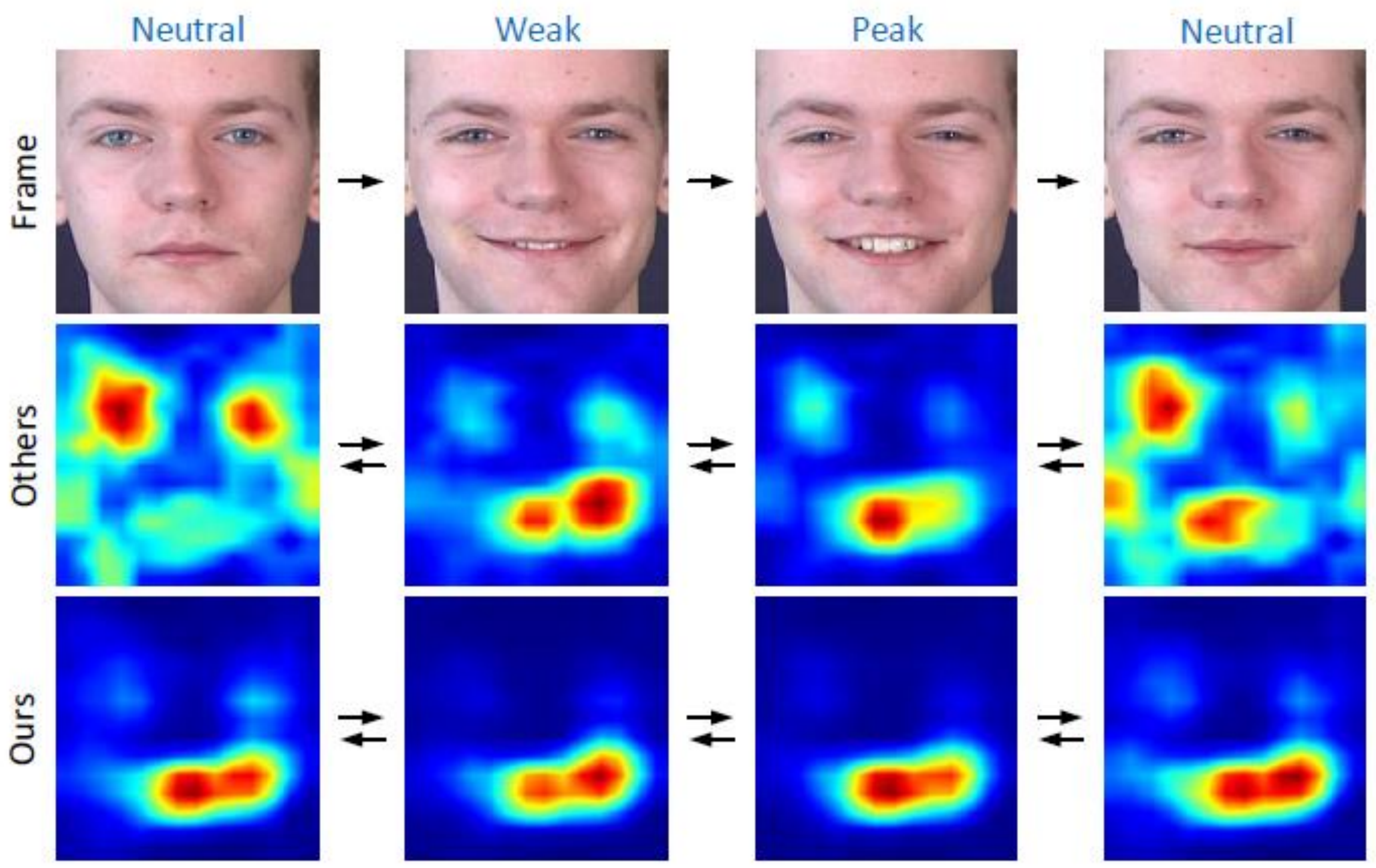
Video-based Facial Expression Recognition using Graph Convolutional Networks

Daizong Liu[1], Hongting Zhang[1], Pan Zhou[1]

[1]Huazhong University of Science and Technology

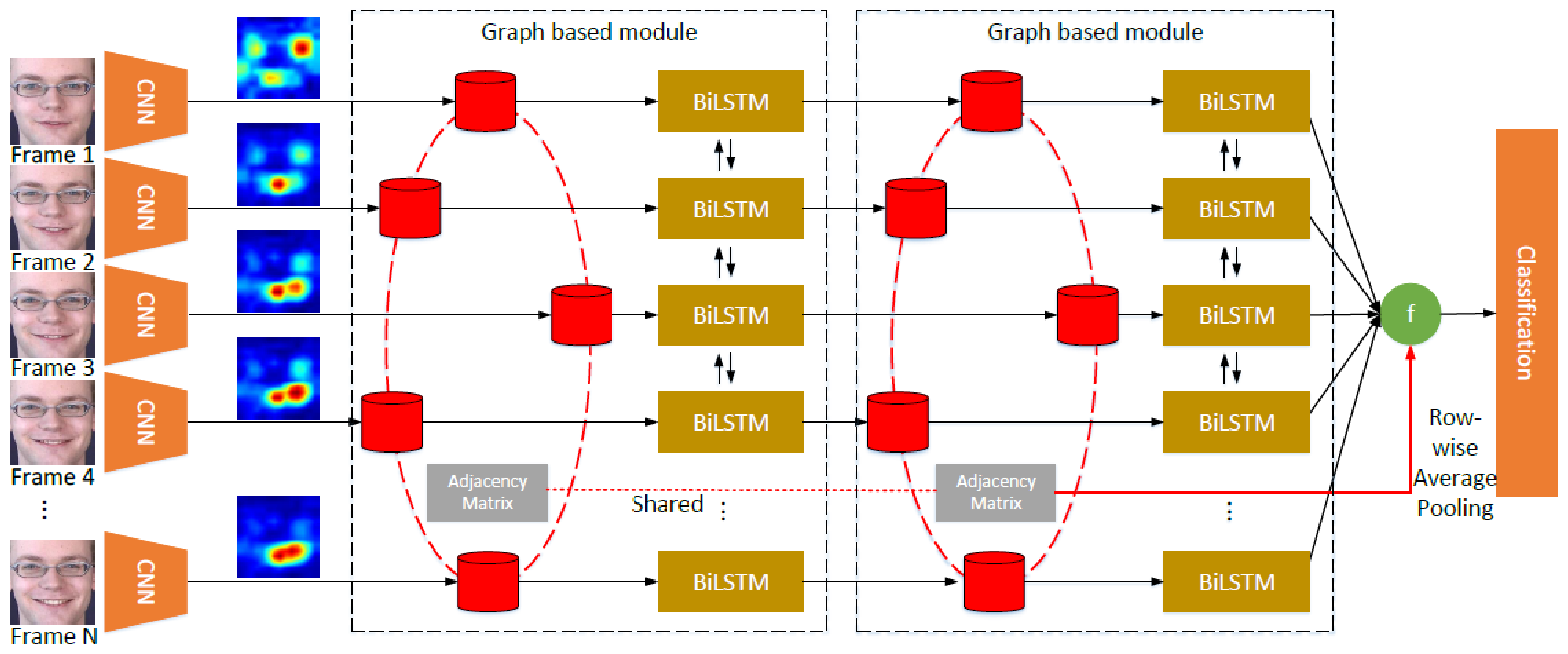
Introduction:



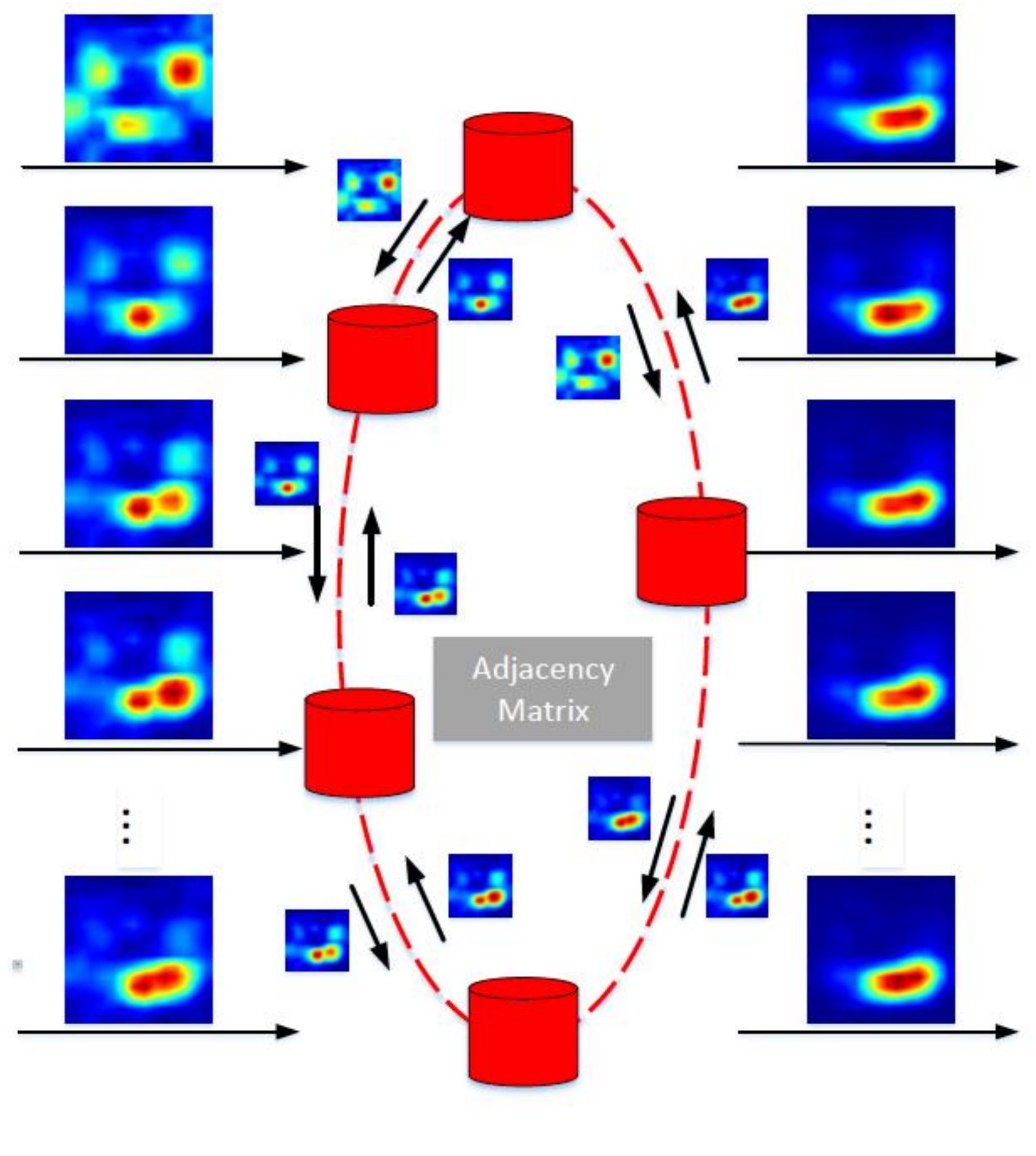
Most of existing works aim to find out the most contributing expression features with each frame and take it as an image-based task by assembling these features to model the facial activation. However, the individual features they learned from each frame focus on different regions, because the facial expression intensity on different regions is dynamically changing among the video frames. Such features can only contribute limited strength to explore the dynamic variation of expression as they do not concentrate on the facial activation in an certain expression region.

Moreover, the features coming from peak frames usually focus on important regions which have more contributing information than those of non-peak frames. They should be considered more for final recognition.

Pipeline:



Pipeline



Graph learning

$$n_i = [H_1^{\mathrm{T}}\ H_2^{\mathrm{T}}\ \ldots\ H_{i-1}^{\mathrm{T}}\ H_{i+1}^{\mathrm{T}}\ \ldots\ H_N^{\mathrm{T}}]^{\mathrm{T}}$$

$$M_i^l = n_i \boldsymbol{W}^l$$

$$\boldsymbol{A}_{i\bar{i}} = [\boldsymbol{A}_{i1}, \boldsymbol{A}_{i2}, \cdots, \boldsymbol{A}_{i(i-1)}, \boldsymbol{A}_{i(i+1)}, \cdots, \boldsymbol{A}_{in}]$$

$$o_i^{l+1} = f(\boldsymbol{A}_{i\bar{i}} M_i^l \oplus \boldsymbol{A}_{ii} H_i \boldsymbol{W}^l)$$

$$\boldsymbol{A}^{l+1} = \boldsymbol{A}^l - lr * \partial loss/\partial \boldsymbol{A}^l$$

$$weight = softmax(mean(\boldsymbol{A}, dim = 0))$$

$$r = \sum_{i=1}^{N} weight_i H_i$$

Experiments:

TABLE I: Average accuracy on the CK+, Oulu-CASIA and MMI datasets respectively.

Method	CK+	Oulu	MMI	Feature
Inception [13]	93.20%	-	77.60%	static
IACNN [22]	95.37%	-	71.55%	static
DLP-CNN [23]	95.78%	-	-	static
FN2EN [24]	96.80%	87.71%	-	static
DeRL [25]	97.30%	88.00%	73.23%	static
PPDN [15]	99.30%	84.59%	-	static
3DCNN [14]	85.90%	-	53.20%	Dynamic
ITBN [26]	86.30%	-	59.70%	Dynamic
HOG 3D [27]	91.44%	70.63%	60.89%	Dynamic
TMS [28]	91.89%	-	-	Dynamic
3DCNN-DAP [14]	92.40%	-	63.40%	Dynamic
STM-ExpLet [29]	94.19%	74.59%	75.12%	Dynamic
LOMo [30]	95.10%	82.10%	-	Dynamic
3D Inception-Resnet [31]	95.53%	-	79.26%	Dynamic
Traj. on S+(2, n) [32]	96.87%	83.13%	79.19%	Dynamic
DTAGN [33]	97.25%	81.46%	70.24%	Dynamic
GCNet [34]	97.93%	86.11%	81.53%	Dynamic
PHRNN-MSCNN [6]	98.50%	86.25%	81.18%	Dynamic
Ours	**99.54%**	**91.04%**	**85.89%**	Dynamic

TABLE V: Ablation study on the individual components.

Experiment model	CK+	Oulu-CASIA	MMI
VGG16	97.78%	85.83%	80.75%
VGG16 + graph based spatial-temporal module×1	98.39%	88.33%	84.37%
VGG16 + graph based spatial-temporal module×2	99.09%	89.79%	84.64%
VGG16 + graph based spatial-temporal module×3	99.00%	87.71%	83.07%
VGG16 + graph based spatial-temporal module×2 + weighted feature fusion	**99.54%**	**91.04%**	**85.89%**

TABLE II: Confusion matrix of recognizing four expressions on CK+ dataset.

	An	Co	Di	Fe	Ha	Sa	Su
An	**100%**	0%	0%	0%	0%	0%	0%
Co	0%	**100%**	0%	0%	0%	0%	0%
Di	0%	0%	**100%**	0%	0%	0%	0%
Fe	0%	0%	0%	**100%**	0%	0%	0%
Ha	0%	0%	0%	0%	**100%**	0%	0%
Sa	0%	0%	0%	0%	0%	**100%**	0%
Su	0%	1%	0%	0%	0%	0%	**99%**

TABLE III: Confusion matrix of recognizing four expressions on Oulu-CASIA dataset.

	An	Di	Fe	Ha	Sa	Su
An	**88%**	7%	1%	0%	3%	1%
Di	10%	**84%**	2%	0%	3%	1%
Fe	0%	0%	**91%**	4%	1%	4%
Ha	0%	0%	2%	**98%**	0%	0%
Sa	4%	4%	1%	0%	**90%**	1%
Su	0%	0%	4%	0%	1%	**95%**

TABLE IV: Confusion matrix of recognizing four expressions on MMI dataset.

	An	Di	Fe	Ha	Sa	Su
An	**77%**	13%	0%	0%	10%	0%
Di	3%	**91%**	3%	0%	3%	0%
Fe	4%	0%	**68%**	4%	4%	20%
Ha	0%	0%	2%	**98%**	0%	0%
Sa	9%	0%	0%	0%	**91%**	0%
Su	0%	0%	10%	0%	2%	**88%**

TABLE VI: Recognition accuracy of each single model on the validation dataset of AFEW 8.0.

Method	Accuracy
Emotiw2018 (baseline) [37]	38.81%
HoloNet [39]	46.50%
DSN-VGG-Face [40]	48.04%
Resne50-LSTM [38]	49.31%
DenseNet161-pool5 [41]	51.44%
VGG-Face-LSTM [38]	53.91%
Ours	**55.67%**

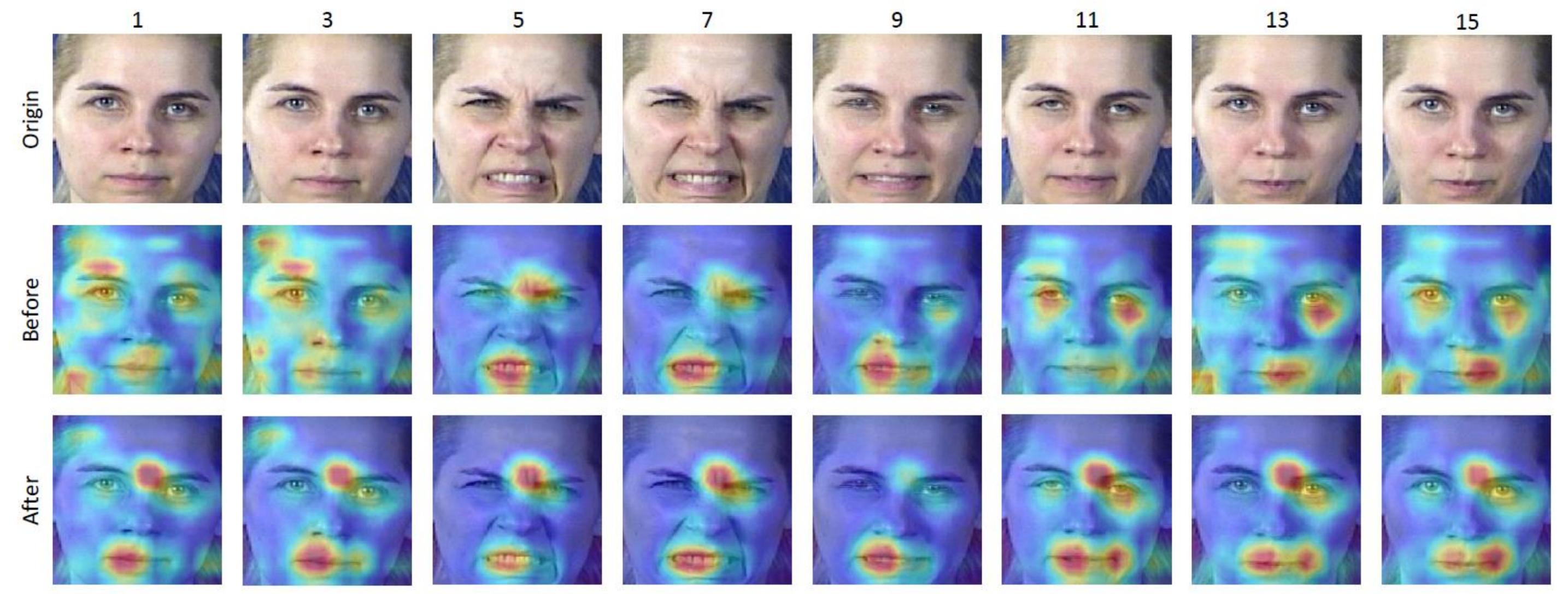
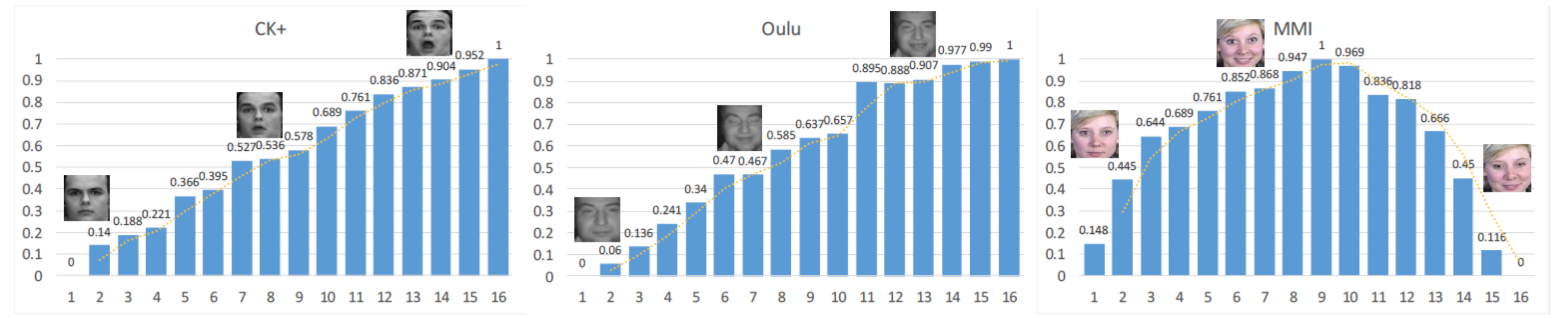


Fig. 3: Example of the feature reconstruction in our GCN layer. First row: Origin facial images of "Disgust" in MMI dataset; Second row: input features of GCN layer; Third row: output features of GCN layer. It clarifies that our GCN layer shares most contributing expression features among frames to helps model focus more on the corresponding expression regions (such as mouth and nose here).



(a) CK+ (b) Oulu-CASIA (c) MMI

Fig. 4: Visualization of expression intensity weights for 16 steps on three datasets respectively. The horizontal axis represents the step number in each video sequence. The values of temporal weighs are given in the vertical axis through a sigmoid function, which refer to the expression intensity of each frame in the dynamic expression variation.