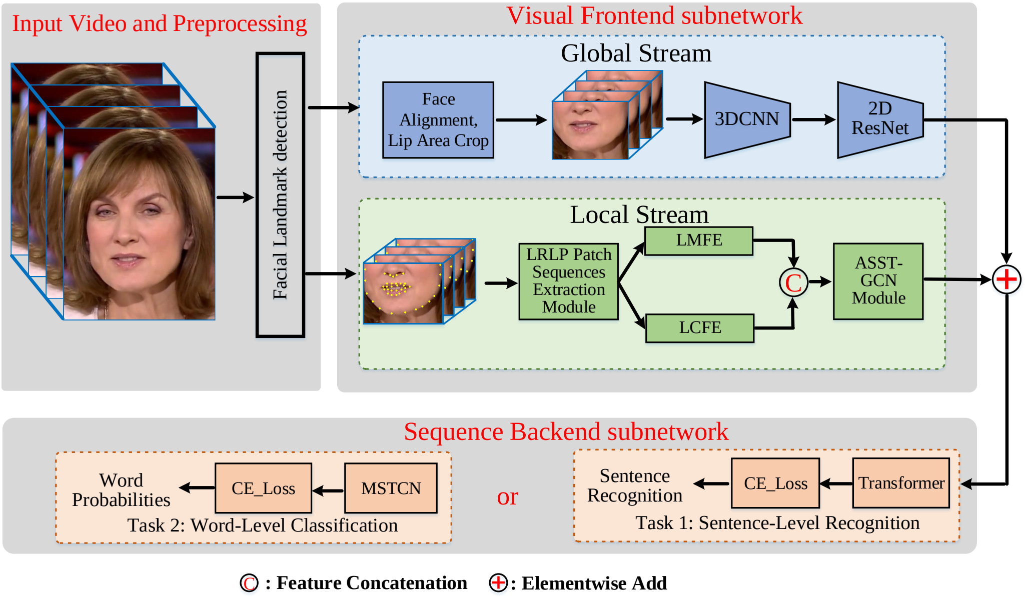

Input Video and Preprocessing
Facial Landmark detection
Visual Frontend subnetwork
Global Stream
Face Alignment, Lip Area Crop
3DCNN
2D ResNet
Local Stream
LRLP Patch Sequences Extraction Module
LMFE
LCFE
C
ASST-GCN Module
Sequence Backend subnetwork
Word Probabilities
CE_Loss
MSTCN
Task 2: Word-Level Classification
or
Sentence Recognition
CE_Loss
Transformer
Task 1: Sentence-Level Recognition
© : Feature Concatenation
⊕: Elementwise Add

Fig. 1. The overall framework of the proposed lip reading model. For the input video, facial landmark detection is first processed. Based on the video and landmark points, the global stream extracts global visual features from aligned and lip centered cropped video. Meanwhile, the local stream extracts local visual features from landmarks centred patch video and landmark coordinates. Then, the local and global visual features are concatenated together as the input of sequence back-end networks.

appropriate fusion of the two streams promotes robustness and accuracy.

Global Stream. The goal of the global stream in the visual front-end network is to extract global features of lip motions. This module consists of deep CNNs (C3D_ResNet18), which are directly applied to lip-centered video. The pooling layers are added to reduce the spatial size and improve the receptive field, which results in the extracted visual features contain more global lip motion information but little local subtle motion information and lip contour information. The layers are listed in full detail in Tab. I.

Layer Type	Filters	Output dimension
Conv3d	$5 \times 7 \times 7, 64, /[1, 2, 2]$	$T \times 64 \times \frac{H}{4} \times \frac{W}{4}$
MaxPool3d	$1 \times 3 \times 3/[1, 2, 2]$	
Residual Conv2d	$[3 \times 3, 64] \times 2/1$	$T \times 64 \times \frac{H}{4} \times \frac{W}{4}$
Residual Conv2d	$[3 \times 3, 64] \times 2/1$	
Residual Conv2d	$[3 \times 3, 128] \times 2/2$	$T \times 128 \times \frac{H}{8} \times \frac{W}{8}$
Residual Conv2d	$[3 \times 3, 128] \times 2/1$	
Residual Conv2d	$[3 \times 3, 256] \times 2/2$	$T \times 256 \times \frac{H}{16} \times \frac{W}{16}$
Residual Conv2d	$[3 \times 3, 256] \times 2/1$	
Residual Conv2d	$[3 \times 3, 512] \times 2/2$	$T \times 512 \times \frac{H}{32} \times \frac{W}{32}$
Residual Conv2d	$[3 \times 3, 512] \times 2/1$	
GlobalPool2d	-	$T \times 512 \times 1 \times 1$

TABLE I
ARCHITECTURE DETAILS FOR THE SPATIO-TEMPORAL VISUAL FRONT-END (C3D_RESNET18).

Local Stream. Complementary to the global stream, this stream is to learn local motion features based on these semantical local patches and lip contour points. It consists of several modules: LRLPs patch sequence extraction module, local motion feature extraction (LMFE) module, landmark coordinate feature extraction (LCFE) module, and ASST-GCN module, which will be presented in detail in the following subsections.

B. Modules

LRLP Patch Sequence Extraction. In the proposed framework, facial landmark detection frame-by-frame is essential for both the global and local streams. Facial landmark detection aims automatically identifying the locations of the facial key landmark points on facial images. Those key points are either the dominant points describing the unique location of a facial component (*i.e.*, eye corner, mouth corner) or an interpolated point connecting those dominant points around the facial components and facial contour [18]. In the global stream, Facial landmark points are only used to determine the mouth position for alignment and cropping.

In the local stream, for each video frame, a total of 68 facial landmark points are firstly detected with the facial landmark detection algorithms [65], [66]. K ($K = 38$ in this paper) facial landmark points from the lip region are selected as Lip Reading related Landmark Points (LRLPs), as illustrated in Fig. 2 (a). We empirically show that those 38 LRLPs (located in the the below half of the face) can sufficiently retain the visual features related to speech. Then centered at each facial landmark point, a local patch of size 32×32 pixels is extracted to describe this point. As a result, for an input gray video, we extract K LRLPs patch sequences which will be fed into later steps for further processing. This is clearly shown in Fig. 2 (a). In addition, the coordinates of each LRLP are extracted as well.

Local Motion Feature Extraction (LMFE). A lightweight 3D CNN is used to extract spatio-temporal feature vectors from the extracted LRLP patch sequences, as shown in Fig. 2 (b1). In specific, a 3-layer 3D CNN (More details are given in

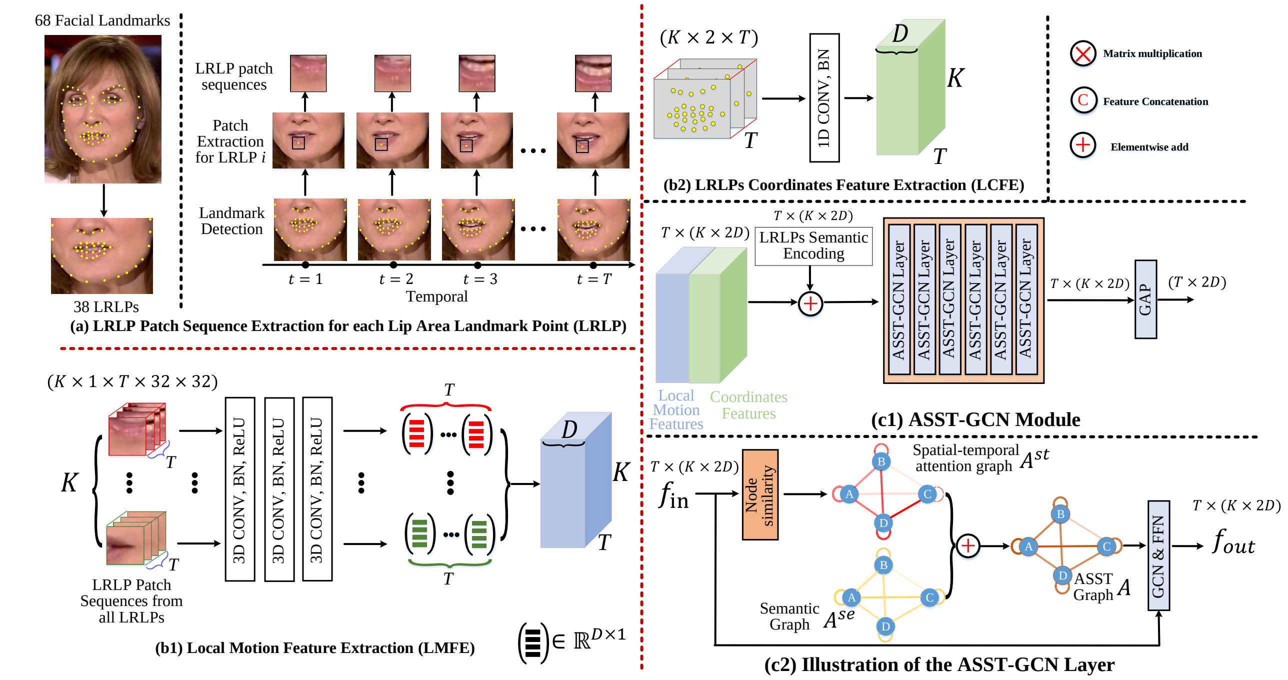


Fig. 2. Building blocks of the local stream in Fig. 1. **(a)** LRLPs patch sequence extraction module aims to crop landmark points-centred patch videos from raw input video. **(b1)** Local motion feature extraction module consists of 3 3D CNN layers, aiming to extract low dimension local features from the patch videos. **(b2)** Landmark coordinates feature extraction module extracts LRLPs motion features using a 1D CNN layer. **(c1)** The proposed ASST-GCN module, including 6 ASST-GCN layers. GAP represents the global average pooling layer. **(c2)** Simplified diagram of the ASST-GCN Layer. FFN means the feed-forward network. It contains two types of graph architectures: Semantic graph A^{se} and spatio-temporal attention graph A^{st}. Among them, A^{st} is calculated by the similarity of graph node representations, and A^{se} is obtained by training with random initialization parameters. The ASST graph structure is the sum of A^{se} and A^{st}. More details of the ASST-GCN layer are given in Fig. 3.

Tab. II) is applied on all LRLP patch sequences. By this means, each input LRLP patch sequence of T frames (suppose gray image here) of size $T \times 32 \times 32$ is transformed into a $D \times T$ ($D = 256$) dimensional feature vector. In other words, the total extracted K LRLP patch sequences result in K output feature vectors of D dimension, which are organized into a feature tensor of size $K \times D \times T$.

Layer Type	Filters	Output dimension
Conv3d MaxPool3d	$5 \times 7 \times 7, 64, /[1, 2, 2]$ $1 \times 3 \times 3/[1, 2, 2]$	$T \times 64 \times \frac{H}{4} \times \frac{W}{4}$
Conv3d	$1 \times 3 \times 3$, 128, /[1,2,2]	$T \times 128 \times \frac{H}{8} \times \frac{W}{8}$
Conv3d	$1 \times 3 \times 3$, 256, /[1,2,2]	$T \times 256 \times \frac{H}{16} \times \frac{W}{16}$
AveragePool2d	-	$T \times 256$

TABLE II
ARCHITECTURE DETAILS FOR THE 3-LAYER 3DCNN USED IN THE LMFE MODULE.

LRLPs Coordinates Feature Extraction (LCFE). As shown in Fig. 2 (b2). The size of the input LRLPs coordinates is $K \times 2 \times T$, and a lightweight 1D CNN is introduced to extract spatio-temporal LRLPs coordinates features. To ensure consistency with the LMFE module, the 1D CNN layer has the same temporal receptive field (5 frames) and output channels ($D = 256$) as the 3D CNN layers used in the LMFE module, resulting in the output of size $K \times D \times T$. Finally, the concatenation of coordinates features and local motion features are treated as the feature representation of those LRLPs.

LRLPs Semantic Encoding. Besides the local motion information and coordinates information, each LRLP also contains different semantic information, *i.e.* facial components. This kind of information is generally ignored in the previous work. The semantic information can be encoded as the landmark index, and we introduce an embedding layer to model the semantic information. This is a simple but effective method to introduce semantic information. The semantic encodings have the same dimension as the fused features from the LMFE module and LCFE module so that the two can be summed directly.

C. Adaptive Semantic-Spatio-Temporal Graph Convolution Network

The remaining key issue is how to explore lip reading-related visual features from these LRLPs features. Because LRLPs are discrete facial feature points, conventional CNNs are not well suited for this case since CNNs can only operate on regular grid data like images. Graph neural networks can collectively aggregate information from graph nodes, and are used to model our selected LRLPs consisting of discrete points and their dependency. Motivated by this, we propose a novel adaptive graph convolution framework to model the semantic and spatio-temporal relationships of LRLPs. Based on the adaptive learning strategy, we explicitly model the relations between LRLPs and show that they are useful for improving ALR performance.

Below, we firstly illustrate the background of spatial based GCNs. Then we elaborate the construction of lip graph, the detailed design of ASST-GCN module (illustrated in Figure 2 (c)) and how it works in lip reading.

(1) Spatial based Graph Convolution Network. Spatial based Graph Convolutional Networks (GCNs) define graph convolution on graph nodes based on their spatial adjacency relations. Here, we follow the same definition of spatial based GCN layer proposed in [29]. Let $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ denotes a graph, where $\mathbf{V}$ is the set of graph nodes and $\mathbf{E}$ are edges. The graph convolution operation on input feature map $\boldsymbol{f}_{in} \in \mathbb{R}^{D_{in} \times K}$ (D_{in} is the input feature dimension of each graph node and K is the total number of graph nodes) can be represented as:

$$\boldsymbol{f}_{out} = \mathbf{\Lambda}^{-\frac{1}{2}}(\mathbf{A} + \mathbf{I})\mathbf{\Lambda}^{-\frac{1}{2}}\boldsymbol{f}_{in}\mathbf{W}, \tag{1}$$

Where $\boldsymbol{f}_{out} \in \mathbb{R}^{D_{out} \times K}$ is the output feature map, $\mathbf{A} \in \mathbb{R}^{K \times K}$ denotes the adjacency matrix, $\mathbf{I}$ represents the identity matrix, and $\mathbf{W} \in \mathbb{R}^{D_{in} \times D_{out}}$ represents for feature transformation matrix. $\Lambda_{ii} = \sum_j (A_{ij} + I_{ij})$. Let the normalized adjacency matrix be $\overline{\mathbf{A}} = \mathbf{\Lambda}^{-\frac{1}{2}}(\mathbf{A} + \mathbf{I})\mathbf{\Lambda}^{-\frac{1}{2}}$. Eq. 1 can be rewritten as:

$$\boldsymbol{f}_{out} = \overline{\mathbf{A}}\boldsymbol{f}_{in}\mathbf{W} \tag{2}$$

From Eq. 2, GCN calculation can be divided into two steps: node features are transformed by a learnable parameter matrix $\mathbf{W}$ and nodes features are aggregated by a specific normalized adjacency matrix $\overline{\mathbf{A}}$.

To further improve the representation ability of GCN, [67] proposes Partition Graph Convolution (PGC) that partitions the neighbors of a node into Q groups based on certain criteria. PGC constructs Q sub-adjacency matrices according to the defined neighborhood by each group. Meanwhile, PGC applies GCN with a different parameter matrix to each neighbor group and sums the results:

$$\boldsymbol{f}_{out} = \sum_{q=1}^{Q} \overline{\mathbf{A}}_q \boldsymbol{f}_{in}\mathbf{W}_q \tag{3}$$

Based on the natural graph structure of human skeleton and joints, Yan *et al.* adopt Eq. 3 as the basic GCN to achieve skeleton-based action recognition [67]. In the case of lip reading, specific facial landmark points, defined as LRLPs in this paper, can be treated as graph nodes, just like the those joints in skeleton graph. However, it is difficult to define a specific lip graph structure like the bones in skeleton graph. To effectively learn visual features with GCNs, the topology of the graph is critical.

(2) Adaptive Semantic-Spatio-Temporal Graph Convolution. The semantic information of LRLPs is human-defined. Therefore, the semantic graph architectures of LRLPs are entirely determined by semantic information for lip reading tasks. In other words, during the test phase, the semantic graph adjacency matrices are unrelated to the current input data, and these adjacency matrices remain fixed after the end of training. Therefore, we define these matrices as sample-independent graph adjacency matrices. However, the spatio-temporal graph architectures of LRLPs are utterly dependent on what is said. Hence, we define the spatio-temporal graph adjacency matrices as sample-dependent graph adjacency matrices.

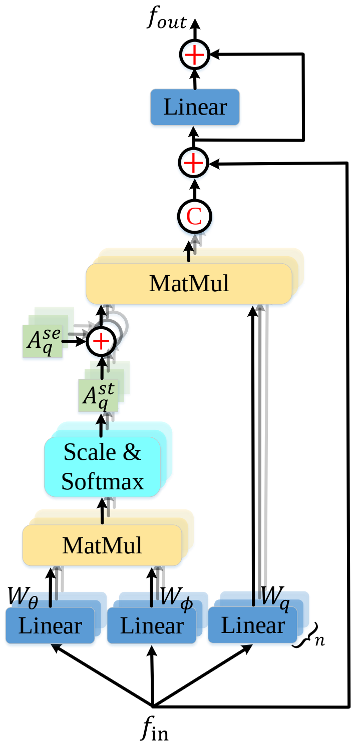


Fig. 3. Illustration of the adaptive semantic-spatio-temporal graph convolution (ASST-GCN) layer. Where n denotes the number of subgraphs.It consists of Two parts: a GCN unit and a feed forward unit. The GCN unit contains n subgraphs. For each subgraph, $\mathbf{A}_q^{se}$ and $\mathbf{A}_q^{st}$ are adjacency matrices. $\mathbf{W}_q^{st}$ is feature transformation matrices of the subgraph. $\mathbf{W}_\theta, \mathbf{W}_\phi$ are projection matrices for node similarity calculation. The output of the all subgraphs are concatenated together as the input of feed forward unit. Multiple residual connections are added to make the network easy to train. All parameters in the units are learnable.

We argue that the optimal lip graph structure should be able to make full use of the semantic relationships and spatio-temporal relationships of LRLPs. Based on this assumption, we define semantic graph and spatio-temporal graph respectively, and make these graph structure parameters to be adaptively learnable. Given K LRLPs as the the lip graph nodes, we adopt the basic graph convolution unit in Eq. 3. The problem is that no predefined lip graph topology can be directly applied to this case. Based on the analysis above, we firstly define two types of fully connected graph structures: sample independent semantic graph $\mathbf{A}^{se} \in \mathbb{R}^{K \times K}$ and sample-dependent spatio-temporal attention graph $\mathbf{A}^{st} \in \mathbb{R}^{K \times K}$. Both of these two graph structure parameters can be optimized together with the other network parameters in an end-to-end manner. In this way, we change Eq. 3 into the following formula:

$$\boldsymbol{f}_{out} = \sum_{q=1}^{Q} (\mathbf{A}_q^{se} + \mathbf{A}_q^{st})\boldsymbol{f}_{in}\mathbf{W}_q. \tag{4}$$

The main difference of Eq. 4 from Eq. 3 is that we introduce two different graph topologies.

Semantic Graph ($\mathbf{A}^{se}$). LRLPs contain rich semantic information which is ignored in previous work. We believe that semantic relations between LRLPs is inherent of the talking mouth. To model this kind of relations, we assume that the semantic graph is a fully connected graph and all of the parameters of $\mathbf{A}^{se}$ is learnable. In addition, there are no constraints on the parameters, which means that the semantic graph is completely learned from the training data without any prior information. To make the semantic graph more flexible and capable, we construct layer-specific semantic graphs for different layers, as we argue that different hierarchical abstract features of the nodes contain different semantic relations. In detail, we independently initialize the semantic subgraph matrices A_q^{se} with the constant value 10^{-6} in each subgraph of each layer, all the semantic subgraph matrices are learned by the network automatically.

Spatio-temporal attention graph ($\mathbf{A}^{st}$). In addition to semantics, there also exist abstract relations between LRLPs corresponding to spoken words. Therefore, we additionally define a sample-dependent spatio-temporal attention graph $\mathbf{A}^{st}$. To determine the connection strength of two nodes, we utilize the soft self-attention mechanism [68], [69] to calculate the similarity of the two nodes in embedding space. In detail, the similarity of two nodes (v_i, v_j) can be defined as follow:

$$s(v_i, v_j) = \frac{(\mathbf{W}_\theta v_i)^T(\mathbf{W}_\phi v_j)}{\sum_{j=1}^{K}(\mathbf{W}_\theta v_i)^T(\mathbf{W}_\phi v_j)}, \tag{5}$$

where $\mathbf{W}_\theta, \mathbf{W}_\phi \in \mathbb{R}^{D_e \times D_{in}}$ are the parameters of the embedding spaces θ and ϕ, respectively. D_e is the dimension of embedding space. So we define the normalized spatio-temporal graph adjacency matrix $\mathbf{A}^{st}$ as follow:

$$\mathbf{A}^{st} = softmax((\mathbf{W}_\theta \boldsymbol{f}_{in})^T(\mathbf{W}_\phi \boldsymbol{f}_{in})). \tag{6}$$

The overall architecture of our ASST-GCN layer is shown in Fig. 3. In addition to the ASST-GCN unit, a fully connected feed forward layer is also utilized to improve the feature representation. Meanwhile, to make the whole network easy to train, we also add multiple residual connections [13] in the ASST-GCN layer.

The ASST-GCN Module (Fig. 2 (c1)) stacks 6 ASST-GCN layers. Each ASST-GCN layer contains 8 subgraphs The number of output channels for all ASST-GCN layers are 512.

D. Sequence Back-end Subnetwork

The goal of sequence back-end subnetwork is to map the fused local and global visual features extracted from the two-stream visual front-end network to natural language. For fair performance comparison of front-end network, we adopt the same back-end network with the baseline models.

As for word-level lip reading task, the baseline model [42] consists of Multi-Scale dilated TCN layers, a fully connected (FC) layer and a softmax layer. The detail of this network architecture is illustrated in Fig. 4. Where k means kernel size, d means dilation size, and s is the strides. It contains three blocks of Multi-scale dilated TCN, where the dilation size of each block is 1,2,4 respectively. Every block consists of three TCN layers, whose kernel sizes are 3,5,7 respectively.

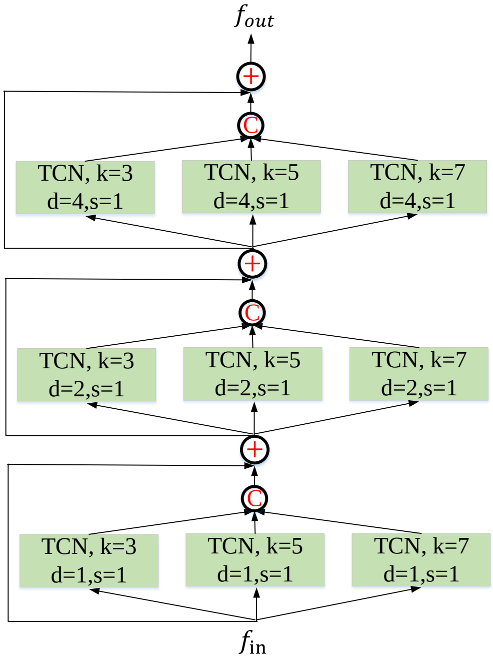


Fig. 4. The details of MSTCN blocks. Where k means kernel size, d means dilation size, and s is the strides.

For the challenging sentence-level lip reading task, the back-end subnetwork produces character probabilities that are directly matched to the ground truth labels. The commonly used transformer variant (transformer-seq2seq [1], [68]) network is adopted as the sequence back-end network. In this variant, we remove the embedding layer in the transformer encoder part because the input is visual representations instead of word class indexes. In addition, the output dimension of the last fully-connected layer of the decoder is changed to 39 to fit the size of the vocabulary.

IV. Experiments

In this section, we provide experiments for both word level and sentence level lip reading tasks to demonstrate the effectiveness of the proposed method. Ablation study is also provided to show the effect of each module.

A. Datasets and Experimental Setup

Large scale datasets play a key role for the research on lip reading. There are several large-scale lip reading datasets [6], [8], [16], [70], [71], such as LRW [8], MVLRS [1], LRS2 [1], LRS3 [6], LSVSR [70]. However, some of these datasets are not publicly available, such as MVLRS and LSVSR. Almost all of the current state-of-the-art sentence-level lip reading models are pretrained on the private datasets [70] or with multiple datasets including MVLRS, LRS2 and LRS3 [1]. Therefore, in order to verify the effectiveness of our method fairly, we reimplement the baseline method on specific public datasets, and compare the results of our method with those implemented on our own.

LRW. The LRW dataset is commonly used for word-level visual speech classification task. It consists of up to 1000 utterances of 500 different English words, spoken by hundreds of different speakers.

Dataset	Subset	#Utter.	Word instances	Vocabularies
LRW	Trainval	514k	514k	500
	Test	25k	25k	500
LRS2	Pretrain	96k	2M	41k
	Main	48k	344k	20k
LRS3	Pretrain	132k	4.2M	52k
	Trainval	32k	358k	17k
	Test	1,452	11k	2,136

TABLE III
THE STATISTICS OF THE DATASETS USED FOR TRAINING AND TESTING.

LRS2. The Lip Reading Sentences BBC dataset (LRS2) is a large-scale lip reading dataset composed of over 140k utterances that selected from BBC television. Each video contains a sentence with variable length. It contains over 2.3 million words with a vocabulary size of 41,000.

LRS3. The LRS3 dataset contains three subsets: *pretrain*, *trainval* and *test*. All videos are selected from TED and TEDx videos. Totally, it contains over 4.2 million words and the vocabulary size is about 51k.

The statistics of the datasets used in this paper is given in Tab. III.

B. Training Details

Preprocessing. For all the datasets, we use dlib or face-alignment detector [65], [66] to detect 68 facial landmark points for each video frame. For each frame of the input video, a lip-centered region of size 112 × 112 pixels is cropped. For the local stream of the visual front-end network, $K = 38$ LRLPs located in the below half of a face are selected. Around each selected facial feature point *i.e.* LRLP, a patch of size 32 × 32 pixels is extracted to represent this point. Moreover, the tip of nose (one of the 68 facial landmark points) is selected to be datum point to do LRLPs coordinates alignment. The videos are converted to grayscale and all frames are normalized with respect to the overall mean and variance of all videos.

Evaluation protocol. For the word-level classification task, classification accuracy (Acc) is reported. In the sentence-level task, Character Error Rate (CER) and Word Error Rate (WER) [72] are reported. CER is defined as $CER = (S + D + I)/N$, where S, D and I are the numbers of substitutions, deletions, and insertions respectively to get from the reference to the hypothesis, and N is the number of characters in the reference. WER and CER are calculated in the same way. The difference lies in whether the formula is applied to character level or word level.

Implementation details. For the sentence-level lip reading task, the output dimension is 39, including the 26 letters, 10 digitals, one punctuation "'" and [*SPACE*] and [*EOS*].

The training proceeds in three stages: first, the two streams of the visual front-end network are separately trained with the LRW dataset. Then the sentence-level back-end network is trained using the LRS2 and LRS3 pretrain dataset, while the parameters of pretrained two stream visual front-end network remain fixed. Finally, the whole network is finetuned with the LRS3 trainval dataset.

We use the same data augmentation technique as that in [1] for training the global stream, such as horizontal flipping and random shifts. In the training phase, the Adam [73] with the default parameters is employed as the optimizer. When training on the LRS3 pretrain dataset, we adopt the similar curriculum learning scheme as that in [16].

For the input of global stream, we also perform data augmentation in the form of horizontal flipping, random shifts of up to ±4 pixels in the spatial dimension and ±1 frame in the temporal dimension. For the local stream, we do not perform any data augmentation.

After training on the LRW dataset, we extract visual features of the LRS2 and LRS3 pretrain dataset, using the trained front-end network. We then train the transformer back-end model directly on the frozen features. The transformer model is trained with the learning rate schedule strategy as Eq. 7. The transformer back-end is trained using teacher forcing - we supply the ground truth of the previous decoding step as the input to the decoder in the training process, while during inference we feed back the decoder prediction.

$$\begin{aligned} lr = d_{model}^{-0.5} \times min((factor \times step)^{-0.5}, \\ (factor \times step) \times warmupStep^{-1.5}) \end{aligned} \tag{7}$$

Finally, the whole model is fine-tuned end-to-end on the trainval set of the LRS3 dataset for one epoch, with learning rate of 10^{-6}. For all the models we use dropout with p = 0.1 and label smoothing.

In the test phase, the beam search decoder is applied to the transformer decoder and the beam width is set to 6. Note that we do not introduce any language model to improve the final result, in order to make a fair comparison.

C. Ablation Study

To investigate the effectiveness of different parts of the proposed two-stream approach, in particular the local stream and the novel ASST-GCN module, we conducted several ablation experiments on the LRW and LRS3 datasets.

Global Stream vs. Local Stream. We evaluate the performance of each individual stream and the two-stream on the LRW and LRS3 datasets, with result listed in Tab. IV. The results clearly demonstrate the effectiveness of the novel local stream for ALR. Importantly, the proposed two-stream approach significantly outperforms each individual stream. This indicates that the lip contour deformation conveys significant information that is complementary to appearance features.

For single-stream methods, The global stream performs slightly better than the local stream. As the performance of the local stream depends heavily on the image resolution and accuracy of facial landmark detection, we believe that the performance of the local stream can be further improved with higher image resolution.

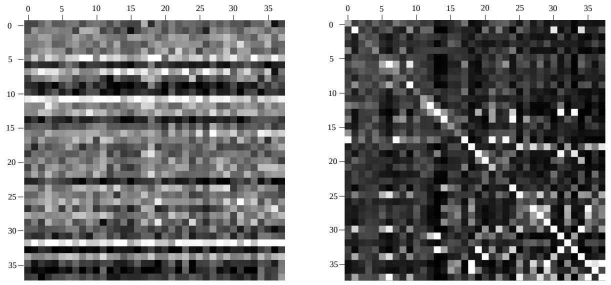
Semantic Graph vs. spatio-temporal Graph. We have defined two types of fully-connected graph convolution for local

Methods	LRW Acc	LRS2 CER	LRS2 WER	LRS3 CER	LRS3 WER
Two Stream	**85.7**	**36.2**	**55.7**	**42.9**	**62.7**
Only Global	85.3	40.1	58.7	48.1	68.8
Only Local	82.6	45.3	61.6	52.1	73.5

TABLE IV

PERFORMANCE (% OMITTED) COMPARISON (LOWER IS BETTER): GLOBAL STREAM *vs.* LOCAL STREAM *vs.* TWO-STREAM. LRW TRAINING WITH MSTCN BACK-END, LRS3 TRAINING WITH TRANSFORMER BACK-END.

visual feature extraction. We argue that the semantic relationship of LRLPs are sample independent, and spatio-temporal relationship of LRLPs are sample dependent. Only the local stream was used to test the performance of each individual graph.



(a) Semantic graph adjacency matrix

(b) Spatial-temporal graph adjacency matrix

Fig. 5. Example of the learned adjacency matrices. (a) Learned Semantic graph adjacency matrix after training with LRW dataset. (b) spatio-temporal adjacency matrix based on one video sample from LRW dataset.

The LRW dataset was used to report results. As shown in Tab. V, it can be observed that both graphs are equally important for lip reading and the combination of them significantly improves performance. Fig. 5 shows an example of the learned adjacency matrices (the adjacency matrices of subgraph 1) in the last GCN layer of the ASST-GCN module. The gray scale of each element in the matrix represents the strength of the connection between LRLPs. It can be seen from the learned semantic graph that some specific points (*e.g.*, the 11_{th} and 32_{nd} points in LRLPs) have stronger connection than some other points (*e.g.*, the 10_{th}, 36_{th} and 37_{th} points in LRLPs). In contrast, LRLPs focus more on themselves in spatio-temporal graph.

Configurations	LRW Acc
Full model	**82.6**
Model w/o $\mathbf{A}^{se}$	79.1
Model w/o $\mathbf{A}^{st}$	78.3
Model w/o LMFE	60.7
Model w/o LCFE	75.9
Model w/o LRLPs SE	79.7
Model with sparse LRLPs	75.4

TABLE V

PERFORMANCE (% OMITTED) COMPARISON: SEMANTIC GRAPH *vs.* SPATIO-TEMPORAL GRAPH, LOCAL MOTION *vs.* COORDINATES. LRLPS SE MEANS LRLPS SEMANTIC ENCODING. W/O X MEANS DELETING THE X MODULE.

Local motion features vs. Coordinate features. As introduced above, the representation of each LRLP is the concatenation of local motion features and coordinates. To verify whether both of them are useful for visual feature extraction, we evaluate the performance of individual features. The results in Tab. V clearly show that the local motion features only outperforms the coordinate feature significantly. However, they are both significantly outperformed by their combination.

Method	TOP-1 Acc
MT [8]	61.1
WAS [16]	76.2
ResNet+BLSTM [10]	83.0
ResNet+BGRU [10]	83.4
Two-Stream 3DCNN+BLSTM [9]	84.1
ResNet+MSTCN [42]	85.3
Ours (ASST-GCN+MSTCN)	**85.7**

TABLE VI

PERFORMANCE (% OMITTED) COMPARISON WITH STATE-OF-THE-ART MODELS ON THE LRW.

LRLPs semantic encoding. Every point in LRLPs conveys human-defined semantic information. We integrate semantic information into the proposed model through the semantic encoding module. To verify its effectiveness, the ablation study on this module is performed on the LRW dataset. As we can see from Tab. V, the performance drops significantly (from 82.6 to 79.7) without the semantic encoding module. The results demonstrate that the introduction of the semantic encoding module is quite useful for lip reading.

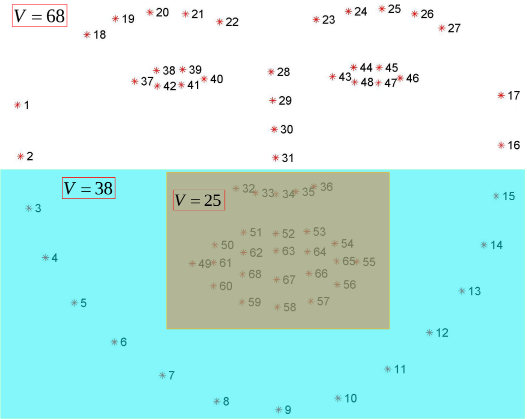


Fig. 6. The selection of LRLPs. The number of the facial landmark points is 68. The number of the facial landmark points located in the below half of a face (blue box) is 38. The number of the facial landmark points located in the mouth area (orange box) is 25.

The selection of LRLPs. the dlib face detector [65], [66] is used to extract facial landmark points as shown in Fig. 6. It is common sense that the upper part of the face is redundant for lip reading. Therefore, we select the 38 landmark points located in the below half of a face as LRLPs. To verify the selection of LRLPs is effective, the ablation study on more sparse LRLPs (25 landmark points located in the mouth area) is conducted on the LRW dataset. The severe performance

Methods	Front-end	Back-end	LRS2 CER	LRS2 WER	LRS3 CER	LRS3 WER
WAS [16]	VGG-M	LSTM	-	70.4	-	-
TM-CTC [1]	C3D_ResNet18	Transformer	-	65.0	-	74.7
TM-seq2seq (Baseline) [1]	C3D_ResNet18	Transformer	40.1*	58.7*	48.1*	68.8*
Zhang et al. [11]	C3D_ResNet18	Transformer (TF Block)	-	**51.7**	-	**60.1**
Ours	ASST-GCN C3D_ResNet18	Transformer	**36.2**	55.7	**42.9**	62.7

TABLE VII

PERFORMANCE (% OMITTED) COMPARISON WITH STATE-OF-THE-ART MODELS ON THE LRS2 DATASET AND THE LRS3 DATASET (LOWER IS BETTER). * MEANS THAT THE RESULTS OF TM-SEQ2SEQ (BASELINE) ARE PRODUCED ONLY WITH PUBLICLY AVAILABLE DATASETS. OUR REPRODUCED RESULT (WER 68.8) IS SIMILAR TO THE RESULT REPRODUCED IN [11] (WER 70.8), THE ORIGINAL RESULT (WER 59.9) REPORTED IN [1] IS BASED ON THE PRIVATE DATASET MVLRS.

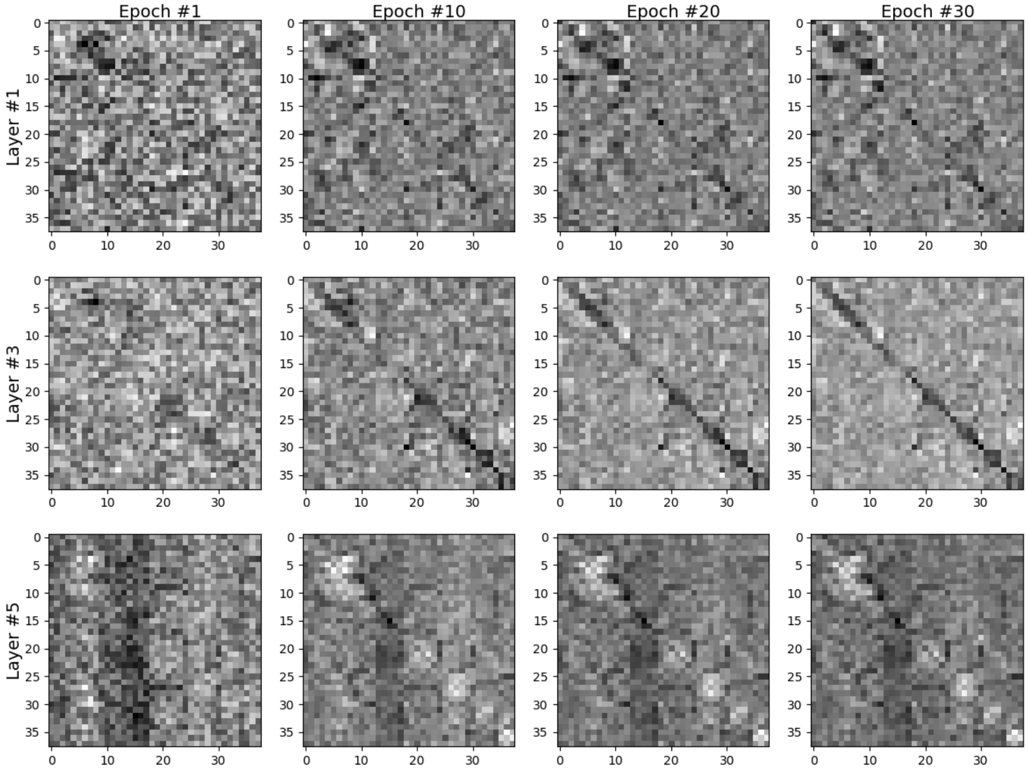


Fig. 7. Some examples of the learned adjacency matrices of the semantic graph. Rows show different training epochs, and columns show different layers.

degradation (from 82.6 to 75.4) proves that the facial contour points located in the below half of a face are pretty crucial for lip reading tasks.

D. Comparative Evaluation

We compare the results of the proposed method with state-of-the-art on LRW, LRS2 and LRS3 datasets. Results are presented in Tab. VII and Tab. VI. In the word-level ALR task, our proposed method outperforms the baseline approach by a large margin. Compared with the recent approach [9], our performance improvement is significant. Their approach is also two-stream, combining appearance with optical flow. Optical flow calculation is very time consuming. In addition, we can also include the optical flow stream to further boost performance.

For the sentence-level ALR task, we have to point out that the state-of-the-art models have been pretrained on a large scale private datasets [70]. Therefore, it is unfair to directly compare with the state-of-the-art. In addition, we have no permission of MVLRS dataset. In order to demonstrate the effectiveness of our proposed approach, we reimplemented the baseline method on the LRS3 dataset. The results in Tab. VII shows that our proposed method significantly outperforms the baseline on the challenging sentence-level ALR task.

E. Visualization and Discussion

Graph Visualisation. The are two kinds of graphs in the ASST-GCN model: the semantic graph and the spatio-temporal graph. Fig. 7 shows some examples of the learned adjacency matrices (subgraph #4) of the semantic graph for different epochs and different layers. During the training phase, all the semantic graph matrices will slowly converge to the optimal values, proving that the hidden semantic relations of LRLPs exist and can be obtained through

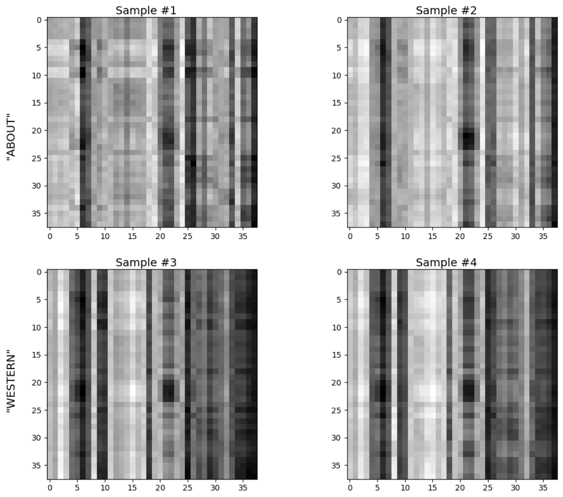


Fig. 8. Average spatio-temporal adjacency matrices of all the subgraphs over all the top ASST-GCN layers. Rows show different test samples, and columns show different word classes.

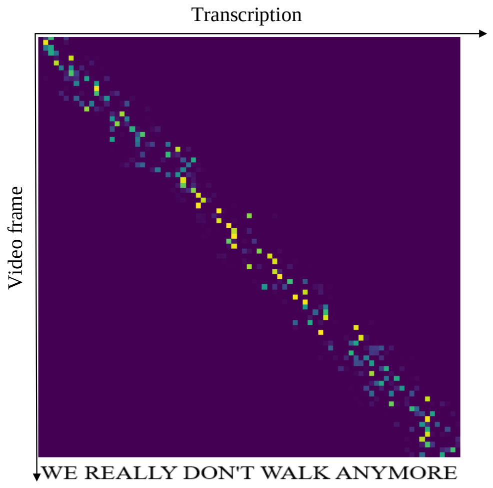


Fig. 9. Average attention map of all the encoder-decoder attention heads over all the decoder layers.

Timestep	Decoded string
01	W
02	WE
03	WE
04	WE R
05	WE RE
06	WE REA
07	WE REAL
08	WE REALL
09	WE REALLY
10	WE REALLY
11	WE REALLY D
12	**THEY** REALLY
13	WE REALLY DON
14	WE REALLY DON’
15	WE REALLY DON’T
16	WE REALLY DON’T
17	WE REALLY DON’T W
18	WE REALLY DON’T WA
19	WE REALLY DON’T WA**N**
20	WE REALLY DON’T WA**NT**
21	WE REALLY DON’T WA**NT**
22	WE REALLY DON’T WA**NT** A
23	WE REALLY DON’T WA**NT** AN
24	WE REALLY DON’T WA**NT** ANY
25	WE REALLY DON’T WA**NT** ANYM
26	WE REALLY DON’T WA**NT** ANYMO
27	WE REALLY DON’T WA**NT** ANYMOR
28	WE REALLY DON’T WA**NT** ANYMORE
Ground Truth: WE REALLY DON’T WALK ANYMORE	

TABLE VII
A DECODING EXAMPLE WITH BEAM SEARCH ALGORITHM.

training. Moreover, the learned semantic graphs are pretty complex and abstract, which confirms our motivation that the semantic graphs for lip reading tasks can not be predefined initially. Besides, the learned semantic graphs are totally different in different layers. It proves the effectiveness of the layer-specific semantic graphs.

As we have discussed, that spatio-temporal graphs mainly depend on what the speaker says (sample-dependent). To prove this statement, we conducted a comparative experiment about spatio-temporal graphs visualization on the LRW dataset. There are totally 500 word categories in the LRW dataset, and we selected two test set samples from two of those categories (*ABOUT* & *WESTERN*). As shown in Fig. 8, the spatio-temporal graph matrices of sample #1 and sample #2 are highly similar, as well as that of sample #3 and sample #4. On the contrary, the spatio-temporal graph matrices of samples of different categories are very different. Meanwhile, the visualization results also convinced us that it makes sense to construct sample-independent semantic graphs and sample-dependent spatio-temporal graphs separately.

Decoding Example. Tab. VII shows an decoding example with beam search algorithm. We list the best prediction of each timestep. Where red color denotes the error output of the current prediction results compared with ground truth. The last line contains the ground truth transcriptions of the example. During timestep 11 to timestep 13, the decoder predicts an error word and then correct it, demonstrating that beam search can effectively improve prediction. The pronunciation of the word *want* and the word *walk* produces similar lip movements, so the final prediction of the decoder makes a small error.

Attention Visualisation. The encoder-decoder attention mechanism of the TM-seq2seq model generates explicit alignment between the input video frames and the ground truth character output. Fig. 9 visualises the dependency of the characters “WE REALLY DON’T WALK ANYMORE” and the corresponding video frames. Since the architecture contains multiple attention heads, we obtain the alignment by averaging the attention masks over all the decoder layers in the log domain [1]. The attention map shows that the prediction of the decoder has obvious short-term dependency.

V. Conclusion

In this work, we introduce graph convolution to capture lip contour, local subtle motion information and semantic information, aiming to improve the visual feature representation capability for lip reading task. And we propose ASST-GCN module to learn semantic preserved local visual features. It assumes that both the semantic graph and spatio-temporal graph structure parameters can be adaptively learned with other parameters in the network. Furthermore, the local visual feature can be easily fused with global visual features with a two stream framework. The two stream visual front-end network framework is proved to be effective on both word-level and sentence-level lip reading task. The final model achieves state-of-the-art performance on LRW dataset, and significantly improve the baseline performance on the LRS2 and the LRS3 dataset.

References

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

[2] W. H. Organization, "Deafness and hearing loss," 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

[3] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and depth-based cnns," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 319–328.

[4] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 389–398.

[5] A. Rekik, A. Ben-Hamadou, and W. Mahdi, "Human machine interaction via visual speech spotting," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2015, pp. 566–574.

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

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

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

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

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

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

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

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

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

[15] T. Afouras, J. S. Chung, and A. Zisserman, "Asr is all you need: Cross-modal distillation for lip reading," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2143–2147.

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

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

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

[19] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1096–1104.

[20] W. Wang, Y. Xu, J. Shen, and S.-C. Zhu, "Attentive fashion grammar network for fashion landmark detection and clothing category classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4271–4280.

[21] L. Fan, W. Wang, S. Huang, X. Tang, and S.-C. Zhu, "Understanding human gaze communication by spatio-temporal graph reasoning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5724–5733.

[22] W. Wang, H. Zhu, J. Dai, Y. Pang, J. Shen, and L. Shao, "Hierarchical human parsing with typed part-relation reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8929–8939.

[23] W. Wang, Z. Zhang, S. Qi, J. Shen, Y. Pang, and L. Shao, "Learning compositional neural information fusion for human parsing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5703–5713.

[24] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, 2015, pp. 2224–2232.

[25] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 1025–1035.

[26] F. Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, and M. M. Bronstein, "Geometric deep learning on graphs and manifolds using mixture model cnns," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5115–5124.

[27] T. Kipf, E. Fetaya, K.-C. Wang, M. Welling, and R. Zemel, "Neural relational inference for interacting systems," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2688–2697.

[28] J. Atwood and D. Towsley, "Diffusion-convolutional neural networks," in *NIPS*, 2016, pp. 1993–2001.

[29] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[30] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 9532–9545, 2020.

[31] S. J. Cox, R. W. Harvey, Y. Lan, J. L. Newman, and B.-J. Theobald, "The challenge of multispeaker lip-reading." in *AVSP*, 2008, pp. 179–184.

[32] P. Lucey, G. Potamianos, and S. Sridharan, "A unified approach to multi-pose audio-visual asr," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[33] I. Almajai, S. Cox, R. Harvey, and Y. Lan, "Improved speaker independent lip reading using speaker adaptive training and deep neural networks," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2722–2726.

[34] G. Zhao, M. Barnard, and M. Pietikainen, "Lipreading with local spatiotemporal descriptors," *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1254–1265, 2009.

[35] Y. Fu, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. S. Huang, "Lipreading by locality discriminant graph," in *2007 IEEE International Conference on Image Processing*, vol. 3. IEEE, 2007, pp. III–325.

[36] Y. Pei, T.-K. Kim, and H. Zha, "Unsupervised random forest manifold alignment for lipreading," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 129–136.

[37] H. E. Cetingul, Y. Yemez, E. Erzin, and A. M. Tekalp, "Discriminative analysis of lip motion features for speaker identification and speech-reading," *IEEE Transactions on Image Processing*, vol. 15, no. 10, pp. 2879–2891, 2006.

[38] K. Saenko, K. Livescu, M. Siracusa, K. Wilson, J. Glass, and T. Darrell, "Visual speech recognition with loosely synchronized feature streams," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2. IEEE, 2005, pp. 1424–1431.

[39] J. Luettin and N. A. Thacker, "Speechreading using probabilistic models," *Computer vision and image understanding*, vol. 65, no. 2, pp. 163–178, 1997.

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

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

[42] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6319–6323.

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

[44] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *NIPS*, 2014, pp. 3104–3112.

[45] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *ICML*. JMLR. org, 2017, pp. 1263–1272.

[46] M. Niepert, M. Ahmed, and K. Kutzkov, "Learning convolutional neural networks for graphs," in *ICML*, 2016, pp. 2014–2023.

[47] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and deep locally connected networks on graphs," in *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

[48] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *NIPS*, 2016, pp. 3844–3852.

[49] M. Henaff, J. Bruna, and Y. LeCun, "Deep convolutional networks on graph-structured data," *arXiv preprint arXiv:1506.05163*, 2015.

[50] R. Levie, F. Monti, X. Bresson, and M. M. Bronstein, "Cayleynets: Graph convolutional neural networks with complex rational spectral filters," *IEEE Transactions on Signal Processing*, vol. 67, no. 1, pp. 97–109, 2018.

[51] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 401–417.

[52] R. Li, S. Wang, F. Zhu, and J. Huang, "Adaptive graph convolutional neural networks," in *AAAI*, 2018.

[53] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *CVPR*, 2019, pp. 12 026–12 035.

[54] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 570–586.

[55] H. Zhao, C. Gan, W.-C. Ma, and A. Torralba, "The sound of motions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1735–1744.

[56] C. Gan, D. Huang, H. Zhao, J. B. Tenenbaum, and A. Torralba, "Music gesture for visual sound separation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 478–10 487.

[57] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, 2017, pp. 349–357.

[58] P. Chen, Y. Zhang, M. Tan, H. Xiao, D. Huang, and C. Gan, "Generating visually aligned sound from videos," *IEEE Transactions on Image Processing*, vol. 29, pp. 8292–8302, 2020.

[59] C. Gan, D. Huang, P. Chen, J. B. Tenenbaum, and A. Torralba, "Foley music: Learning to generate music from videos," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 758–775.

[60] R. Arandjelovic and A. Zisserman, "Objects that sound," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 435–451.

[61] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 631–648.

[62] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 251–263.

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

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

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

[66] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1021–1030.

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

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

[69] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018, pp. 7794–7803.

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

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

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

[73] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

Changchong Sheng received the B.E degree in information engineering and the M.E degree in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2015 and 2017, respectively. He is currently serving as a visiting Ph.D at the Machine Vision Group at the University of Oulu, Finland. His current research interests include lip reading and deep learning.

Xinzhong Zhu received the PhD degree from XIDIAN University and MS degree from National University of Defense Technology (NUDT), China. He is a professor with the College of Mathematics and Computer Science, Zhejiang Normal University, and also the president of Research Institute of Ningbo Cixing Co. Ltd, China. His research interests include machine learning, deep learning, computer vision, manufacturing informatization, robotics and system integration, and intelligent manufacturing. He is a member of the ACM and certified as CCF senior member. Dr. Zhu has published more than 30 peer-reviewed papers, including those in highly regarded journals and conferences such as the IEEE Transactions on Pattern Analysis and Machine Intelligence, the IEEE Transactions on Multimedia, the IEEE Transactions on Knowledge and Data Engineering, AAAI, IJCAI, etc. He served on the Technical Program Committees of IJCAI 2020 and AAAI 2020.

Huiying Xu is an associate professor with the College of Mathematics and Computer Science, Zhejiang Normal University, and also the researcher of Research Institute of Ningbo Cixing Co. Ltd, China. She received the MS degree from National University of Defense Technology(NUDT), China. Her research interests include kernel learning and feature selection, machine learning, deep learning, computer vision, image processing, pattern recognition, computer simulation, digital watermarking, and their applications. She is a member of the China Computer Federation.

Matti Pietikäinen received the doctor of science degree in technology from the University of Oulu, Finland. He is an emeritus professor with the Center for Machine Vision and Signal Analysis, University of Oulu. From 1980 to 1981 and from 1984 to 1985, he visited the Computer Vision Laboratory, University of Maryland. He has made fundamental contributions, *e.g.*, to Local Binary Pattern (LBP) methodology, texture based image and video analysis, and facial image analysis. He has authored more than 350 refereed papers in international journals, books, and conferences. His papers have about 57,000 citations in Google Scholar (hindex 82). He was associate editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), the Pattern Recognition, the IEEE Transactions on Forensics and Security, and the Image and Vision Computing journals. Currently, he serves as associate editor of the IEEE Transactions on Biometrics, Behavior and Identity Science, and served guest editor for special issues of the IEEE Transactions on Pattern Analysis and Machine Intelligence and the International Journal of Computer Vision. He is serving as Guest Editor for IEEE TPAMI special issue on "Learning with Fewer Labels in Computer Vision". He was president of the Pattern Recognition Society of Finland from 1989 to 1992, and was named its honorary member in 2014. From 1989 to 2007, he served as member of the Governing Board of International Association for Pattern Recognition (IAPR), and became one of the founding fellows of the IAPR in 1994. In 2014, his research on LBP-based face description was awarded the Koenderink Prize for fundamental contributions in computer vision. He was the recipient of the IAPR King-Sun Fu Prize 2018 for fundamental contributions to texture analysis and facial image analysis. In 2018, he was named a highly cited researcher by Clarivate Analytics, by producing multiple highly cited papers in 2006-2016 that rank in the top 1 percent by citation for his field in web of science. He is a fellow of the IEEE for contributions to texture and facial image analysis for machine vision.

Li Liu received her Ph.D. degree in information and communication engineering from the National University of Defense Technology (NUDT), China, in 2012. During her PhD study, she spent more than two years as a Visiting Student at the University of Waterloo, Canada, from 2008 to 2010. From 2015 to 2016, she spent ten months visiting the Multimedia Laboratory at the Chinese University of Hong Kong. From 2016.12 to 2018.11, she worked as a senior researcher at the Machine Vision Group at the University of Oulu, Finland. She was a cochair of nine International Workshops at CVPR, ICCV, and ECCV. She served as the Leading Guest Editor for special issues in IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI) and International Journal of Computer Vision. She is serving as the Leading Guest Editor for IEEE TPAMI special issue on "Learning with Fewer Labels in Computer Vision". Her current research interests include computer vision, pattern recognition and machine learning. Her papers have currently over 4,100 citations in Google Scholar. She currently serves as Associate Editor for Pattern Recognition and Pattern Recognition Letter. She serves as Area Chair of ICME 2020, ICME 2021 and ACCV 2020.