Deformation Flow Based Two-Stream Network for Lip Reading

Jingyun Xiao[1,2], Shuang Yang[1], Yuanhang Zhang[1,2], Shiguang Shan[1,2], Xilin Chen[1,2]
[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China

***Abstract*— Lip reading is the task of recognizing the speech content by analyzing movements in the lip region when people are speaking. Observing on the continuity in adjacent frames in the speaking process, and the consistency of the motion patterns among different speakers when they pronounce the same phoneme, we model the lip movements in the speaking process as a sequence of apparent deformations in the lip region. Specifically, we introduce a Deformation Flow Network (DFN) to learn the deformation flow between adjacent frames, which directly captures the motion information within the lip region. The learned deformation flow is then combined with the original grayscale frames with a two-stream network to perform lip reading. Different from previous two-stream networks, we make the two streams learn from each other in the learning process by introducing a bidirectional knowledge distillation loss to train the two branches jointly. Owing to the complementary cues provided by different branches, the two-stream network shows a substantial improvement over using either single branch. A thorough experimental evaluation on two large-scale lip reading benchmarks is presented with detailed analysis. The results accord with our motivation, and show that our method achieves state-of-the-art or comparable performance on these two challenging datasets.**

I. INTRODUCTION

Visual speech recognition, also known as lip reading, is the task of decoding speech content based on the visual cues of a speaker's lip motion. Lip reading is a developing topic that has received growing attention in recent years. It has broad application prospects in hearing aids, special education for hearing impaired people, complementing acoustic speech recognition in noisy environments, new human-machine interaction methods, among many other potential applications.

The field of video understanding has progressed significantly in recent years. However, lip reading, a special task of video understanding, remains a challenging task. Different from coarse-grained video analysis tasks, such as action detection and action recognition, lip reading is a fine-grained video analysis task, and requires subtle spatial information in the lip region as well as continuous and discriminative temporal information of lip motion. While humans outperform machines in action recognition, machines have already exceeded humans in lip reading. This is partly because the visual details and lip motions are too subtle for humans to capture and analyze, while machines have an innate advantage in this respect.

Recent lip reading methods are based on deep learning and often conducted in end-to-end fashion. Although promising performance has been achieved by these methods, there are still several issues that demand more consideration. First, most existing lip reading methods extract frame-wise features and then model the temporal relationships with RNNs, with less consideration of the innate spatiotemporal correlation of adjacent frames. Second, one main difference between lip reading and other video analysis tasks is that the input video is focused on the face, and usually a crop of the lip region. It sets higher demands on the discriminative power of subtle facial information in the videos.

In this paper, we propose a Deformation Flow Network (DFN) to generate the deformation flow of the face in a video. It is trained in a completely self-supervised manner, with no need for labeled data. The deformation flow is a sequence of deformation fields. A deformation field is a mapping of the correlated pixels from the source frame to the target frame, which directly represents the motion information from the source frame to the target frame. By computing the deformation field between each pair of adjacent frames, we can capture and represent the motion of the face in the video.

For effective lip reading, we use both the computed deformation flow and the raw videos as the input to a two-stream network. The two branches predict the probabilities of each word class independently. To make the two branches exchange information during training, we adopt knowledge distillation, and utilize a bidirectional knowledge distillation loss to help the two branches learn from each other's predictions during training. At test time, we fuse predictions from both branches to make the final prediction. We observe that a simple average of the predictions produces more accurate predictions, compared with results of using either single branch. It suggests that the two sources of input, the raw video and the deformation flows, provide complementary cues for the lip reading task.

Our contributions are threefold: (a) we propose a Deformation Flow Network (DFN) to generate deformation flows that can capture the motion information of the faces, which is trained in a self-supervised manner; (b) we use the deformation flows and the raw videos as the inputs to a two-stream network, which provide complementary cues for lip reading, and utilize a bidirectional knowledge distillation loss to train the two branches jointly; (c) we conduct extensive experiments on LRW [4] and LRW-1000 [17], demonstrating the effectiveness of our methods.

This work was done by Jingyun Xiao during his internship at the Institute of Computing Technology, Chinese Academy of Sciences.

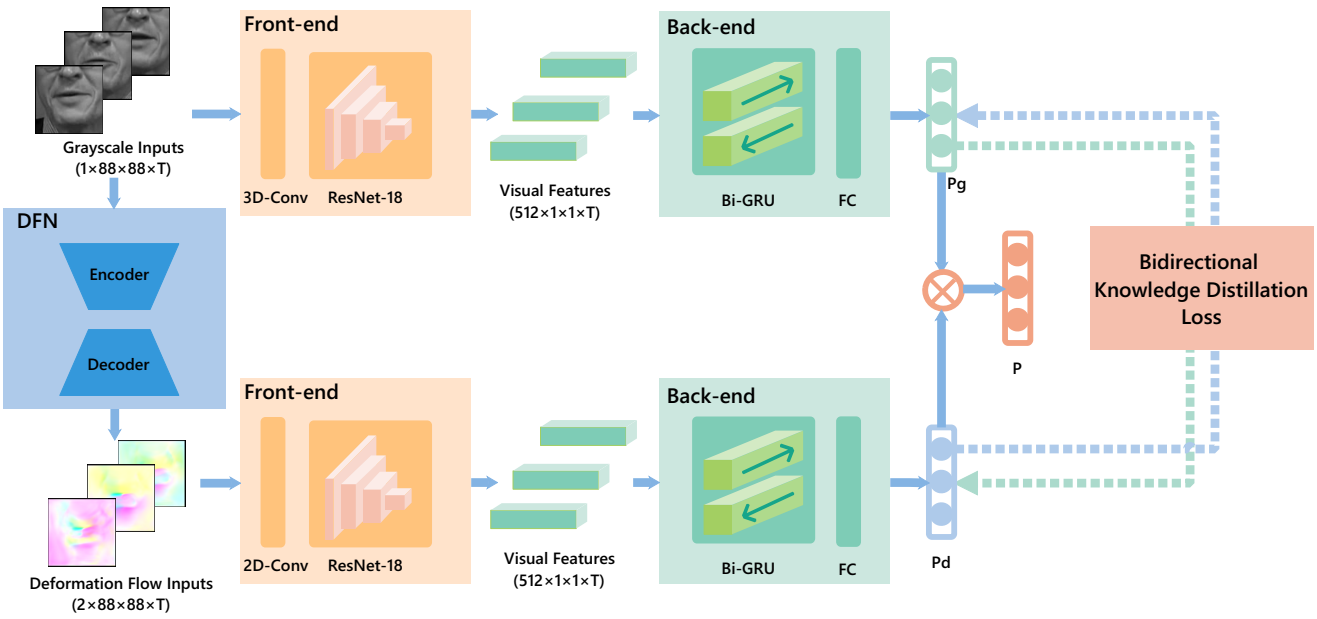

Fig. 1. The overview of Deformation Flow Based Two-stream Network. Given an input video, we first feed it to the Deformation Flow Network to generate the deformation flow. Then the raw video and the deformation flow are fed into the two branches separately. Each branch predicts the probability of each word class independently. At test time, we fuse the results of each branch to improve classification performance. During training, we propose a bidirectional knowledge distillation loss to enable the two branches exchange learned knowledge.

II. Related Works

In this section, we briefly review previous works on deep learning methods for lip reading, as well as self-supervised methods for facial deformation modeling.

A. Deep Learning Methods for Lip Reading

With the rapid development of deep learning in recent years, some works have begun to apply deep learning methods to lip reading and obtained considerable improvements over traditional methods using hand-engineered features. Noda et al. [7] first employed a convolutional neural network (CNN) to extract the features for lip reading. Wand et al. [12] used Long Short-Term Memory (LSTM) to replace the traditional classifier for lip reading, and achieved considerable improvement. In 2016, Chung et al. [4] proposed an end-to-end lip reading model and compared several strategies of processing the frames for word classification, which has founded a solid base for the subsequent progress for lip reading. Since then, more recent lip reading approaches have followed an end-to-end paradigm. Concurrently, Assael et al. [1] proposed LipNet, which is the first end-to-end sentence-level lip reading model.

In 2017, Stafylakis et.al. [10] proposed a new word-level lip reading model that attains 83.0% classification accuracy on the LRW dataset, which is a significant improvement over prior art. It uses a combination of a single 3D convolution layer, ResNet [5], and bidirectional LSTM networks [6]. The proposed architecture shows a strong spatiotemporal modeling power, successfully copes with many in-the-wild variations that LRW presents. Inspired by the success of deep spatiotemporal convolutional networks and two-stream architectures in action recognition, Weng et al. [14] introduced deep spatiotemporal convolutional networks to lip reading. They also employ optical flow and two-stream networks. However, the optical flow is hard to obtain and it costs considerable time and storage. Moreover, most existing optical flow methods are unable to capture the fine-grained motion information of the lip region.

B. Self-supervised Facial Deformation

Recently, there have been a series of works using the deformation field and warping methods for face manipulation, facial attributes learning and other face-related tasks.

The Deforming Autoencoder (DAE) [9] presents an unsupervised method to disentangle shape (in the form of a deformation field) and appearance (texture information disregarding the pose variations) of a face. The learned features are demonstrated to be effective for face manipulation, landmarks localization and emotion estimation.

X2Face [16] is a network that can generate face images with a target pose and expression. In the evaluation stage, given a source face and a driving face, the network is able to generate a new face that preserves the identity, appearance, hairstyle and other attributes of the source face, while possessing the pose and expression of the driving face. In the training stage, it uses a pixel-wise L1 loss between the generated frame and the driving frame to supervise the training process. In this way, the training process of the network does not need any annotations.

FAb-Net [15] has a similar architecture to X2Face. However, it aims to learn the facial attributes in a self-supervised manner. The learned facial attributes are demonstrated to achieve results comparable to and even surpassing the features learned by supervised methods in several tasks.

Inspired by these works, we propose the Deformation Flow Network (DFN) in our work to model the lip motion in the

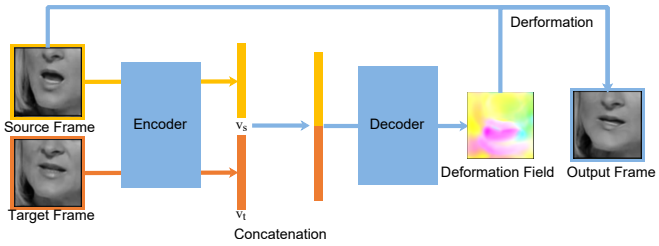

Fig. 2. The architecture of DFN. It consists of an encoder and a decoder. Given a source frame and a target frame, the encoder encodes them into two feature vectors, v_s and v_t. The decoder takes the concatenation of v_s and v_t as input, and generates a deformation field. The source frame is warped by the deformation field, and generates an output frame. A pixel-wise L1 loss between the output frame and target frame can supervise the network effectively. The DFN is trained in a completely self-supervised manner.

speaking process for lip reading, which is also trained in a self-supervised manner.

III. Methods

In this section, we introduce our Deformation Flow Network (DFN) for generating the deformation flow, Deformation Flow Based Two-stream Network (DFTN) for word-level lip reading, and the bidirectional knowledge distillation loss for training the two-stream network jointly.

An overview of the pipeline is shown in Fig. 1. Given an input video (i.e., cropped grayscale image sequence of the lip region), we first feed it to the Deformation Flow Network to generate a series of deformation fields, one for each pair of adjacent frames. This resulting deformation field sequence is the deformation flow of the original video. Next, the grayscale video and the deformation flow are fed into the two branches separately for recognition. The two branches are optimized with individual classification losses, and a bidirectional knowledge distillation loss, which helps the two branches learn from each other. At test time, we fuse the results of each branch to make the final prediction for the input video.

A. Deformation Flow Network

The architecture of the Deformation Flow Network (DFN) is shown in Fig. 2. The input to the DFN is a pair of frames (i.e., a source frame and a target frame). The output of the DFN is a deformation field, which is a 2-channel map with the same size as the input frames. The DFN consists of an encoder and a decoder. The encoder encodes the source frame s and target frame t into a source vector v_s, and a target vector v_t. The decoder takes the concatenation of v_s and v_t as input, and generates a deformation field d, which predicts the relative offsets $(\delta x, \delta y)$ for each pixel location (x,y) in the target frame relative to the source frame. An output frame o is generated by sampling from the source frame s with the offsets $(\delta x, \delta y)$ of the deformation field d:

$$o(x,y) = s(x+\delta x, y+\delta y) \tag{1}$$

The output frame $o = D(s,t)$, is expected to be identical to the target frame t, which can be supervised by a pixel-wise L1 loss between the output frame and target frame:

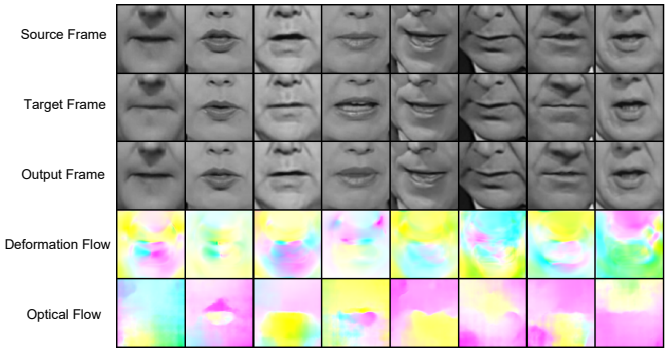

Fig. 3. Examples of the source frames, target frames, output frames, deformation flow generated by DFN, and optical flow generated by PWC-Net [11]. The color variations indicate that the deformation flow captures more details of the face than the optical flow.

Fig. 4. Examples of the difference images of output frame and target frame.

$$\mathscr{L}_1 = \frac{1}{n}\sum_{(x,y)} |o(x,y) - t(x,y)| \tag{2}$$

Given the above optimization target, the DFN can be trained in a completely self-supervised manner, with no need for any extra manual annotations. Examples of the source frames, target frames and output frames are shown in Fig. 3. It is worth noting that since the deformation field is estimated at the pixel level, it can capture very subtle variations of faces and directly represent the motion information, which means it also has great potentials in other face-related tasks beyond lip reading.

B. Deformation Flow Based Two-stream Network

In this subsection, we introduce the two branches (i.e., the grayscale branch and the deformation flow branch) of the Deformation Flow Based Two-stream Network, as well as the fusion strategy of the two branches in detail.

Firstly, we introduce the baseline model in this paper. The grayscale branch adopts the widely used architecture proposed by [10], which is a combination of CNN and RNN, except that we use Gated Recurrent Units (GRU) [2] instead of LSTMs. Specifically, it consists of a front-end (i.e., a single layer of 3D CNN followed by ResNet-18 [5] and a back-end (i.e., a 2-layer bidirectional RNN with GRUs).

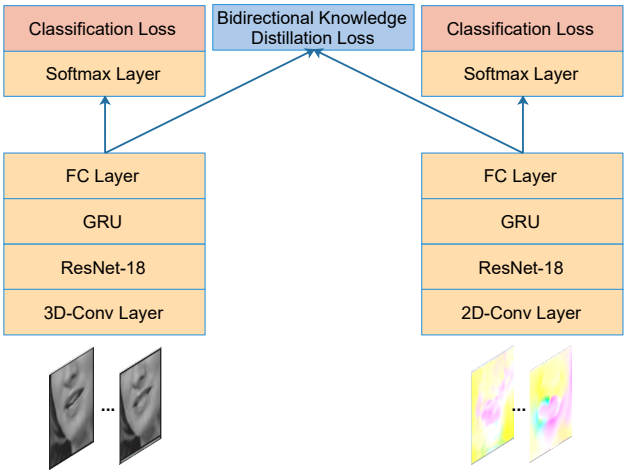

Fig. 5. The architecture of the two-stream network for lip reading. The learning propcess is guided by both the classification loss and the bidirectional knowledge distillation loss.

The front-end extracts the visual features for each frame, and outputs a sequence of feature vectors. The back-end decodes the feature sequences, and predicts the probability of each word class. The deformation flow branch mostly mirrors the structure of the grayscale branch. The only difference is that the first layer of this branch is a 2D convolution layer, while it is a 3D convolution layer in the grayscale branch. The detailed architecture is shown in Fig. 5.

Massive amount of works on two-stream networks have explored methods to fuse the two branches. In this work, we experimented with different fusion strategies, and the results with different fusion strategies are presented in IV-C. Among all the strategies, we find that fusing the output probabilities of the two branches gives the best performance.

However, the problem with fusing the predicted probabilities from individual branches is that the two branches are optimized separately, and lack interaction in the training stage. We wish to design a method that can help the two branches exchange the knowledge they learned during the training process. Therefore, we propose the bidirectional knowledge distillation loss.

C. Bidirectional Knowledge Distillation Loss

In this subsection, we introduce the bidirectional knowledge distillation loss as an additional supervision for training the two branches jointly.

Fusion strategies for the two-stream architecture have been widely explored in the field of action recognition. Here, we adopt the method of knowledge distillation. The two branches are able to make word-level classification as two independent models respectively. The outputs of the fully-connected (FC) layers of the grayscale branch and the deformation branch are denoted as $\mathbf{z}_g$ and $\mathbf{z}_d$ respectively. We then obtain the predicted probability distribution over all classes, q_g and q_d as:

$$q^{(i)} = \frac{\exp(z^{(i)}/T)}{\sum_j \exp(z^{(j)}/T)}, \tag{3}$$

where T is a parameter known as *temperature*. T is usually set to 1 for classification tasks, and the equation becomes the softmax function. In knowledge distillation, a large T makes the probability distribution q "softer", which is easier for a student network to learn than a one hot vector corresponding to the ground truth. In our work, we set T to 20. The knowledge distillation loss is defined as:

$$L_{KD}(q_t, q_s) = -\sum_{i=1}^{N} q_t^{(i)} log q_s^{(i)}, \tag{4}$$

where q_t and q_s denotes the soft probability distributions of the teacher network and student network, respectively, and N denotes the number of classes.

Since we expect the two branches to learn from each other, we adopt a bidirectional knowledge distillation loss:

$$L_{BiKD}(q_g, q_d) = L_{KD}(q_g, q_d) + L_{KD}(q_d, q_g) \tag{5}$$

Therefore the final objective function of the two-stream network is:

$$L = L_{CE}(z_g, y) + L_{CE}(z_d, y) + \lambda L_{BiKD}(q_g, q_d), \tag{6}$$

where L_{CE} represents the standard cross-entropy loss for classification tasks, y is the one hot vector indicating the word class label of the video, and λ is a hyper-parameter indicating the weight of L_{BiKD}.

IV. Experiments

A. Datasets

The proposed methods are evaluated on two large-scale public lip reading datasets, LRW [4] and LRW-1000 [17]. Here we give a brief overview of the two datasets.

LRW. LRW [4] is a large and challenging word-level lip reading dataset. Each sample of LRW is a video snippet of 29 frames captured from BBC programs. The label is the corresponding word class of the video snippet. The dataset has 500 word classes and each class has around 1000 training samples, 50 validation samples and 50 testing samples. The total duration of LRW is approximately 173 hours. The main challenges of LRW are: (a) the variability of appearance and pose of the speakers, (b) similar word classes such as "benefit" and "benefits", "allow" and "allowed", which demands strong discriminative power of models, and (c) the target words do not exist independently in the videos; rather, they are presented with surrounding context, which requires the model to focus on the correct keyframes.

LRW-1000. LRW-1000 [17] is the first public large-scale Mandarin lip reading dataset. It is a naturally-distributed large-scale benchmark for lip reading in the wild which contains 1,000 word classes with more than 700,000 samples from more than 2,000 individual speakers. Each class corresponds to the syllables of a Mandarin word composed of one or several Chinese characters. It is a challenging dataset, marked by the following properties: (a) it contains significant

image quality variations such as lighting conditions and scale, as well as speakers' attribute variations in pose, speech rate, age, make-up and so on, (b) the frequency of each word class is imbalanced, which is consistent with the natural case that some words occur more frequently than others in the everyday life, and (c) the samples of the same word are not limited to a constant length range to allow for modeling of different speech rates. These factors make LRW-1000 a challenging lip reading benchmark with a large lexicon.

B. Implementation Details

Data preprocessing. For both LRW and LRW-1000, we resize the cropped images of lip region to 96×96 as input. For LRW, we randomly crop the input to 88×88 during training and apply random horizontal flipping. For LRW-1000, we take a central 88×88 crop, and do not apply random flipping.

Network architecture. For DFN, we employ a ResNet-18 [5] as the encoder, and 7 cascaded pairs of deconvolutional layers and bilinear upsampling layers as the decoder. The encoder yields a 256-dimensional vector for each frame. The decoder takes the concatenation of a source vector v_s and a target vector v_t as input, which is 512-dimensional, and then generates a 2-channel deformation field with the same size as the input frames. The two channels of the deformation field denote the offsets along the x and y axis at each pixel location.

For the lip reading model, as mentioned earlier, we employ ResNet-18 as the front-end and GRU as the back-end. More specifically, for the grayscale branch, the front-end is a single 3D convolution layer followed by a powerful ResNet-18 network which yields a 512-dimensional vector for each frame. For the deformation branch, we use a single 2D convolution layer on top of the ResNet-18 network. As for the back-end, we use a 2-layer bidirectional Gated Recurrent Unit (Bi-GRU) RNN with 1024 hidden units to process the sequence of the 512-dimensional vectors, each vector extracted from a frame.

Training strategies. We use the three-stage training strategy proposed in [10]. We use the Adam optimizer with default hyperparameters. For LRW, the learning rate is initialized to 0.0001 and reduced by half every time when the validation loss stagnates, until the model reaches convergence. For LRW-1000, the learning rate is initialized to 0.001. In all of our experiments, when the validation loss stagnates for the first time, we reduce the learning rate of the back-end to 10% of the learning rate of the front-end. This policy works well in alleviating the overfitting problem. As for the weight of bidirectional knowledge distillation loss, we initialize it to be 100, and reduce it by half every time when the validation loss stagnates.

C. Evaluation of DFN

We performed a thorough evaluation of DFN over several aspects on LRW [4].

Firstly, the source frames, target frames, output frames, and generated deformation fields are shown in Fig. 3. As can be seen, the output frame matches the target frame quite well. Visualizations of the deformation field shows clear discrimination of the lip region, which carries the motion information we wish to capture, from neighboring regions. This indicates that DFN can generate precise deformation fields, which meets our expectation of directly capturing motion in the speakers' faces, especially in the lip region.

Secondly, we also studied the reconstruction quality of the output frames qualitatively and quantitatively. As shown in Fig. 4, DFN is able to reconstruct faces of varying poses by warping the source frames. We randomly chose 2000 pairs of target frames and output frames to evaluate the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) index. The average PSNR is 26.86 and the SSIM index is 0.82, which also proves the effectiveness of our method.

Inspired by the observations in [8], we further experiment with replacing the L1 loss with classification loss to supervise the DFN. This should help the DFN learn to generate task-specific deformation flows which better suits the lip reading task. Specifically, we freeze the decoder and unfreeze the encoder of DFN when training the deformation flow branch with classification loss, after pretraining in the self-supervised manner. As shown in Fig. 6, the action of mouth opening or closing is slightly amplified in the output frames compared with the motion in the target frames. The classification accuracy is also improved, as shown in Table I.

TABLE I
EVALUATION OF DIFFERENT INPUTS ON LRW.

Input	Accuracy (%)
Grayscale	81.91
Deformaion Flow	77.24
Deformaion Flow (optimized by classification loss)	79.43
Optical Flow	67.81

Finally, we compared DFN with a state-of-the-art optical flow method, PWC-Net [11], on the task of lip reading qualitatively and quantitatively in the following aspects.

(1) We utilize the pretrained model in [11] to generate the optical flow of the adjacent frames in the video, and use the optical flow for lip reading. The generated optical flow and deformation flow are shown in Fig. 3. It shows that the deformation flow reflects more fine-grained details.

(2) We use the deformation flow generated by DFN and optical flow generated by PWC-Net as inputs to evaluate their lip reading performance on LRW respectively. The results of are presented in Table I. It indicates that our task-specific deformation flow is more effective for the lip reading task.

(3) We also compared the network complexity (i.e., floating point operations (FLOPs) and the number of params) of DFN and PWC-Net, which is shown in Table II. The result shows that the computational complexity of DFN is much lower than PWC-Net, which is one of our motivations to propose DFN. The greatly reduced complexity makes it possible to use DFN in real-time applications.

TABLE II

COMPUTATION EXPENSE OF DIFFERENT NETWORKS.

Network	GFLOPS	# Params
DFN	14.5	7.95M
PWC-Net	635	9.37M
Lip Reading Model	18.4	40.5M

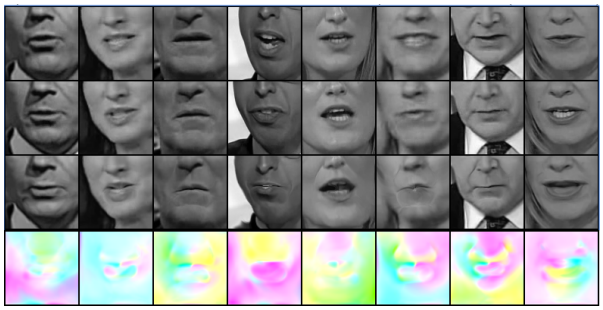

Fig. 6. The output frames and deformation fields generated by DFN, where the encoder is optimized with the classification loss instead of the L1 loss. The output frames have slight differences from the target frames. According to the views in [8], optical flow learned for action recognition in a task-specific manner differs from traditional optical flow and improves the performance of action recognition. This is also the case with the deformation flow.

D. Evaluation of DFTN

In this subsection, we present the ablation studies of DFTN on LRW and LRW-1000.

Evaluation of each single branch. We pretrained the two branches (i.e., the grayscale branch and the deformation flow branch) of the two-stream networks independently. The inputs of the two branches are shown in Fig. 7. The grayscale branch alone is also the baseline model in this paper. The results in terms of recognition accuracy on LRW and LRW-1000 are shown in Table III.

Evaluation of the two-stream network. We fused the probabilities predicted by the two branches to make the final classification of the testing samples. The results are shown in Table III. Empirically, we found using multiplicative fusion, i.e. taking the product of the probabilities results in higher recognition accuracy than additive fusion, i.e. taking the average of the probabilities of the two branches.

TABLE III

EVALUATION OF DFTN ON LRW AND LRW-1000.

Method	LRW (%)	LRW-1000 (%)
Grayscale branch (baseline)	81.91	38.56
Deformaion flow branch	79.43	36.44
Two-stream	83.03	41.46
Grayscale branch (with L_{BiKD})	82.93	38.76
Deformation flow branch (with L_{BiKD})	80.85	37.47
Two-stream(with L_{BiKD})	**84.13**	**41.93**

TABLE IV

EVALUATION OF DIFFERENT STRATEGIES ON LRW.

Method	Accuracy (%)
Grayscale branch	81.91
Deformation flow branch	79.43
Avg (FC)	82.13
Add (Res4)	82.52
Mul (probabilities)	83.03
Mul (probabilities) (with $L_{KD(d->g)}$)	82.14
Mul (probabilities) (with $L_{KD(g->d)}$)	82.92
Mul (probabilities) (with L_{BiKD})	**84.13**

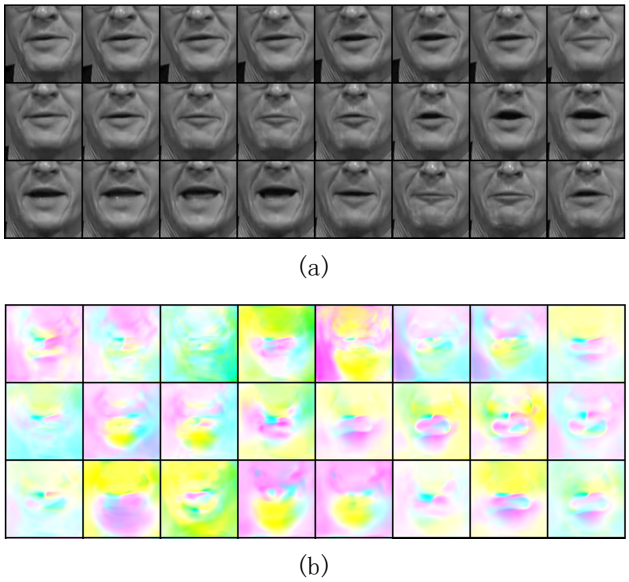

Fig. 7. Examples of the inputs of the grayscale branch and the deformation flow branch.

Evaluation of the bidirectional knowledge distillation loss. To make the two branches exchange the learned knowledge and further improve the performance of DFTN, we trained the two-stream network with the bidirectional knowledge distillation loss as an additional supervision. The results are presented in Table III. It is shown that the bidirectional knowledge distillation not only improves the accuracy of the joint prediction, but also improves the prediction accuracy of each branch when they work independently.

Evaluation of different fusion strategies and distillation strategies. To further validate the effectiveness of the bidirectional knowledge distillation loss, we conducted experiments to compare the performance of different fusion strategies and distillation strategies. We experimented with two fusion methods that fuse the intermediate features of the two branches rather than the probabilities:

1) Average the outputs of FC layers of the two branches, feed the vector to a softmax layer to get the probability distribution, and then compute the cross-entropy loss;
2) Adopt the fusion method in [14], i.e. sum the outputs of the last layers of ResNet of the two branches, feed the resulting vector to the back-end to get the probability distribution, and then compute the cross-entropy loss.

Besides the above fusion strategies, we also experimented with two unidirectional knowledge distillation strategies to compare with the bidirectional knowledge distilling strategy:

1) Distill knowledge from the grayscale branch to the deformation flow branch.
2) Distill knowledge from the deformation flow branch to the grayscale branch.

The results are presented in Table IV. It indicates that the fusion of the output probabilities performs better than the fusion of the intermediate features of the two branches (mid-fusion). Also, the bidirectional knowledge distillation

TABLE V

COMPARISON WITH OTHER METHODS ON LRW.

Method	Accuracy (%)
Chung16 [4]	61.10
Chung17 [3]	76.20
Stafylakis17 [10]	83.00
Stafylakis17 [10] (reproduced)	77.80
Weng19 [14]	84.07
DFTN	**84.13**

TABLE VI

COMPARISON WITH OTHER METHODS ON LRW-1000.

Method	Accuracy (%)
Yang19 [17]	38.19
Wang19 [13]	36.91
DFTN	**41.93**

outperforms unidirectional knowledge distillation with an obvious improvement.

E. Comparison with State-of-the-Art

Comparison with other methods on LRW. We compared our method with other word-level lip reading methods [4], [10], [14] on LRW. The results are presented in Table V. The model in [14] employs deep 3D CNNs and optical flow based two-stream networks, which achieved the existing state-of-the-art performance. Our method outperforms it, and establishes the new state-of-the-art performance.

Comparison with other methods on LRW-1000. We compared our method with other word-level lip reading methods [17], [13] on LRW-1000. The results are presented in Table VI. Our method shows a considerable improvement over all previous methods on LRW-1000 and achieves state-of-the-art performance.

V. CONCLUSION

In this paper, we propose a Deformation Flow Network (DFN) to generate the deformation flow, a way to model the lip movements in the speaking process as a sequence of deformations over the lip region. Notably, the network is lightweight and trained in a self-supervised manner. To take advantages of the complementary cues provided by the deformation flow and the raw videos, we propose a Deformation Flow Based Two-stream Network (DFTN) for word-level lip reading. Different from previous methods that fuse the features of the two branches, we employ the bidirectional knowledge distillation loss to help the two branches interact with each other, and exchange knowledge during training. Finally, we compare our method with other word-level lip reading methods, and show that our method achieves state-of-the-art performance. Our work makes a first attempt to introduce facial deformation to generate a new modality. It provides potential applications and possibilities for not only lip reading, but also other face analysis tasks.

VI. ACKNOWLEDGMENTS

This work is partially supported by National Key R&D Program of China (No. 2017YFA0700804) and National Natural Science Foundation of China (No. 61702486, 61876171).

REFERENCES

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

[2] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[3] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3453. IEEE, 2017.

[4] J. S. Chung and A. Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103. Springer, 2016.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[7] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata. Lipreading using convolutional neural network. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[8] L. Sevilla-Lara, Y. Liao, F. Güney, V. Jampani, A. Geiger, and M. J. Black. On the integration of optical flow and action recognition. In *German Conference on Pattern Recognition*, pages 281–297. Springer, 2018.

[9] Z. Shu, M. Sahasrabudhe, R. Alp Guler, D. Samaras, N. Paragios, and I. Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 650–665, 2018.

[10] T. Stafylakis and G. Tzimiropoulos. Combining residual networks with lstms for lipreading. *arXiv preprint arXiv:1703.04105*, 2017.

[11] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018.

[12] M. Wand, J. Koutník, and J. Schmidhuber. Lipreading with long short-term memory. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6115–6119. IEEE, 2016.

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

[14] X. Weng and K. Kitani. Learning spatio-temporal features with two-stream deep 3d cnns for lipreading. *arXiv preprint arXiv:1905.02540*, 2019.

[15] O. Wiles, A. Koepke, and A. Zisserman. Self-supervised learning of a facial attribute embedding from video. *arXiv preprint arXiv:1808.06882*, 2018.

[16] O. Wiles, A. Sophia Koepke, and A. Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 670–686, 2018.

[17] 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 *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.