Multi-grained Spatio-Temporal Features Perceived Network for Event-based Lip-Reading

Ganchao Tan[1]*, Yang Wang[1]*, Han Han[1], Yang Cao[12], Feng Wu[1], Zheng-Jun Zha[1]†
[1]University of Science and Technology of China, Hefei, China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{tgc1997, hanh}@mail.ustc.edu.cn, {ywang120, forrest, fengwu, zhazj}@ustc.edu.cn
Project page: https://sites.google.com/view/event-based-lipreading

Abstract

Automatic lip-reading (ALR) aims to recognize words using visual information from the speaker's lip movements. In this work, we introduce a novel type of sensing device, event cameras, for the task of ALR. Event cameras have both technical and application advantages over conventional cameras for the ALR task because they have higher temporal resolution, less redundant visual information, and lower power consumption. To recognize words from the event data, we propose a novel Multi-grained Spatio-Temporal Features Perceived Network (MSTP) to perceive fine-grained spatio-temporal features from microsecond time-resolved event data. Specifically, a multi-branch network architecture is designed, in which different grained spatio-temporal features are learned by operating at different frame rates. The branch operating on the low frame rate can perceive spatial complete but temporal coarse features. While the branch operating on the high frame rate can perceive spatial coarse but temporal refinement features. And a message flow module is devised to integrate the features from different branches, leading to perceiving more discriminative spatio-temporal features. In addition, we present the first event-based lip-reading dataset (DVS-Lip) captured by the event camera. Experimental results demonstrated the superiority of the proposed model compared to the state-of-the-art event-based action recognition models and video-based lip-reading models.

1. Introduction

Automatic lip-reading (ALR), also known as visual language recognition, aims to decode the text content through the visual information of the speaker's lip movements. ALR has great applications in biometric identification [2], improved hearing aids [40], speech recognition in noisy environments [29], so that it has attracted much attention in the field of computer vision and pattern recognition over a long period.

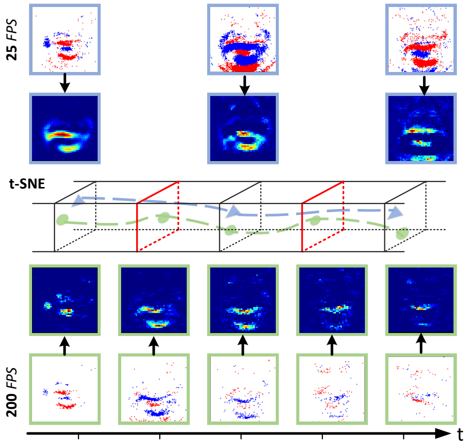


Figure 1. Visualization of event frames of different temporal resolutions (25FPS and 200FPS, respectively) and their corresponding feature maps and feature points that have been dimensionally reduced by t-SNE [41]. The low-rate event frames contain complete spatial features but coarse temporal features, while the high-rate event frames contain fine temporal features but incomplete spatial features.

In this paper, we introduce a novel type of optical sensor, *event cameras* [18], to tackle automatic lip-reading problem. Event cameras are biologically inspired optical sensors. Unlike conventional cameras that capture images at a fixed rate, event cameras capture per-pixel brightness changes asynchronously in the microsecond level. For the ALR task that requires the perception of fine-grained spatio-temporal features, event cameras have significant advantages over conventional cameras in terms of technology and applications: 1) the high temporal resolution of event cameras allow them to record finer-grained movements; 2) their output does not contain much redundant visual information since only brightness changes of the scene are recorded; 3) they are low-power and can work on challenging lighting conditions which are essential in real-world applications.

*Co-first author. †Corresponding author.

To correctly recognize words from the event stream, it is necessary to perceive fine-grained spatio-temporal features from the event stream. There are already a number of event-based action recognition methods [3, 6, 38, 42–44]. Point-cloud-based [42] and graph-based methods [6, 44] treat events data as point clouds and graph nodes, respectively. However, during the conversion from the original event data to the point clouds or graph nodes, the fine-grained temporal and spatial information contained in the event data is discarded. SNN-based methods [3, 38] process the input events stream asynchronously with spiking neural networks, but they are difficult to train since no efficient back-propagation algorithm exists. Existing CNN-based methods [19, 43] convert the asynchronous event data into fixed-rate frame-like representations and feed them into standard deep neural networks. They also lose varying degrees of spatial or temporal information depending on the temporal resolution of the event frames. In summary, existing event-based action recognition methods are not suitable for the ALR task, which requires the perception of fine-grained spatio-temporal features from the event data.

In this work, we choose to convert the event data into multi-grained event frames. As illustrated in Figure 1, the low-rate event frames contain complete spatial features but coarse temporal features due to high temporal compression, while the high-rate event frames contain fine temporal features but incomplete spatial features since each event frame is composed of a small number of events. To take full advantage of the abundant spatio-temporal information in the event data, we propose a Multi-grained Spatio-Temporal Features Perceived Network (MSTP) that takes multi-grained event frames as input to recognize words. The proposed model contains two branches, of which the first branch takes the low-rate event frames as input, allowing the model to perceive complete spatial structure information. In contrast, the second branch takes the high-rate event frames as input, enabling the model to perceive fine temporal features. Furthermore, we devise a message flow module (MFM) to merge multi-grained spatio-temporal features learned by different branches, leading to perceiving more discriminative spatio-temporal features.

Due to the lack of available datasets for event-based lip-reading, we collected the first event-based lip-reading dataset (called DVS-Lip) using event camera *DAVIS346*. The DVS-Lip contains a total of 19,871 samples, an example of which is shown in Figure 3. To explore the advantages of event cameras in capturing fine-grained movement evolution information, we divided the vocabulary of the DVS-Lip dataset into two parts. The first part consists of the 25 pairs of visually similar words selected from the LRW dataset [11], and the second part consists of another 50 randomly selected words from the vocabulary of the LRW dataset. More details can be found in Sect. 4.1.

We validate the effectiveness of the proposed MSTP by conducting extensive experiments on the DVS-Lip dataset. Quantitative results show that: 1) MSTP outperforms existing event-based action recognition models and the state-of-the-art video-based lip-reading models on the proposed dataset for both common and visually similar words; 2) The proposed message flow module has a more significant improvement on visually similar words recognition, thus it is beneficial to perceive fine-grained spatio-temporal features.

The major contributions of our work can be summarized in the following four aspects:

- To the best of our knowledge, it is the first work to study event-based automatic lip-reading. And we propose a novel event-based automatic lip-reading framework MSTP to perceive multi-grained spatio-temporal features for words recognition.
- We devise a message flow module to merge multi-grained spatio-temporal features for more discriminative features perceiving.
- Considering the lack of a relevant benchmark, we collected the first event-based lip-reading dataset (DVS-Lip), which will be made available to the community.
- Extensive experiments conducted on the DVS-Lip dataset show that the proposed method outperforms the state-of-the-art event-based and video-based methods.

2. Related Work

In this section, we first investigate existing lip-reading datasets and relevant lip-reading methods. Then we introduce event cameras and existing event-based action recognition methods.

Lip-Reading Datasets. Depending on the recognition object, existing lip-reading datasets can be divided into alphabet recognition datasets [13, 27], digit recognition datasets [4, 28, 31], word recognition datasets [11, 14, 48], and sentence recognition datasets [1, 10, 12, 37]. All of these datasets are recorded by conventional cameras. In this paper, we focus on word recognition with datasets recorded by the event camera.

Lip-Reading Methods. State-of-the-art lip-reading methods are based on deep learning techniques [5, 8, 10, 11, 23, 26, 36, 37]. They leverage end-to-end deep neural networks to extract visual features and then perform word classification. [11] propose a multiple towers architecture that they first extract shallow visual features of each frame with 2D convolutions, then concatenate all features. Finally, several 3D convolution layers are used to extract the global visual features of the video. [17, 26] employ convolutional neural networks (CNN) as visual feature extractor and then leverage recurrent neural networks (RNN) [17] or temporal convolutional networks (TCN) [26] to model long-term dependency.

Event Cameras. Event cameras are neuromorphically inspired dynamic vision sensors (DVS) [18], different from conventional cameras that capture images at a fixed-rate, event cameras capture per-pixel brightness changes asynchronously in the microsecond level. Event cameras have attractive advantages over conventional cameras: high temporal resolution (order of μs), low power consumption (order of 10mW), high dynamic range (140 dB), and high pixel bandwidth (order of kHz). The type of event camera we used to collect data is *DAVIS346*, which can simultaneously output the event stream and intensity images.

Event-based Action Recognition Methods. Event cameras have been used in a wide range of applications [3, 21, 30, 33, 39, 43], among which the most relevant to ours are gesture recognition [3] and gait recognition [43]. Same as lip-reading, they are also a type of action recognition [7, 24, 45, 47] task. Point-cloud-based method [42] treats events as space-time event clouds and then leverages PointNet++ [32] as the feature extractor to extract events features. Graph-based methods [6, 44] transform events into a set of connected nodes and then use GNNs to extract the spatio-temporal features of events. SNN-based methods [3, 38] process the input events stream asynchronously with spiking neural networks, but they are difficult to train since no efficient back-propagation algorithm exists. CNN-based methods [19,43] convert the asynchronous events into fixed-rate frames and feed them into standard deep neural networks.

3. Method

In this section, we first briefly describe the output format of the event camera in Sect. 3.1. Then, we introduce the proposed Multi-grained Spatio-Temporal Features Perceived Network (MSTP) in Sect. 3.2.

3.1. Event Data

An event camera outputs asynchronous event data independently at each pixel, it will trigger an event at a specific pixel $\mathbf{u} = (x, y)$ whenever the log brightness change at (x, y) reaches the contrast threshold, *i.e.*,

$$log\mathcal{I}(x, y, t) - log\mathcal{I}(x, y, t - \Delta t) = pC, \tag{1}$$

where $\mathcal{I}$ is the brightness of the scene, $p \in \{-1, 1\}$ is the polarity of the change in brightness, C is the contrast threshold, and Δt is the time elapsed since the last event triggered at (x, y). In a given period of time $[T_0, T_1]$, the output of the event camera can be formulated as:

$$\mathcal{E} = \{e_k\}_{k=1}^{N} = \{(x_k, y_k, t_k, p_k)\}_{k=1}^{N}, T_0 \leq t_k \leq T_1, \tag{2}$$

which are scattered points in the space-time dimension.

3.2. Framework

As shown in Figure 2, we propose a novel event-based lip-reading framework MSTP. Our framework contains three components: 1) projection between the raw event streams and frame-like representations; 2) a multi-branch network with message flow modules (MFM) between different branches; 3) a sequence model that decodes the visual features into words.

3.2.1 Event Representation

One of the most challenging problems for event-based tasks is how to design an effective event representation. As analyzed in Sect. 1, it is not optimal to treat event data as point clouds or graph nodes or directly process them using SNN for ALR. In this paper, we choose to convert asynchronous event data into synchronous frame-like representations.

Here we adopt the voxel grid [49], where each event distributes its polarity p to the two closest spatio-temporal voxels, to represent the event data. Given a set of N input events $\mathcal{E} = \{(x_k, y_k, t_k, p_k)\}_{k=1}^{N}$ and the temporal bin T, the voxel grid approach first scales the timestamps to the range $[0, T-1]$, then generates event frames $\mathcal{V}$ with dimension $T \times H \times W$ as follows:

$$t_k^* = \frac{T-1}{t_N - t_1}(t_k - t_1), \tag{3}$$

$$\mathcal{V}(t, y, x) = \sum_k p_k max(0, 1 - |t - t_k^*|). \tag{4}$$

In contrast to previous works [33, 49] that use one fixed temporal bin in their voxel grid representation, we use multiple temporal bins to keep the spatio-temporal information of the event stream better. According to Eq.3 and Eq.4, we convert the input events $\mathcal{E} = \{(x_k, y_k, t_k, p_k)\}_{k=1}^{N}$ into low-rate event frames $\mathcal{V}^{low}$ with temporal bin set to T^{low} and high-rate event frames $\mathcal{V}^{high}$ with temporal bin set to T^{high}.

3.2.2 Multi-branch Networks

Event frames with different temporal resolutions belong to the same modality but contain spatio-temporal information with different granularities. The high-rate event frames contain fine temporal information but incomplete spatial information, and the opposite is true for the low-rate event frames as shown in Figure 1. To take full advantage of the abundant spatio-temporal information in the event data, we propose a multi-branch network with message flow modules between different branches, each branch taking event frames of different temporal resolutions as input. This network architecture allows the model to perceive both complete spatial features and fine-grained temporal features simultaneously.

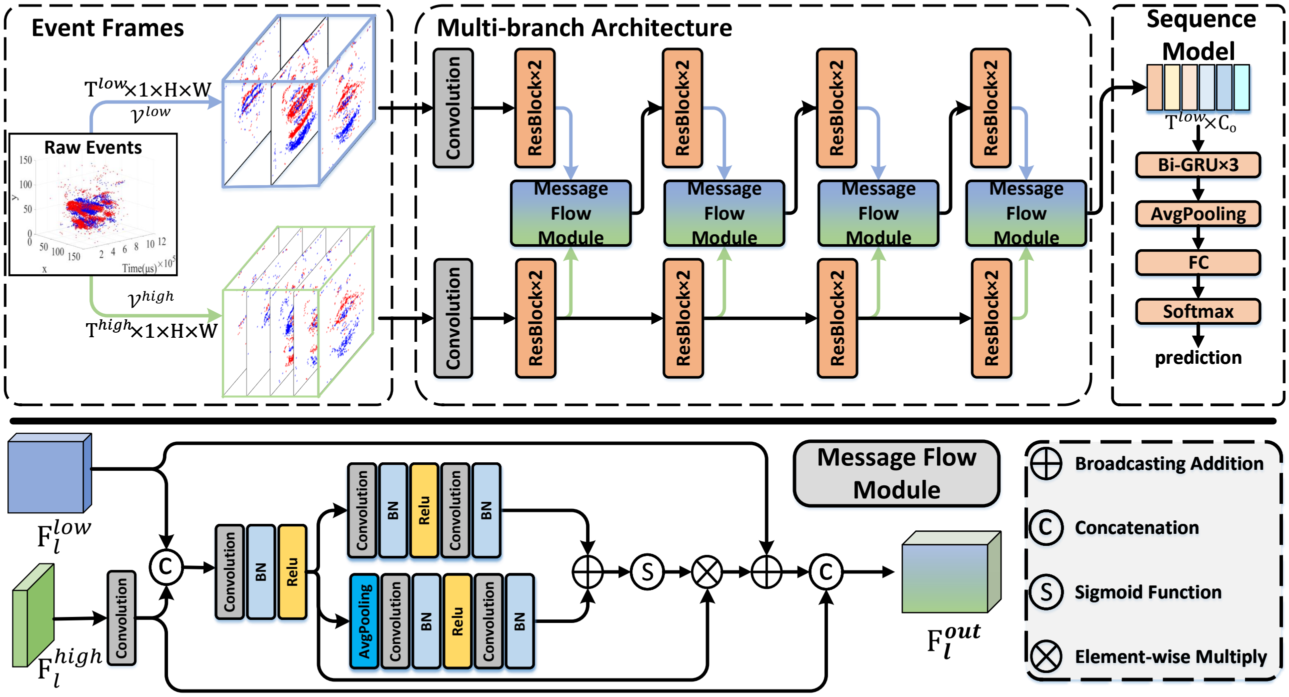


Figure 2. (**Top**) The architecture of our proposed Multi-grained Spatio-Temporal Features Perceived Network (MSTP). MSTP consists of three components: 1) projection from asynchronous raw events to multi-grained synchronous event frames; 2) a multi-branch network with message flow modules between different branches, which is designed to perceive both complete spatial features and fine temporal features from the event data; 3) a sequence model for decoding the visual features into words. (**Bottom**) Message Flow Module (MFM). MFM uses features learned by different branches to obtain an attention weight map to merge multi-grained spatio-temporal features.

Specifically, each branch of the model is composed of a 3D convolutional layer and several residual blocks [20]. To maintain the fine-grained temporal features, we preserve the temporal dimension of each branch. The difference between them is that the number of channels in the convolutional layers is not the same. Since the high time dimension leads to a significant increase in computation, we use a smaller channel in the high-rate branch to reduce the computation. As studied in [15], convolutional layers with lower channel capacity can weaken the spatial modeling ability while strengthening the temporal modeling ability and significantly reducing the computation. The number of channels in the convolutional layers of the low-rate branch is four times higher than that of the high-rate branch in our implementation.

Different branches of the model focus on learning spatio-temporal features at different granularities. To perceive more discriminative spatio-temporal features for lip-reading, we devise a message flow module to enable message flow between different branches as illustrated at the bottom of Figure 2. Let $\{\mathbf{F}_l^{low} \in \mathbb{R}^{T^{low}\times C_l^{low}\times H_l\times W_l}\}_{l=1}^{L}$ and $\{\mathbf{F}_l^{high} \in \mathbb{R}^{T^{high}\times C_l^{high}\times H_l\times W_l}\}_{l=1}^{L}$ be the output features of the low-rate branch and the high-rate branch respectively, where l denotes the layer in which we obtain the features, L denotes the total layer numbers, T^{low} and T^{high} denote the temporal bins of each branch, C_l^{low} and C_l^{high} denote the number of channels of the l-th layer of each branch, H_l and W_l denote the height and width of the output features from l-th layer. At l-th layer, we first downsample $\mathbf{F}_l^{high}$ along the temporal dimension by a temporal convolutional layer. Then concatenate the features from two branches and use a convolutional layer to fuse them, the fused features $\mathbf{F}_l^{fuse}$ are obtained by:

$$\mathbf{F}_l^{fuse} = \delta(BN(g_{lf}([g_{ld}(\mathbf{F}_l^{high});\mathbf{F}_l^{low}]))), \tag{5}$$

where g_* is the convolutional layer and the same blow, δ is the $Relu$ function, BN is the batch normalization layer, and $[;]$ denotes the concatenation operation.

The features from the higher-rate branch contain finer temporal information that can guide the learning of the lower-rate branch to focus on key parts of the features. To merge features learned by different branches, we use the fused features $\mathbf{F}_l^{fuse}$ to compute two attention contexts. The first attention context is a local one:

$$\mathbf{A}_l^{tchw} = BN(g_{l2}(\delta(BN(g_{l1}(\mathbf{F}_l^{fuse}))))), \tag{6}$$

where $\mathbf{A}_l^{tchw} \in \mathbb{R}^{T^{low}\times C_l^f\times H_l\times W_l}$ is the local attention context, C_l^f is the number of channels of $\mathbf{F}_l^{fuse}$.

The second attention context is computed in temporal-channel-wise, where a global average pooling is applied in

$\mathbf{F}_l^{fuse}$ along spatial dimension before computing the attention context, formally:

$$\mathbf{A}_l^{tc} = BN(g_{l4}(\delta(BN(g_{l3}(GAP_{hw}(\mathbf{F}_l^{fuse})))))), \tag{7}$$

where $\mathbf{A}_l^{tc} \in \mathbb{R}^{T^{low} \times C_l^f \times 1 \times 1}$ is the global attention context, and GAP_{hw} is the global average pooling along spatial dimension.

Then the attention map is computed by:

$$\mathbf{Att}_l = \sigma(\mathbf{A}_l^{tchw} \bigoplus \mathbf{A}_l^{tc}), \tag{8}$$

where σ denotes the $Sigmoid$ function, $\bigoplus$ denotes the broadcasting addition.

The attention based augmented features $\widetilde{\mathbf{F}}_l$ is computed by:

$$\widetilde{\mathbf{F}}_l = \mathbf{F}_l^{low} + \mathbf{F}_l^{fuse} \bigotimes \mathbf{Att}_l, \tag{9}$$

where $\bigotimes$ denotes the element-wise multiplication.

To keep the original information from the higher-rate branch, the downsampled features of $\mathbf{F}_l^{high}$ are concatenate with $\widetilde{\mathbf{F}}_l$ as the output features of the message flow module:

$$\mathbf{F}_l^{out} = [\widetilde{\mathbf{F}}_l; g_{ld}(\mathbf{F}_l^{high})]. \tag{10}$$

3.2.3 Sequence Model

Different from previous multi-branch networks [15, 16], we employ a sequence model as the backend of our framework.The sequence model we use is a bidirectional Gate Recurrent Unit [9], which takes the output of the multi-branch network, *i.e.*, $\mathbf{F}_L^{out} \in \mathbb{R}^{T^{low} \times C_o}$, as input. $\mathbf{F}_L^{out}$ is treated as a sequence of length T^{low}, and the dimension of the feature at each time step is C_o. Let $\mathcal{P} \in \mathbb{R}^{voc\_size}$ be the output probabilities of each word in the vocabulary, where voc_size is the vocabulary size. Then $\mathcal{P}$ is computed by

$$\mathcal{P} = Softmax(FC(GAP_t(BiGRU(\mathbf{F}_L^{out})))), \tag{11}$$

where $BiGRU$ is a three-layer bidirectional Gate Recurrent Unit, GAP_t is the global average pooling along temporal dimension of the output features of the $BiGRU$, FC is a fully connection layer that converts the final visual feature into probability logit. Finally, $\mathcal{P}$ is computed by the $Softmax$ function.

4. Experiments

In this section, we first detail the dataset collection process in Sect. 4.1. Then, we describe the implementation details of our proposed model in Sect. 4.2. Next, we compare our proposed method with several existing event-based action recognition methods and the state-of-the-art video-based method in Sect. 4.3.1. We then conduct some ablation studies to verify the effectiveness of each part of the proposed model and study the impact of the temporal bin on the model in Sect. 4.3.2. Finally, we show some qualitative results of our model in Sect. 4.3.3.

4.1. Dataset Collection

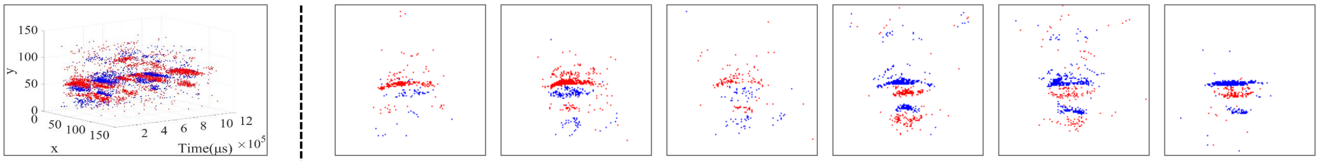
Due to the lack of available datasets for event-based lip-reading, we collected the first event-based lip-reading dataset DVS-Lip, using the event camera. The type of event camera we used is *DAVIS346*, which can simultaneously output the event stream and intensity images with a spatial resolution of 346×260 without viewpoint differences. To explore the advantages of event cameras in capturing fine-grained movement evolution information, we divide the vocabulary of the DVS-Lip dataset into two parts, where the first part is composed of visually similar word pairs and the second part is composed of common words. The first part of the vocabulary consists of the 25 most frequently confusing word pairs that are selected from the vocabulary (500 words in total) of the LRW dataset [11]. Refer to the supplementary materials for a more detailed selection of words included in the first part of the vocabulary. For the second part of our vocabulary, we randomly select another 50 words from the vocabulary of the LRW dataset. Combining the two parts, the vocabulary of the DVS-Lip dataset contains a total of 100 words. We have listed all the words contained in the vocabulary in the supplementary materials.

We recruited 40 volunteers to participate in the recording of our dataset in indoor scene, 20 of each gender. We first constructed 5 sequences, each containing all words in the vocabulary. To avoid the volunteers reading the same word too similarly, each sequence was randomly disrupted. Thus for each word, the words before and after it in each sequence are different. The volunteers were then asked to sit in front of the event camera and read each of the five word sequences once. If a word is mispronounced, we ask the volunteers to reread the sequence until there are no errors. We simultaneously recorded the audios corresponding to these words, which were used to split the data of each sequence into word-level samples. We used the Montreal Forced Aligner[*] to get the start and end time of each word according to the corresponding audio. For comparison with the video-based approach, we kept both the event streams and the intensity images (25FPS) output from the event camera. And we used the face detection tool[†] to obtain the position of the face and mouth and then extracted a mouth-centered crop of size 128×128 pixels.

A total of 200 word sequences from 40 volunteers were recorded, each containing 100 words after word-level segmentation. However, due to the damage of a small part of the event data files, we finally obtained 19,871 valid word samples, an example of which is shown in Figure 3. We use the data of 30 volunteers (including 14,896 samples of 15 males and 15 females) for training and the remainder (including 4,975 samples of 5 males and 5 females) for evalu-

[*] https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

[†] https://github.com/ageitgey/face_recognition



(a) Raw Events

(b) Event Frames

Figure 3. Visualization of an example in the DVS-Lip dataset. Red points denote the positive ($p = 1$) events while blue points denote the negative ($p = -1$) events. (a) Visualization of asynchronous raw events. (b) Visualization of synchronous frame-like event representation corresponding to the raw events.

Dataset	classes	Speakers	Utterances	Language	Source
MIRACL-VC [34]	10	15	1500	English	Recorded by RGB-D Camera
MODALITY [14]	182	35	231	English	Recorded by stereo camera
LRW [11]	500	1000+	550,000	English	Crawling from TV programs
LRW1000 [48]	1000	2000+	745,187	Chinese	Crawling from TV programs
DVS-Lip	100	40	19,871	English	Recorded by event camera

Table 1. Statistics of the DVS-Lip dataset, compared with previous video-based word-level lip-reading datasets.

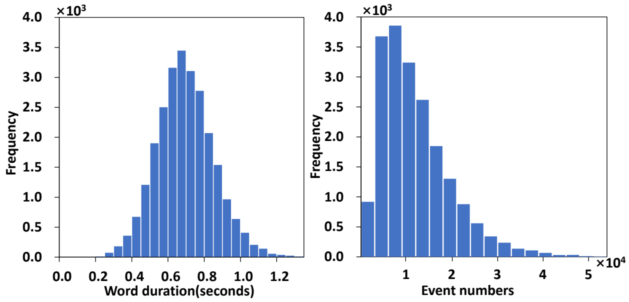


Figure 4. Word statistics. (left) Word duration. (right) Event numbers of each sample.

ation. Therefore, the speakers corresponding to the training set and the test set do not overlap. The training set and test set can also be divided into two parts according to which part of the vocabulary the word comes from. As a result, 7,441 samples belong to the first part and 7,455 samples belong to the second part of the training set, while 2,493 samples belong to the first part and 2,482 samples belong to the second part of the test set. Table 1 shows the comparison between our dataset and previous word-level lip-reading datasets. In addition, we counted the duration of words and the number of events contained in each word, as illustrated in Figure 4.

4.2. Implementation Details

All the experiments in this work are conducted on our DVS-Lip dataset. The input spatial dimension of our proposed model is 88×88, so we first perform central cropping of size 96×96 on the original data. And then, we random crop the event frames to 88×88 and random flip them with a probability of 0.5 horizontally for data argumentation in the training phase. While for testing, we center crop the test data to 88×88. For the training and evaluation of video-based methods, if the number of frames contained in each video clip is larger than 30, we linearly sample 30 of them. Otherwise, we pad them to the length of 30 with the zero-padding operation. We set the maximum number of frames per video to 30 for that most of the videos are shorter than 1.20 seconds according to Figure 4 and all videos have a frame rate of 25FPS. To keep consistent with the number of video frames, we set the temporal bin of the input event frames of the low-rate branch to 30. Thus, our model's low-rate branch input has the same temporal resolution as the video-based methods.

We use the PyTorch[‡] framework to implement all the methods used in this work. Our model is optimized by Adam Optimizer [22] with standard settings. We use the cosine annealing scheduler [25] to control the learning rate during training, in which the initial learning rate is set to 3e-4 and the minimum learning rate is set to 5e-6. We set the batch size to 32 and trained our model for 80 epochs.

4.3. Experimental Results

4.3.1 Comparisons with State-of-the-art Methods

We compare the proposed Multi-grained Spatio-Temporal Features Perceived Network (MSTP) with several relevant action recognition methods, including 1) event-based action recognition methods [19,42–44]; 2) video-based action recognition methods [7,24,45]; 3) video-based lip-reading methods [17,26,46].

Quantitative comparison results of the above methods and our model on the DVS-Lip test set are shown in Table 2. We can find that our proposed MSTP significantly outperforms existing event-based and video-based action recognition methods on both parts of the test set. It indicates that our proposed method has a much stronger ability to perceive fine-grained spatio-temporal features from the event data

[‡] https://pytorch.org/

Model	Input	Temporal Bin	BackBone	Acc1 (%)	Acc2 (%)	Acc (%)
Event Clouds [42]	event	-	PointNet++	35.82	48.51	42.15
EV-Gait-3DGraph [44]	event	-	3DGraph	26.35	37.75	32.04
EV-Gait-IMG [43]	event	-	-	17.20	25.66	21.42
EV-Gait-IMG* [43]	event	30	-	28.80	40.21	34.49
EST [19]	event	30	ResNet-34	40.91	56.45	48.66
I3D [7]	event	30	InceptionV1	58.24	77.68	67.94
TANet [24]	event	30	ResNet-101	58.36	79.17	68.74
ACTION-Net [45]	event	30	ResNet-50	58.32	79.41	68.84
DFTN [46]	video	30	ResNet-18	52.63	73.73	63.16
Feng *et al.* [17]	video	30	ResNet-18	54.23	72.60	63.40
Martinez *et al.* [26]	video	30	ResNet-18	55.60	75.46	65.51
MSTP	event	(30, 210)	ResNet-18	**62.17**	**82.07**	**72.10**

* We use the event frames with the temporal bin set to 30 instead of the four-channel representations in the original paper.

Table 2. Comparisons with existing event-based models and the state-of-the-art video-based models on the DVS-Lip test set. **Temporal bin** denotes the temporal dimension of the input event frames or video clip. **Acc1** denotes the accuracy on the first part of the test set, **Acc2** denotes the accuracy on the second part of the test set, and **Acc** denotes the accuracy on the entire test set.

Model	Temporal Bin	Acc1 (%)	Acc2 (%)	Acc (%)
Low-rate Branch	30	58.84	80.34	69.57
High-rate Branch	210	59.04	79.96	69.49
MSTP(w/o MFM)	(30, 210)	60.69	81.59	71.11
MSTP	(30, 210)	**62.17**	**82.07**	**72.10**

Table 3. Effectiveness of each part of our MSTP.

than existing action recognition methods. There are obvious flaws in these approaches. Point-cloud-based method [42] and graph-based method [44] suffer from downsampling operation. During the conversion from the original event data to the point clouds or graph nodes, the fine-grained temporal and spatial information will be discarded, leading them to perceive indiscriminative spatio-temporal features. Existing CNN-based methods [7, 19, 24, 43, 45] only use fixed frame rate event frames as input, so they are not suitable for tasks such as lip-reading that require the perception of fine-grained spatio-temporal features. In contrast, our MSTP can learn both complete spatial features and fine temporal features by feeding multi-grained event frames into different branches.

It is worth noting that the proposed model also outperforms the state-of-the-art video-based lip-reading models. There are two reasons for it. On the one hand, the event data contains less redundant visual information due to recording only brightness changes. On the other hand, our MSTP can successfully exploit abundant spatio-temporal information in sparse event data.

4.3.2 Ablation Study

To verify the effectiveness of each part of the proposed model and study the effect of the temporal bin on the proposed model, we conducted the following two groups of experiments.

As shown in Table 3, we compare our MSTP with a set of ablated models with various settings: 1) **Low-rate Branch**: using only the low-rate branch of the proposed model; 2) **High-rate Branch**: using only the high-rate branch of the proposed model; 3) **MSTP(w/o MFM)**: discarding the message flow module and using the lateral connection between different branches like previous works [15, 16].

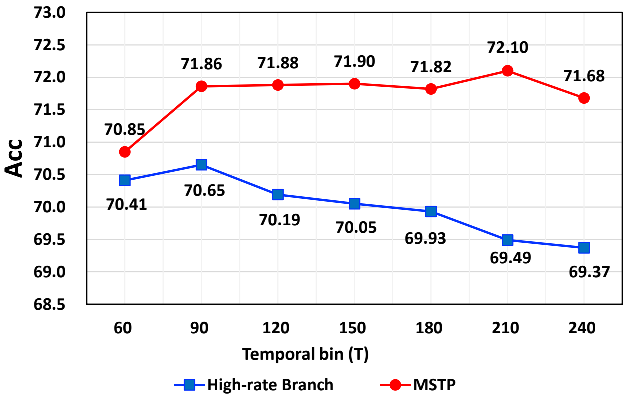
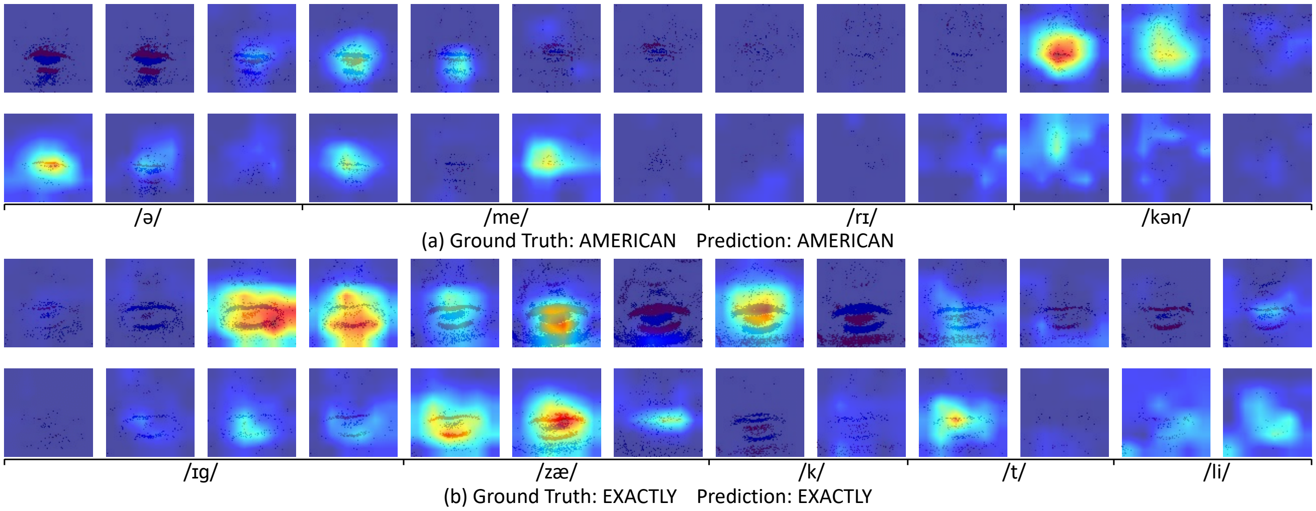


Figure 5. Effect of the temporal bin on High-rate Branch and our full MSTP.

According to the results in Table 3, we have the following observations. Firstly, multi-branch models outperform single-branch models. This is because multi-branch models take event frames of different temporal resolutions as input, leading to learning complete spatial features and fine-grained temporal features simultaneously. Secondly, our full MSTP outperforms MSTP(w/o MFM), which demonstrates that integrating features from different branches with our designed message flow module can help the model to perceive more discriminative spatio-temporal features. Thirdly, the relative improvement on the first part of the test set is more significant than that of the second part. For example, comparing Low-rate Branch with MSTP, the accuracy of the first part improved by 3.33 from 58.84 to 62.17, while the accuracy of the second part improved by only 1.73 from 80.34 to 82.07. This indicates that our pro-



(a) Ground Truth: AMERICAN Prediction: AMERICAN

(b) Ground Truth: EXACTLY Prediction: EXACTLY

Figure 6. Visualization of the saliency maps for words (a) “American” and (b) “exactly”. The first row of each example shows the saliency maps for the low-rate branch’s input event frames ($T^{low} = 30$), and the second row shows the saliency maps for the input event frames ($T^{high} = 210$) of the high-rate branch.

posed MSTP and the message flow module are effective in perceiving fine-grained spatio-temporal features.

In addition, we studied the effect of the temporal bin for High-rate Branch and MSTP as illustrated in Figure 5. We can observe that the performance of High-rate Branch becomes better and then worse as the temporal bin increases. This is because when the temporal bin is low, the event frames will discard fine-grained temporal information due to high temporal compression. When the temporal bin is high, the spatial structure of the event frames will be corrupted because each event frame is composed of a small number of events. And the best compromise between complete spatial information and fine temporal information is achieved when the temporal bin is 90. For our MSTP, the best performance is achieved when the temporal bin of the high-rate branch is 210. The low-rate branch of the MSTP can provide complete spatial features so that the MSTP can perceive finer-grained spatio-temporal features as the temporal bin of the high-rate branch increases. The results from Figure 5 demonstrate that our MSTP can benefit from inputting multi-grained event frames, which enable our model to learn both complete spatial features and fine temporal features.

4.3.3 Qualitative Analysis

For the qualitative analysis, we apply the Grad-CAM [35] to our MSTP using the samples from the DVS-Lip test set. Grad-CAM result shows visual saliency regions clearly by calculating gradients with respect to a unique class. Two examples are shown in Figure 6. For each example, the first row shows the saliency maps for the input of the low-rate branch, and the second row shows the saliency maps for the input of the high-rate branch. The temporal bin for the low-rate branch is 30, and the temporal bin for the high-rate branch is 210. Due to the limited space, we downsample the saliency maps for the high-rate branch by a factor of 7. At the same time, the downsampling operation can make the saliency maps from the two branches aligned in time.

In Figure 6, we can see that the model has correctly learned to focus on phonologically important regions in the event frames and the saliency maps from different branches complement each other. This indicates that our MSTP can automatically select important spatio-temporal regions for word recognition from event frame inputs of different granularities. Accordingly, the model can learn both complete spatial features and fine temporal features.

5. Conclusion

This paper proposes a novel Multi-grained Spatio-Temporal Features Perceived Network (MSTP) for event-based lip-reading. MSTP takes multi-grained event frames as input through the design of multi-branch architecture, enabling it to perceive both complete spatial features and fine-grained temporal features. And a message flow module is devised to perceive more discriminative spatio-temporal features. In addition, we collected the first event-based lip-reading dataset (DVS-Lip) and will make it available to the community. According to the evaluation on the test set of the DVS-Lip, our proposed MSTP is significantly superior to the state-of-the-art event-based and video-based methods.

Acknowledgments This work was supported by the National Key R&D Program of China under Grant 2020AAA0105700, National Natural Science Foundation of China (NSFC) under Grant U19B2038, the University Synergy Innovation Program of Anhui Province under Grants GXXT-2019-025.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018. 2

[2] Zahid Akhtar, Christian Micheloni, and Gian Luca Foresti. Biometric liveness detection: Challenges and research opportunities. *IEEE Security & Privacy*, 13(5):63–72, 2015. 1

[3] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. A low power, fully event-based gesture recognition system. In *CVPR*, pages 7243–7252, 2017. 2, 3

[4] Iryna Anina, Ziheng Zhou, Guoying Zhao, and Matti Pietikäinen. Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–5. IEEE, 2015. 2

[5] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*, 2016. 2

[6] Yin Bi, Aaron Chadha, Alhabib Abbas, Eirina Bourtsoulatze, and Yiannis Andreopoulos. Graph-based spatio-temporal feature learning for neuromorphic vision sensing. *IEEE Transactions on Image Processing*, 29:9084–9098, 2020. 2, 3

[7] Joao Carreira *et al.* Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 3, 6, 7

[8] Weicong Chen, Xu Tan, Yingce Xia, Tao Qin, Yu Wang, and Tie-Yan Liu. Duallip: A system for joint lip reading and generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1985–1993, 2020. 2

[9] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 5

[10] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3453. IEEE, 2017. 2

[11] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian conference on computer vision*, pages 87–103. Springer, 2016. 2, 5, 6

[12] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006. 2

[13] Stephen J Cox, Richard W Harvey, Yuxuan Lan, Jacob L Newman, and Barry-John Theobald. The challenge of multispeaker lip-reading. In *AVSP*, pages 179–184. Citeseer, 2008. 2

[14] Andrzej Czyzewski, Bozena Kostek, Piotr Bratoszewski, Jozef Kotus, and Marcin Szykulski. An audio-visual corpus for multimodal automatic speech recognition. *Journal of Intelligent Information Systems*, 49(2):167–192, 2017. 2, 6

[15] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 4, 5, 7

[16] Christoph Feichtenhofer, Axel Pinz, and Richard Wildes. Spatiotemporal residual networks for video action recognition. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3468–3476, 2016. 5, 7

[17] Dalu Feng, Shuang Yang, Shiguang Shan, and Xilin Chen. Learn an effective lip reading model without pains. *arXiv preprint arXiv:2011.07557*, 2020. 2, 6, 7

[18] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *TPAMI*, 2020. 1, 3

[19] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5633–5643, 2019. 2, 3, 6, 7

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[21] Yongcheng Jing, Yiding Yang, Xinchao Wang, Mingli Song, and Dacheng Tao. Turning frequency to resolution: Video super-resolution via event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7772–7781, 2021. 3

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[23] Zhijie Lin, Zhou Zhao, Haoyuan Li, Jinglin Liu, Meng Zhang, Xingshan Zeng, and Xiaofei He. Simullr: Simultaneous lip reading transducer with attention-guided adaptive memory. *arXiv preprint arXiv:2108.13630*, 2021. 2

[24] Zhaoyang Liu *et al.* Tam: Temporal adaptive module for video recognition. In *ICCV*, 2021. 3, 6, 7

[25] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *ICLR*, 2017. 6

[26] Brais Martinez *et al.* Lipreading using temporal convolutional networks. In *ICASSP*, 2020. 2, 6, 7

[27] Iain Matthews, Timothy F Cootes, J Andrew Bangham, Stephen Cox, and Richard Harvey. Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):198–213, 2002. 2

[28] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, Gilbert Maitre, et al. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, volume 964, pages 965–966. Citeseer, 1999. 2

[29] Chalapathy Neti, Gerasimos Potamianos, Juergen Luettin, Iain Matthews, Herve Glotin, Dimitra Vergyri, June Sison, and Azad Mashari. Audio visual speech recognition. Technical report, IDIAP, 2000. 1

[30] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Bringing a blurry frame

alive at high frame-rate with an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6820–6829, 2019. 3

[31] Eric K Patterson, Sabri Gurbuz, Zekeriya Tufekci, and John N Gowdy. Cuave: A new audio-visual database for multimodal human-computer interface research. In *2002 IEEE International conference on acoustics, speech, and signal processing*, volume 2, pages II–2017. IEEE, 2002. 2

[32] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017. 3

[33] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 3

[34] Ahmed Rekik, Achraf Ben-Hamadou, and Walid Mahdi. A new visual speech recognition approach for rgb-d cameras. In *International conference image analysis and recognition*, pages 21–28. Springer, 2014. 6

[35] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 8

[36] Alireza Sepas-Moghaddam, Fernando Pereira, Paulo Lobato Correia, and Ali Etemad. Multi-perspective lstm for joint visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16540–16548, 2021. 2

[37] Brendan Shillingford, Yannis Assael, Matthew W Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, et al. Large-scale visual speech recognition. *arXiv preprint arXiv:1807.05162*, 2018. 2

[38] Sumit Bam Shrestha and Garrick Orchard. Slayer: Spike layer error reassignment in time. *NeurIPS*, 2018. 2, 3

[39] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16155–16164, 2021. 3

[40] Nancy Tye-Murray, Mitchell S Sommers, and Brent Spehar. Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. *Ear and hearing*, 28(5):656–668, 2007. 1

[41] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 1

[42] Qinyi Wang, Yexin Zhang, Junsong Yuan, and Yilong Lu. Space-time event clouds for gesture recognition: From rgb cameras to event cameras. In *2019 IEEE Winter Conference on Applications of Computer Vision*, pages 1826–1835. IEEE, 2019. 2, 3, 6, 7

[43] Yanxiang Wang, Bowen Du, Yiran Shen, Kai Wu, Guangrong Zhao, Jianguo Sun, and Hongkai Wen. Ev-gait: Event-based robust gait recognition using dynamic vision sensors. In *CVPR*, pages 6358–6367, 2019. 2, 3, 6, 7

[44] Yanxiang Wang, Xian Zhang, Yiran Shen, Bowen Du, Guangrong Zhao, Lizhen Cui Cui Lizhen, and Hongkai Wen. Event-stream representation for human gaits identification using deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 3, 6, 7

[45] Zhengwei *et al.* Wang. Action-net: Multipath excitation for action recognition. In *CVPR*, 2021. 3, 6, 7

[46] Jingyun Xiao *et al.* Deformation flow based two-stream network for lip reading. In *FG*, 2020. 6, 7

[47] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[48] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin 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. 2, 6

[49] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019. 3