IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 June 2022, accepted 19 July 2022, date of publication 21 July 2022, date of current version 27 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3193231

RESEARCH ARTICLE

A Lip Reading Method Based on 3D Convolutional Vision Transformer

HUIJUAN WANG, GANGQIANG PU, AND TINGYU CHEN
Department of Computer, North China Institute of Aerospace Engineering, Langfang, Hebei 065000, China
Corresponding author: Huijuan Wang (wanghj323@126.com)

This work was supported in part by the Scientific Research Key Project of Hebei Provincial Department of Education under Grant ZD2020161, and in part by the Natural Science Foundation of Hebei Province under Grant F2021409007.

ABSTRACT Lip reading has received increasing attention in recent years. It judges the content of speech based on the movement of the speaker's lips. The rapid development of deep learning has promoted progress in lip reading. However, due to lip reading needs to process the information of continuous video frames, it is necessary to consider the correlation information between adjacent images and the correlation between long-distance images. Moreover, lip reading recognition mainly focuses on the subtle changes of lips and their surrounding environment, and it is necessary to extract the subtle features of small-size images. Therefore, the performance of machine lip reading is generally not high, and the research progress is slow. In order to improve the performance of machine lip reading, we propose a lip reading method based on 3D convolutional vision transformer (3DCvT), which combines vision transformer and 3D convolution to extract the spatio-temporal feature of continuous images, and take full advantage of the properties of convolutions and transformers to extract local and global features from continuous images effectively. The extracted features are then sent to a Bidirectional Gated Recurrent Unit (BiGRU) for sequence modeling. We proved the effectiveness of our method on large-scale lip reading datasets LRW and LRW-1000 and achieved state-of-the-art performance.

INDEX TERMS Lip reading, 3D convolution, vision transformer, sequence modeling.

I. INTRODUCTION

Lip reading is to recognize speech information based on the change of lip movement, also called visual speech recognition. Its research process involves computer vision, natural language processing, and other related fields and has broad application prospects in identity authentication [1], [2], improving speech recognition performance [3], synthesizing talking face [4], [5], speaker recognition [6], helping hearing-impaired people communicate, etc.

Speech plays a leading role in human communication. It is generally regarded as a process of multi-sensory cooperation, including acoustic and visual cues [7], and the influence of vision on language perception has long been proved [8]. Lip reading is complicated. For example, words with different pronunciations have different lip reading similarities, and different people have different lip movements when they say the same word. Lip reading is very challenging with the influence of angle, light, and background [9]. The lip reading system was first proposed in [10] and starts with face detection, locates the lip and its surrounding area, and then carries out feature extraction and classification. In traditional lip reading models, lip-movement feature extraction methods include geometry, motion, statistical model, or image transforms [11], [12]. Then Most methods utilized Hidden Markov Models(HMMs) to transform visual features into speech units [13]–[15]. With the development of deep learning, machine lip reading has made significant progress in recent years. Lip reading is a fine-grained feature recognition problem, which focuses on the subtle changes of lips, and the effect of lip reading is determined by capturing and analyzing the subtle changes. Architecture innovations will significantly improve the capabilities of deep learning models [16], [17]. Therefore, many new lip reading models

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenhua Guo.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

have been proposed recently. Initially, a neural network is only used for feature extraction, which is combined with HHM [18]. Later, with the development of recursive networks, Long Short Term Memory(LSTM) and Gated Recurrent Unit(GRU) gradually replaced HMMs [19], [20]. Recently, It is proposed to use temporal convolution instead of LSTM or GRU to obtain feature temporal correlation [21], [22]. When 3D convolution and ResNet are proposed, their combination becomes the most commonly used feature extractor [23], [24]. Lip reading researchers have begun to focus on self-attention mechanisms [25], [26]. The core of the mechanism is determining the part we should pay attention to based on the goal and further analyzing after finding the details.

However, there are still some problems that need our attention. Firstly, as the input is continuous video frames, we need to attach importance to the temporal and spatial information between adjacent frames. Secondly, the recognition mainly focuses on the lips and their surroundings, which makes the network model have more essential requirements for subtle feature extraction. Additionally, with the deepening of the network, the loss of information caused by the resolution reduction should also be a concerned [16], [17]. Based on the above problems and the complexity of lip reading itself, the accuracy of lip reading tasks has not been high. The related work has made slow progress in improving the accuracy of lip reading.

In view of the above problems, this paper improves the previous lip reading method using ResNet as pipe network [27], [28], and proposes a lip reading method based on 3D convolution visual transformer (3DCvT). In our method, the lip movement features are extracted by a 3D convolutional vision transformer. The extracted feature information is processed by Bidirectional Gated Recurrent Unit(BiGRU) and Full Connection(FC) layer for sequence modeling and classification. The method can effectively extract the high-dimensional features of an image sequence, enhance the semantic representation between video keyframes, and reduce the loss caused by the global average of an image sequence.

Specifically, 3D convolution can obtain the spatio-temporal correlation information between continuous images. The transformer divides the image into noncoincident patches conducive to capturing global information. The addition of convolution operation improves the fine-grained extraction of local regions of interest. We add the Squeezing and Excitation(SE) [29] module to the Convolutional Token Embedding [30]. This structure makes the weights of valid feature maps more significant and the weights of invalid feature maps smaller, which makes the model achieve better results. In addition, we should gradually reduce the number of tokens and increase the width of tokens in each stage to realize spatial downsampling and richer representation and make up for the information loss caused by the reduction of resolution. It maintains the characteristics of convolution and adds the advantages of a transformer. For example, local receptive field, shared weight, dynamic attention, and global context fusion. These features help us better extract the temporal and spatial information of continuous video frames and more accurately identify the changes in lip movement. The proposed model is tested on LRW [31] and LRW-1000 [32] datasets, the state-of-the-art performance is obtained, and the related comparative experiments are carried out to prove our model's effectiveness.

Our contributions are summarized as follows: (1) 3D convolution is added to the CvT [30] to capture the spatio-temporal feature information of video frames. (2) SE structure is introduced into the Convolutional Token Embedding layer of CvT to improve the extraction of effective features on the channel. (3) we have provided some useful tricks for lip reading tasks, and have done ablation experiments to prove their effectiveness. (4) On LRW and LRW-1000 lip reading datasets, Our proposed 3DCvT achieved an accuracy of 88.5% and 57.5% respectively. The method's effectiveness outperforms the current state-of-the-art performance.

The rest of this paper is organized as follows: the Section II introduces the methods of feature extraction and time series modeling. The Section III describes the 3DCvT method proposed in this paper. The Section IV describes the experimental content and result analysis. The final section gives our conclusions and prospects.

II. RELATED WORK

A. TRANSFORMER

The transformer was proposed in 2017 [33], and its performance in machine translation tasks is better than that of RNN and CNN. Only encoder-decoder and attention mechanism can achieve good results. The most significant advantage is that it can be parallelized efficiently, which was initially used in natural language processing and achieved considerable results [34], [35]. Recently, the transformer has been introduced into the field of visual recognition. For example [36]–[38], they have achieved better results than before in visual processing. We have reason to believe that transformer can also achieve good results in lip reading, and we have done it.

Vision transformer (ViT) [36] introduces a transformer into visual recognition with a large amount of data. It decomposes each input image into non-overlapping sub-images to form a token with a fixed length. Then, several standard transformer layers are applied to model these tokens. Only transformer is used in the ViT model, and good results are achieved in classification, but large datasets are required to achieve good convergence. Recently proposed by Haiping Wu *et al.* Convolutional vision Transformer (CvT) [30] is used in image vision processing and has achieved high accuracy in the Imagenet dataset. In this architecture, Convolutional Token Embedding and Convolutional Projection, two convolution-based operations, are introduced into the ViT. Convolutional Token Embedding models local spatial context in a multi-stage method. The purpose of Convolutional Projection is

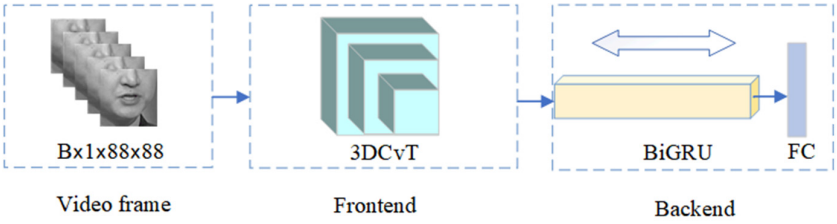


FIGURE 1. The overview of working pipeline using 3DCvT.

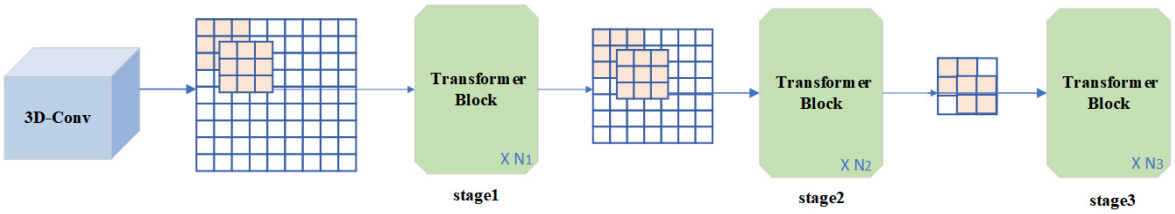


FIGURE 2. The architecture of the 3DCvT.

to obtain the local spatial context better. The convolution operation is used to model the local spatial context in multi-stage and realize the additional modeling of the local spatial environment. However, CvT cannot obtain the timing information of continuous video frames. It only captures features in the spatial dimension, without considering the channel and time dimension information.

B. BiGRU

Gated Recurrent Unit(GRU) is a variant of traditional RNN [39], which can effectively capture the semantic association between long sequences and alleviate the phenomenon of gradient disappearance or explosion. BiGRU is a neural network model composed of two GRU in opposite directions. The GRU formula is as follows:

$$\begin{aligned} z_t &= \sigma\left(W^z x_t + U^z h_{t-1}\right) \\ r_t &= \sigma\left(W^r x_t + U^r h_{t-1}\right) \\ \tilde{h} &= \tanh(W^h x_t + U^h(h_{t-1} \odot r_t) \\ h_t &= (1 - z_t) \odot \tilde{h}_t + z_t \odot h_{t-1} \end{aligned} \tag{1}$$

where z_t is update gate, r_t is reset gate, $\tilde{h}$ is the unit value, h_t is the hidden value, W and U are the input and hidden weight matrix, respectively.

III. MOTHED

This section describes the overall approach to lip reading. The overall architecture is shown in Figure 1. Our improved 3DCvT constitutes the front-end network for feature extraction. The back-end network is composed of BiGRU and carries out sequence modeling on the extracted features. Finally, the output is classified by FC layer.

A. FRONT-END: 3DCVT

In order to better obtain the temporal and spatial information between video frames, this paper constructs a 3D convolutional vision transformer(3DCvT), and its pipeline is shown in Figure 2.

1) 3DCNN

The traditional CNN only processes a single image and cannot obtain the relevant information between continuous video frames in lip reading. In order to solve this problem,

3DCNN was added before the transformer block to deal with the time characteristics of input data. 3DCNN is first proposed and used in human action recognition [40]. It adds a new dimension of information, named the time dimension, and mainly extracts the relevant information between pictures. The 3DCNN formula is as follows:

$$\begin{aligned} v_{ij}^{xyz} &= ReLU(b_{ij} \\ &\quad + \sum_m \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} \sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x-p)(y-q)(z-r)}) \end{aligned} \tag{2}$$

where v_{ij}^{xyz} represents the value in the jth feature map at position (x, y, z) in the ith layer. The ReLU represents the activation function, and b is the offset, and m is the index of the feature map in the $i-1$ layer connected to the feature map in the current layer. The p, q and r in w_{ijm}^{pqr} are the width, height, and spatial information of the convolution kernel, respectively.

The input of lip reading task is a continuous video frame. The input size is 88×88. The 3D convolution kernel size is $5 \times 7 \times 7$, the stride set to $1 \times 2 \times 2$, and the channels are 64.

2) TRANSFORMER BLOCK

The lip movement features extracted by 3DCNN are input to the Transformer Block. The Transformer Block extracts the global feature information through the self-attention mechanism and captures the fine-grained local features

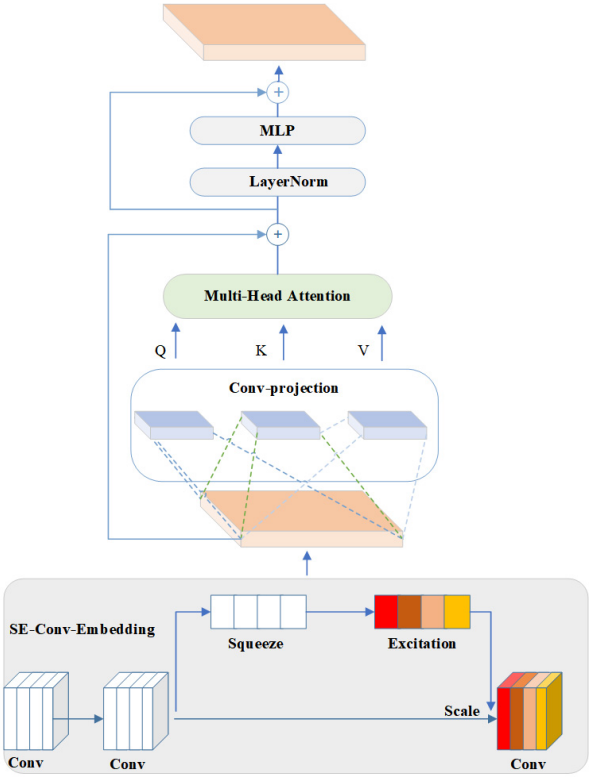


FIGURE 3. The transformer block.

through the convolution structure. The specific structure is shown in Figure 3.

a: SE-CONV-EMBEDDING

With the stacking of Transformer Block and the deepening of the network, the resolution of the image decreases and the ability of the network to capture spatial features information decreases. To solve this problem, we improve the Convolutional Token Embedding layer, add Squeezing and Excitation structure into it. We define the new network layer as SE-Conv-Embedding. It weights the channels to obtain the importance of each feature channel, and correlates the channel information. When the features of spatial dimension is reduced, it will be more difficult to capture the feature. At this time, the number of channels is increasing, and the extraction of feature on channels is particularly important. The feature map is entered into the Squeeze to get the statistical information of the channel through, then the statistical information is entered into the Excitation to get the correlation of the channel, and finally the correlation is entered into the Scale to get the new feature map. The formula is as follows:

$$\hat{x} = F_{scale}\left(F_{ex}\left(F_{sq}\left(conv\left(x\right)\right)\right), x\right) \tag{3}$$

where x is the input feature map, F_{sq} is the squeeze function, F_{ex} is the excitation function, F_{scale} is the scale function, and $\hat{x}$ is the new feature map. This layer can change the parameters to increase the token feature dimension and reduce the token sequence length, compensating for the loss of information caused by the deepening of the network and the reduction of resolution.

The features map in the previous layer is sent to the Conv-projection layer in this block. The purpose of this layer is to get the local spatial context better. The specific formula can be described as:

$$x_i^{q/k/v} = Flatten(Conv2d\left(Reshape2D\left(x_i\right), s\right)) \tag{4}$$

where $x_i^{q/k/v}$ is the input of $Q/K/V$ matrix in layer i, x_i is the output of SE-Conv-Embedding layer, and s is the convolution kernel size. Then, we input the features matrix from the Conv-projection layer into the Multi-Head Self-Attention. The number of heads in each step increases and is normalized by layer normalization, finally, through the MLP layer, the output result is obtained.

B. BACK-END:BIGRU

The back-end network consists of BiGRU for sequence modeling of extracted visual features. Lip reading is the processing of continuous video frames. Therefore, after extracting the high-dimensional feature information of the image, it is necessary to model these features in a temporal sequence. In order to better obtain the global correlation of feature sequences and identify essential information, BiGRU structure is applied in the lip reading task in this paper. The visual feature information output from the front-end is directly sent to BiGRU. The input dimension of the unit is 513, the hidden layer dimension is 1024, there are three layers in total, and the output dimension is 2048. Finally, it is sent to the FC layer for classification.

IV. EXPERIMENTS AND RESULT

In this section, we train and evaluate two large lip reading datasets, LRW [30] and LRW-1000 [31]. The effectiveness of our method is proved and compared with other state-of-the-art lip reading methods.

A. DATASETS

1) LRW [30]

The LRW dataset was proposed by the visual geometry team of Oxford University in 2016. Due to the rise of deep learning, the demand for large-scale datasets is growing, and LRW datasets emerge as the times require. Unlike the previous datasets, the data of the LRW dataset comes from BBC Radio and television programs instead of being recorded by volunteers or experimenters, which makes the data volume of this dataset a qualitative leap. The dataset selects the 500 most common words and intercepts the scenes of the speaker saying these words. Therefore, more than 1000 speakers and more than 550 million discourse examples meet the demand of deep learning for the amount of data to a certain extent.

2) LRW-1000 [31]

LRW-1000 dataset was proposed by the team of the Institute of Computing, the Chinese Academy of Sciences, the

TABLE 1. Classification architecture of 3DCvT on LRW-1000. The default input video frame size is 88 × 88. k is the kernel_ size, s is the stride, c is the number of channels. H_i is the number of heads of the ith MHSA module and D_i is the embedded feature dimension of the ith MHSA module. The expansion ratio of feature dimension in the ith MLP layer is expressed as R_i. N_i the number of Transformer Blocks of the ith stage, which is the number of cycles. In addition, in the BiGRU, h is the hidden layer dimension, l is the number of layers, and $Bi = True$ is the bidirectional GRU.

			Layer Name	Parameter setting	Output Size
Front-end	3D-CNN		Conv3d	k=(5×7×7),s=(1×2×2),c=64	44×44
	Stage1		SE-Conv. Embed	k =(7×7),s=2, c=128	22×22
		Transformer Block	Conv.Proj MHSA MLP	k =(3×3), c=128 $H_1 = 3$ $D_1 = 128$ $R_1 = 4$ $N_1 = 2$	22×22
	Stage2		SE-Conv. Embed	k =(3×3),s=2, c=256	11×11
		Transformer Block	Conv.Proj MHSA MLP	k =(3×3), c=256 $H_2 = 12$ $D_2 = 256$ $R_2 = 4$ $N_2 = 2$	11×11
	Stage3		SE-Conv. Embed	k =(3×3),s=2, c=512	5×5
		Transformer Block	Conv. Proj MHSA MLP	k =(3×3), c=512 $H_3 = 16$ $D_3 = 512$ $R_3 = 4$ $N_3 = 20$	5×5
Back-end	BiGRU		GRU	h =1024,$l = 3$, $c = 2048$, $Bi = True$	5×5
	FC		Linear	$c = 1000$	1×1

University of Chinese Academy of Sciences, and Huazhong University of science and technology in 2018. It aims to establish a large-scale benchmark with different image sizes in an outdoor environment. The dataset covers the natural changes of different speech modes and imaging conditions to meet the challenges in practical applications. The dataset comes from Chinese TV programs and contains 1000 classes. Each class corresponds to one or more Chinese words. The dataset is the largest Chinese word lip reading dataset, with more than 2000 speakers and nearly 720000 utterances. The data richness of the dataset ensures that the deep learning model is fully trained. At the same time, the dataset is also the only open dataset of Mandarin lip reading.

B. IMPLEMENTATION DETAILS

1) DATA PREPROCESSING

We shuffle the initial input video and adjust its size to 96 × 96. Then it is cropped to 88 × 88, and the Mixup [41] method is used for data enhancement, which is the final input. We select a batch of video frames with the size of 256 in each epoch training. Then, each video frame is flipped at a probability level of 0.5, and transformed into a grayscale image, finally normalized to [0, 1]. In addition, before the extracted feature information enters the back-end, we expand the data dimension from 521 to 513, called word boundary. This method can provide context and environment information, which is helpful to the classification of lip reading.

2) NETWORK ARCHITECTURE

This paper divides the lip reading network into front-end and back-end. The specific information is shown in Table 1. In the front-end network 3DCvT, the input size is (256 × 1 × 88 × 88). It first passes through the 3DCnov layer and then enters the Transformer Block. The parameter settings of each stage are different. When entering the back-end network, first go through BiGRU with the input size (256 × 513 × 5 × 5). Before the feature information enters the back-end, we adjust the data dimension from 512 to 513 afterword boundary [42] processing and send it to the entire connection layer for classification processing.

3) MODEL VARIANTS

By changing the number of Transformer Blocks of the ith stage (N_i) and the number of magnetic heads (H_i) of the ith MHSA module, we have obtained three groups of models with different effects, which are respectively defined as 3DCvT-I, 3DCvT-II and 3DCvT-III. Where n of model 1 is 1,2,10 and H is 1, 3, 6 respectively. N of model 2 is 1,4,16 and

TABLE 2. The comparison of the recognition rate of the variants after changing the parameters of our model on the datasets.

Model	LRW	LRW-1000
3DCvT-I	87.1%	56.0%
3DCvT-II	87.6%	56.7%
3DCvT-III	88.5%	57.5%

H is 1, 3, 6 respectively. N and h of model 3 are 2, 2, 20 and 3, 12, 16 respectively. Other parameters remain unchanged. Finally, different results are obtained to verify the scalability of the proposed architecture.

4) TRAINING

This experiment utilized Python 3.8 and pytorch1.7 to build a training network on the platform configured as NVIDIA Tesla v100-pice 32GB graphics card. We set the batch size to 256, use Adam optimizer, set the initial learning rate to 6e-4, 120 epochs for training, and use cosine learning rate scheduling strategy. When the verification error is stable in three consecutive epochs, we reduce the learning rate two times. We choose the cross-entropy loss function with label smoothing as the final decoder. The traditional formula for cross-entropy loss function is as follows:

$$L = -\sum_{i=1}^{N} q_i \log(p_i) \begin{cases} q_i = 0, & y \neq i \\ q_i = 1, & y = i \end{cases} \tag{5}$$

where q is the prediction probability, $p = 1 - q$, and y is the real label value. In the cross-entropy loss function with label smoothing mechanism, the q value is changed to:

$$q_i = \begin{cases} \frac{\epsilon}{N}, & y \neq i \\ 1 - \frac{N-1}{N}\epsilon, & y = i \end{cases} \tag{6}$$

where ϵ is a small constant and take the value 0.1, N is the number of categories. Using label smoothing training can produce a better calibration network to generalize the network better and produce a more accurate prediction for invisible output data.

C. RESULT

The effectiveness of our method is evaluated on two large lip reading datasets LRW and LRW-1000. We changed the relevant parameters of 3DCvT, obtained the 3DCvT-I, 3DCvT-II and 3DCvT-III, and made the corresponding comparative experiments. 3DCvT-III achieves the best recognition accuracy on both datasets. The specific results are shown in Table 2.

In addition, we provide some useful tricks for lip reading tasks and demonstrate their effectiveness in the form of ablation experiments. It includes the word boundary input, Mixup data enhancement technology we used in the data preprocessing stage, and the label smoothing technique used in the cross-entropy loss function. We performed ablation experiments with baseline networks that did not use these methods. We do experiments based on 3DCvT-III, and the results are shown in Table 3. Each adjustment is based on the last change, from top to bottom. These results demonstrate the effectiveness of our use of these tricks.

TABLE 3. Increase the effect of various tricks on the basis of the baseline.

Adjustment	LRW	LRW-1000
Baseline	84.8%	52.6%
word boundary	87.1%	56.5%
Mixup	87.6%	57.1%
Label smoothing	88.5%	57.5%

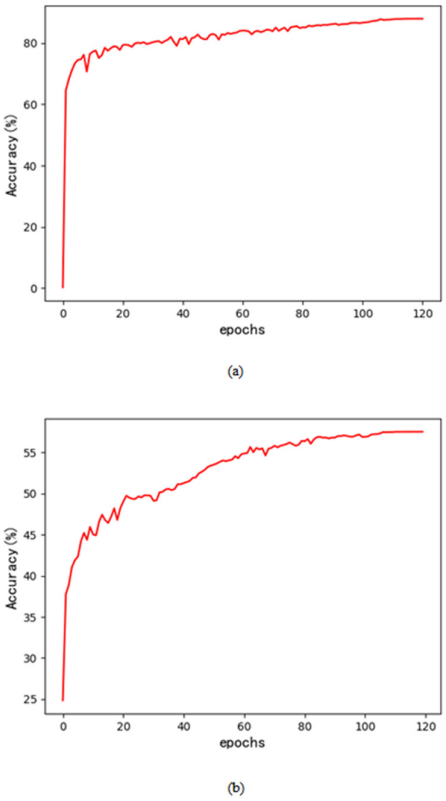


FIGURE 4. (a) LRW dataset recognition accuracy curve. (b) LRW-1000 dataset recognition accuracy curve.

The result verifies the scalability of the proposed model. Finally, the model proposed in this paper achieves the best accuracy of 88.5% and 57.5% on the LRW dataset and LRW-1000 dataset, respectively. The accuracy curve is shown in Figure. 4, in which (a) reflects that the accuracy of the LRW dataset rises sharply after the beginning of training, then fluctuates slightly and tends to be flat after 100 rounds. (b) shows that the accuracy of the LRW-1000 dataset rises rapidly in 20 rounds before training, then rises slowly in 60 rounds, and tends to be flat after 100 rounds. The linguistic features of the two large lip reading datasets are different, which leads to different trends of accuracy.

TABLE 4. Our method is compared with the existing methods.

Work	Method		Data	
	Frond-end	Back-end	LRW	LRW-1000
2019[27]	Tow-Stream	BiLSTM	84.1%	-
2020[22]	3D+ResNet-18	MS-TCN	87.7%	43.2%
2020[28]	3D+SE+ResNet-18	BiGRU	88.4%	55.7%
2020[26]	SBL-ALL-Flag	-	87.3%	56.9%
2021[43]	3D Conv + ResNet-18	Bi-GRU + Visual-Audio Memory	85.4%	50.8%
2022[44]	3D Conv + ResNet-18 + MS-TCN	Multi-Head Visual-Audio Memory	88.5%	53.8%
Ours	3DCvT	BiGRU	**88.5%**	**57.5%**

Generally speaking, the accuracy of the Chinese lip reading dataset LRW-1000 is low because Chinese is relatively complex.

Lip reading has developed rapidly in recent years, and various methods have been put forward. We list several representative methods and compare them with ours, Table 4 for details. It can be seen that most of the networks use ResNet as the baseline, and we introduce the visual application of transformer into lip reading, and the effect is also very obvious. The word recognition rate on LRW is 88.5%, and the word recognition rate on LRW-1000 is 57.5%, which exceeds the current state-of-the-art results.

V. CONCLUSION

This paper proposes a visual change method of a 3D convolutional vision transformer, which take full advantage of convolutions and transformers. In order to achieve the best performance, we have improved the previous vision transformer. The spatio-temporal features of continuous video frames are extracted to effectively obtain the global features of images and the correlation between video frames. Then the extracted features are sent to BiGRU for Sequence modeling. The experiment proved our method's effectiveness on LRW and LRW-1000 and obtained the most advanced performance.

While introducing Transformer Block, this method also increases the number of flops and parameters of the model, and the train and test time is extended. After that, we will continue to carry out lip reading in combination with transformer technology. Optimize the transformer structure and apply it to lip reading with less calculation to facilitate the practical application of lip reading. In addition, audio and multimodal video inputs are used for training to improve the model's performance. Applying transformer technology to the back-end network to replace BiGRU and construct an overall transformer structure is also worth discussing.

REFERENCES

[1] L. Lu, J. Yu, Y. Chen, H. Liu, Y. M. Zhu, L. Kong, and M. Li, "Lip reading-based user authentication through acoustic sensing on smartphones," *IEEE/ACM Trans. Netw.*, vol. 27, no. 1, pp. 447–460, Feb. 2019.

[2] S. Mathulaprangsan, C.-Y. Wang, A. Z. Kusum, T.-C. Tai, and J.-C. Wang, "A survey of visual lip reading and lip-password verification," in *Proc. Int. Conf. Orange Technol. (ICOT)*, Dec. 2015, pp. 22–25.

[3] T. R. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 21, 2018, doi: 10.1109/TPAMI.2018.2889052.

[4] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7832–7841.

[5] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar, "Learning individual speaking styles for accurate lip to speech synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13796–13805.

[6] R. Ding, C. Pang, and H. Liu, "Audio-visual keyword spotting based on multidimensional convolutional neural network," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 4138–4142.

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

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

[9] J. S. Chung and A. Zisserman, "Learning to lip read words by watching videos," *Comput. Vis. Image Understand.*, vol. 173, pp. 76–85, Aug. 2018.

[10] E. D. Petajan, *Automatic Lipreading to Enhance Speech Recognition (Speech Reading)*. Champaign, IL, USA: Univ. Illinois at Urbana-Champaign, 1984.

[11] K. Mase and A. Pentland, "Automatic lipreading by optical-flow analysis," *Syst. Comput. Jpn.*, vol. 22, no. 6, pp. 67–76, 1991.

[12] N. Eveno, A. Caplier, and P.-Y. Coulon, "Accurate and quasi-automatic lip tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 5, pp. 706–715, May 2004.

[13] R. Bowden, S. Cox, R. Harvey, Y. Lan, and B. J. Theobald, "Recent developments in automated lip-reading," *Proc. SPIE*, vol. 8901, Oct. 2013, Art. no. 89010J.

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

[15] R. Seymour, D. Stewart, and J. Ming, "Comparison of image transform-based features for visual speech recognition in clean and corrupted videos," *EURASIP J. Image Video Process.*, vol. 2008, no. 2, pp. 1–9, 2008.

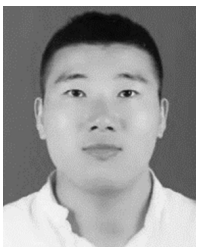
[16] D. Bhatt, C. Patel, H. Talsania, J. Patel, R. Vaghela, S. Pandya, K. Modi, and H. Ghayvat, "CNN variants for computer vision: History, architecture, application, challenges and future scope," *Electronics*, vol. 10, no. 20, p. 2470, Oct. 2021.

[17] C. Patel, D. Bhatt, U. Sharma, R. Patel, S. Pandya, K. Modi, N. Cholli, A. Patel, U. Bhatt, M. A. Khan, S. Majumdar, M. Zuhair, K. Patel, S. A. Shah, and H. Ghayvat, "DBGC: Dimension-based generic convolution block for object recognition," *Sensors*, vol. 22, no. 5, p. 1780, Feb. 2022.

[18] K. Noda, Y. Yamaguchi, K. Nakadai, O. Hg, and T. Ogata, "Lipreading using convolutional neural network," in *Made Available by the Northern Territory Library Via the Publications Act*. Darwin, NT, Australia: Northern Territory Library, 2014.

[19] S. Petridis, Y. Wang, Z. Li, and M. Pantic, "End-to-end multi-view lipreading," in *Proc. Brit. Mach. Vis. Conf.*, 2017.
[20] J. Xiao, S. Yang, Y. Zhang, S. Shan, and X. Chen, "Deformation flow based two-stream network for lip reading," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 364–370.
[21] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6319–6323.
[22] P. Ma, B. Martinez, S. Petridis, and M. Pantic, "Towards practical lipreading with distilled and efficient models," 2020, *arXiv:2007.06504*.
[23] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6548–6552.
[24] K. Xu, D. Li, N. Cassimatis, and X. Wang, "LCANet: End-to-end lipreading with cascaded attention-CTC," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 548–555.
[25] X. Zhang, F. Cheng, and W. Shilin, "Spatio-temporal fusion based convolutional sequence learning for lip reading," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 713–722.
[26] M. Luo, S. Yang, S. Shan, and X. Chen, "Synchronous bidirectional learning for multilingual lip reading," in *Proc. BMVC*, 2020.
[27] X. Weng and K. Kitani, "Learning spatio-temporal features with two-stream deep 3D CNNs for lipreading," in *Proc. 30th Brit. Mach. Vis. Conf.*, 2019.
[28] D. Feng, S. Yang, S. Shan, and X. Chen, "Learn an effective lip reading model without pains," 2020, *arXiv:2011.07557*.
[29] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
[30] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing convolutions to vision transformers," 2021, *arXiv:2103.15808*.
[31] C. Js and A. Zisserman, "Lip reading in the wild," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 87–103.
[32] 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 *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–8.
[33] A. V. Aswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017.
[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019.
[35] D. Mahajan, R. Girshick, V. Ramanathan, K. He, and L. Maaten, "Exploring the limits of weakly supervised pretraining," in *Proc. Eur. Conf. Comput. Vis.*, 2018.
[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, and G. H. Sgju, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021.
[37] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," 2021, *arXiv:2102.12122*.
[38] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," 2021, *arXiv:2101.11986*.
[39] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*.
[40] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3D residual networks for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 3154–3160.
[41] H. Zhang, M. Cisse, D. Yn, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. 18th Int. Conf. Learn. Represent. (ICLR)*, 2018.
[42] T. Stafylakis, M. H. Khan, and G. Tzimiropoulos, "Pushing the boundaries of audiovisual word recognition using residual networks and LSTMs," *Comput. Vis. Image Understand.*, vol. 176, pp. 22–32, Nov. 2018.
[43] M. Kim, J. Hong, S. J. Park, and Y. M. Ro, "Multi-modality associative bridging through memory: Speech sound recollected from face video," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 296–306.
[44] M. Kim, J. H. Yeo, and Y. M. Ro, "Distinguishing homophenes using multi-head visual-audio memory for lip reading," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2022, pp. 1174–1182.

HUIJUAN WANG was born in Dacheng, Hebei, China, in 1982. She received the B.S. and M.S. degrees in computer science and technology from Nankai University, China, in 2005 and 2008, respectively, and the Ph.D. degree from the Hebei University of Technology, China, in 2019. She is currently an Associate Professor with the North China Institute of Aerospace Engineering. She has published more than 20 articles. Her research interests include computer vision, pattern recognition, and deep learning.



GANGQIANG PU was born in Bengbu, Anhui, China, in 1995. He received the bachelor's degree in software engineering from the School of Computer Science, Anhui Polytechnic University, China, in 2019. He is currently pursuing the M.S. degree with the Department of Computer, North China Institute of Aerospace Engineering, China. His research interests include computer vision, image processing, and pattern recognition.

TINGYU CHEN was born in Langfang, Hebei, China, in 1993. She received the bachelor's degree in computer science and technology from the Hebei University of Technology, in 2012, and the M.S. degree in advanced computing with management from the King's College London, U.K., in 2017. She is currently a Lecturer at the North China Institute of Aerospace Engineering. Her research interests include artificial intelligence and pattern recognition.