Supplementary Materials for FERV39k: A Large-Scale Multi-Scene Dataset for Facial Expression Recognition in Videos

Yan Wang[1], Yixuan Sun[1], Yiwen Huang[2], Zhongying Liu[2], Shuyong Gao[2], Wei Zhang[2], Weifeng Ge[2,*] and Wenqiang Zhang[1,2,*]
[1]Academy of Engineering & Technology, Fudan University, Shanghai, China
[2]School of Computer Science, Fudan University, Shanghai, China
{wfge, wqzhang}@fudan.edu.cn

1. Annotation Documentation

In this paper, we build a large-scale multi-scene dataset (FERV39k) for FER in videos. There are four isolated scenarios subdivided into 22 scenes: 6 scenes {Argue, Social, School, Medicine, Conflict, and Daily-Life} designed for Daily Life (DL11k), 6 scenes {Action, Scholar-Reports, Speech, Elegant-Art, Live-Show, and Talk-Show} for Weak-Interactive Shows (WIS9k), 6 scenes {Business, Experiment, Official-Event, Crime, Interview, Contest} for Strong-Interactive Activities (SIA10k) and 4 scenes {History, Terror, War and Crisis} for Anomaly Issues (AI9k). It tends to be ambiguous and remains as a big challenge to build DFER datasets with the complexity and variability of spatial-temporal dynamics. To ensure the consistency of annotation of 7 basic expressions in different scenes, we write a handbook (also called annotation documentation) to clarify the definition of 7 basic expressions ("Angry", "Disgust", "Fear", "Happy", "Sad", "Surprise", and "Neutral"). Afterwards, we formulate representative examples with 7 basic expressions across 22 scenes.

1.1. Daily Life (DL11k)

DL11k scenario is composed of 6 scenes commonly experienced in the daily life. In this scenario, scenes vary from each other a lot due to the complexity of real-life activities. Figure 3 shows an overview of 7 expressions in 6 scenes.

1.2. Weak-Interactive Shows (WIS9k)

WIS9k scenario is composed of 6 kinds of shows. The person in scenes of WIS9k usually maintains a consistent emotional state over a long period of time. Besides, the intensity of expressions is much higher. Figure 4 shows an overview of 7 expressions in 6 show scenes.

1.3. Strong-Interactive Activities (SIA10k)

SIA10k scenario mainly focuses on activities with strong interaction. In these scenes, the emotion of a person is usually influenced by other people and environment. As a result, the distribution of expressions shows great instability and diversity. Figure 5 shows an overview of 7 expressions with great diversity in 6 scenes.

1.4. Anomaly Issues (AI9k)

AI9k scenario contains 4 hardly seen scenes in our daily life. It is difficult for both researchers and DFER methods to distinguish unexpected appearances and changes of an expression in these scenes. Figure 6 shows some unusual appearances with 7 basic expressions.

2. Generation of Candidate Video Clips

After reviewing top-level 22 scenes, we collect scenes corresponding online videos, TV shows and movies from open search engines. The first step of building a dynamic dataset is candidate video clip selection. The main problem for us is to acquire available full context and single face clips. Existing works ask annotators to manually segment video clips with expressions via video editing software. However, it is costly to do so when segmenting many raw videos into several qualified video clips among 0.5∼4 seconds. Hence, we design a four-stage strategy to collect and generate candidate video clips.

2.1. Rule-based Selection Mechanism

At first, we randomly split each raw video into many video clips among 0.5∼4 seconds according to this work [2]. It is impossible to annotate millions of clips randomly generated from thousands of raw videos. Hence, we need to make some rules to discard even $\frac{19}{20}$ raw clips in some scenes such as Interview, Speech, etc. In addition, the selected clips tend to contain many samples with same person. In order to further generate finer candidate video clips, we make a rule list to help our well-designed mechanism to adaptively select satisfactory clips from a twenty-fold

amount of clips than expected scale of final dataset. The rule list is included as:

- Our algorithm can automatically segment a raw video into many video clips containing only one face lasting for 0.5∼4 seconds by removing such clips: multi-face, small face, vitural face, captions and picture in picture. Some examples in Figure 1 show that six kinds of such video clips need to be removed by both our algorithm and annotators.
- Only one person can appear in one video clip, no other people are allowed to appear. We use a face matching technique to completely remove such video clips with identity variation in a frame sequence. Some examples in Figure 2 show that such video clips are expcted to be removed by both our algorithm and annotators.
- More than 90% frames in one video clip must contain detectable faces of the same person.
- The total clip latency should follow the previous statistic researches and the number of selected clips from each video should follow an average total duration distribution. A further random selection method is used to fulfill this purpose.

2.2. FER-based detector

According to the above rule-based algorithm, we can find that the majority expressions of video clips generated from the raw videos are Neutral and Happy. In order to balance the data distribution, we train a well-designed and light-weight FER detector with high accuracy for recognizing in-the-wild facial expressions. As Real-world Affective Face Database (RAF-DB) [6] contains about 30,000 real-world facial images, we use RAF-DB to train our FER detector based on ResNet50 model. The FER-based detector is implemented in PyTorch-GPU using GeForce RTX 2080ti GPUs. The learning rate lr is initialized to 1e-3, and lr is exponentially decayed by a factor of 0.95 every epoch. The stochastic gradient descent (SGD) is optimized with 0.9 Momentum and 1e-3 weight decay. Batch size is fixed at 32. The maximum epoch number is 120. Finally, the FER-based detector achieves the overall accuracy of 87.53% of 7 basic expressions, which shows the comparable performance with other state-of-the-art methods [5, 11] in limited computation sources. As a result, we utilize this FER-based detector to refine these clips with threshold value to generate relatively balanced candidate video clips for 7 basic expressions.

3. Annotation Workflow

In our designed procedure, there are two roles named crowd-sourcing annotators (20 workers) and professional researchers (10 workers), respectively. The crowd-sourcing annotators are relatively cost-less but with lower reliability. And the professional researchers cost more for labeling a clip but with the highest reliability (over the crowd-sourcing annotators, statistics-based inspection methods and even the administrators). Before annotating candidate video clips, two teams are both provided with a handbook. And the personalized designed platform is used for annotation. Five steps and two stages (annotation and judgement) in total are designed for the annotation work-flow:

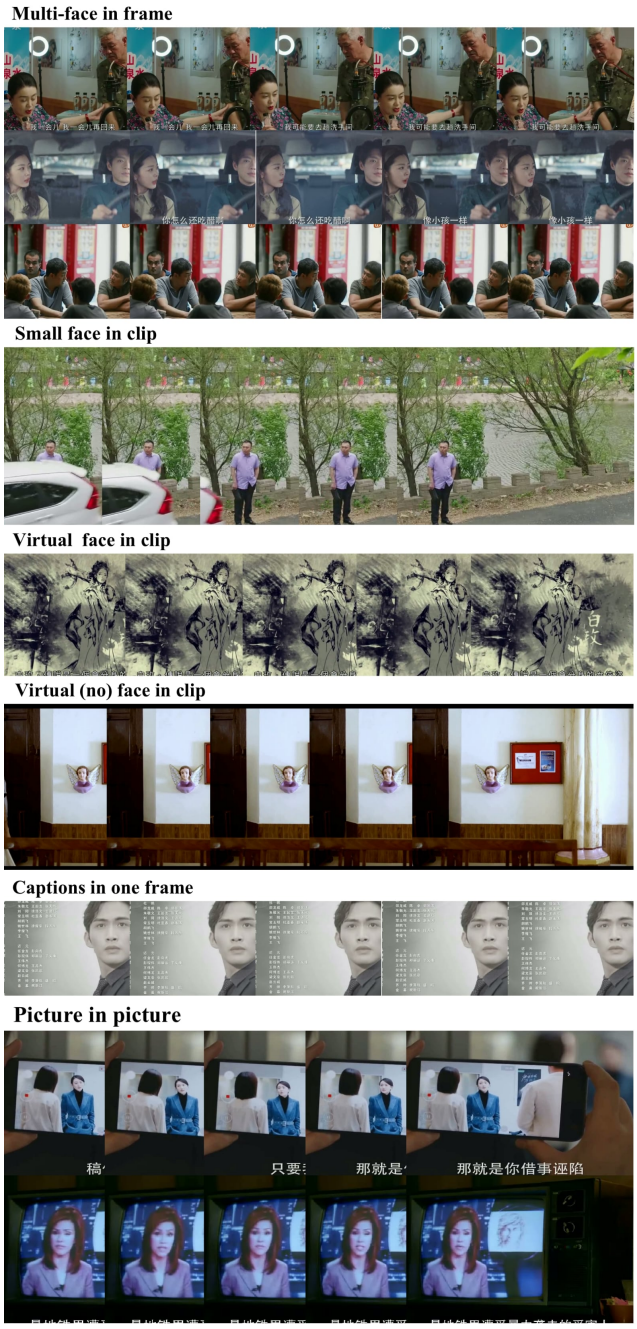

Figure 1. Six kinds of video clips need to be removed due to violating the rule of only one face in one frame.

Figure 2. Multi-face in one video.

- Data grouping and flag presetting (annotation stage): Before annotation, clips are randomly divided into several groups and 5% of clips in each group are annotated by the professional researchers and mixed into the raw materials (labels are hidden to crowd-sourcing annotators). Afterwards, the mixed groups are copied for 3 times. The copies of a specific group have unique group IDs which are invisible to annotators. Then we randomly shuffle the grouped materials and provide them to the crowd-sourcing annotators.
- Annotation (annotation stage): The annotators are asked to choose the most likely expression from 26 fine-grained labels on the platform. Each label corresponds to an expression in "Angry", "Disgust", "Fear", "Happy", "Sad", "Surprise" and "Neutral" and the chosen label can be automatically converted through the platform. For the samples without proper correspondence, annotators can press 'PASS' and the clips will be marked as illegal.
- Auto-checking via error statistics (judgement stage): With the pipeline of annotation, the annotated materials are firstly checked via the flag-recapture based error statistics method [1]. In this step, we collect all the groups and calculate correct rate of preset check data as the total correct rate. We design a two-level threshold of 40% and 80%. The group with correct rate lower than 40% will be marked as unacceptable and retreated to crowd-sourcing annotators; The group with correct rate between 40% and 80% will be marked as improper with additional warnings and passed to professional judgment; And the group with over 80% correct rate will be marked as accept and passed to professional judgment without notes.
- Professional judgment (judgement stage): Instead of choosing the corresponding expression from the multiple labels, the professional researchers only need to decide whether the labels are proper to the clips and whether the annotation of a group is reliable. The unreliable (unaccepted) ones will also be retreated but with extra instructions, and the improper ones will be relabeled by the professional researchers as the supplementary annotation, and a notice and advice will feedback to crowd-sourcing annotators.
- Final decision (judgement stage): A weighted winner-takes-all (WWTA) voting mechanism is used. During the voting, supplementary annotation have twice higher weight than the normal annotation which means that, the professional researchers have higher confidence but can be refuted by a broader consensus.

4. Agreement

- The FERV39k dataset is available to **non-commercial research purposes** only.
- All videos of the FERV39k dataset are obtained from the Internet which are not property of ***. Our group is not responsible for the content nor the meaning of these videos.
- You agree **not to** reproduce, duplicate, copy, sell, trade, resell or exploit for any commercial purposes, any portion of the videos and any portion of derived data including but not limited to frames, and cropped face images
- You agree **not to** further copy, publish or distribute any portion of the FERV39k dataset. Except, for internal use at a single site within the same organization it is allowed to make copies of the dataset.
- Our group reserves the right to terminate your access to the FERV39k dataset at any time.

5. More Results of Comparisons and Confusion Matrices Under Different Scenes

On top of the FERV39k, we systematically evaluate four kinds of baseline architectures following action recognition baselines [3,4,7] and investigate inter-scene and intra-scene performances based on RS50-LSTM network, we design two kinds of experimental schemes: 1) Training all baselines on all data collected from 22 scenes, and test on the whole dataset, four scenarios, and each scene. 2) Training the RS50-LSTM networks on four scenarios, respectively, and test on four scenarios, and their sub-scenes. We present more results of confusion matrices of different methods, i.e. RS18-LSTM, C3D, Two RS18-LSTM, and Two VGG13-LSTM (all networks training from scratch) on four scenarios, i.e., DL11k, WIS9k, SIA10k, and AI9k in Figure 7, Figure 8, Figure 9, and Figure 10, respectively. According to the results, We observe that existing four kinds of baseline architectures fail to perform well in distinguishing challenging expressions such as 'Fear' and 'Disgust' with variable intensities across different scenes.

5.1. Daily Life (DL11k)

Table 1 shows comparison of four kinds of baseline architectures training on all scenes and then testing on all

scenes, DL11k, and its sub-scenes. The confusion matrices of this experiment is also provided in Figure 7. And Table 2 shows the experiment result of intra-scenario performance consistency and inter-scenario invariance via RS50-LSTM training on four scenarios and testing on DL11k, and its sub-scenes.

5.2. Weak-Interactive Shows (WIS9k)

Table 3 shows comparison of four kinds of baseline architectures training on all scenes and then testing on all scenes, WIS9k, and its sub-scenes. The confusion matrices of this experiment is also provided in Figure 8. And Table 4 shows the experiment result of intra-scenario performance consistency and inter-scenario invariance via RS50-LSTM training on four scenarios and then testing on WIS9k, and its sub-scenes.

5.3. Strong-Interactive Activities (SIA10k)

Table 5 shows comparison of four kinds of baseline architectures training on all scenes and then testing on all scenes, SIA10k, and its sub-scenes. The confusion matrices of this experiment is also provided in Figure 9. And Table 6 shows the experiment result of intra-scenario performance consistency and inter-scenario invariance via RS50-LSTM training on four scenarios and then testing on SIA10k, and its sub-scenes.

5.4. Anomaly Issues (AI9k)

Table 7 shows comparison of four kinds of baseline architectures training on all scenes and then testing on all scenes, AI9k, and its sub-scenes. The confusion matrices of this experiment is also provided in Figure 10. And Table 8 shows the experiment result of intra-scenario performance consistency and inter-scenario invariance via RS50-LSTM training on four scenarios and then testing on AI9k, and its sub-scenes.

5.5. Conculsion

According to above results of four kinds of baseline architectures training on all scenes and then testing on all scenes, four scenarios, and their scenes, we figure out two conclusions: 1) C3D [9] shows the worst results, and VGG13-LSTM achieves the best performance in one-stream network. We consider that C3D [9] fails to capture the temporal information of the limited frames (only 8 frames), but the LSTM can model the global information. 2) Our designed two-stream networks can further improve the performance because the scene context plays an important role in DFER and provides supplemtary information for face-only DFER. In different scenes of our buit FERV39k, confusion matrices of different methods demonstrate that most methods perform best on Happy and perform well on Angry, Sad and Neutral, but are confused in Disgust, Fear, and Surprise due to the limited data and large-scale intensity variations. For cross-domain experiments, the result shows the cross-domain performance of a method is directly related to the feature consistency and intensity of an expression in a scene. For example, WIS9k is designed as a scenario with high similarity and obvious appearance of expressions and the experiment result shows an ideal performance and smaller best-worst difference of each scene among 4 scenarios.

References

[1] Graham Bell. Population estimates from recapture studies in which no recaptures have been made. *Nature*, 248(5449):616–616, 1974. 3

[2] Xianye Ben, Yi Ren, Junping Zhang, Su-Jing Wang, Kidiyo Kpalma, Weixiao Meng, and Yong-Jin Liu. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3, 10, 11, 12

[4] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3

[5] Hangyu Li, Nannan Wang, Xinpeng Ding, Xi Yang, and Xinbo Gao. Adaptively learning facial expression representation via cf labels and distillation. *IEEE Transactions on Image Processing*, 30:2016–2028, 2021. 2

[6] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017. 2

[7] Xin Liu, Silvia L Pintea, Fatemeh Karimi Nejadasl, Olaf Booij, and Jan C van Gemert. No frame left behind: Full video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14892–14901, 2021. 3

[8] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017. 10, 11, 12

[9] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 4, 10, 11, 12

[10] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 10, 11, 12

[11] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020. 2

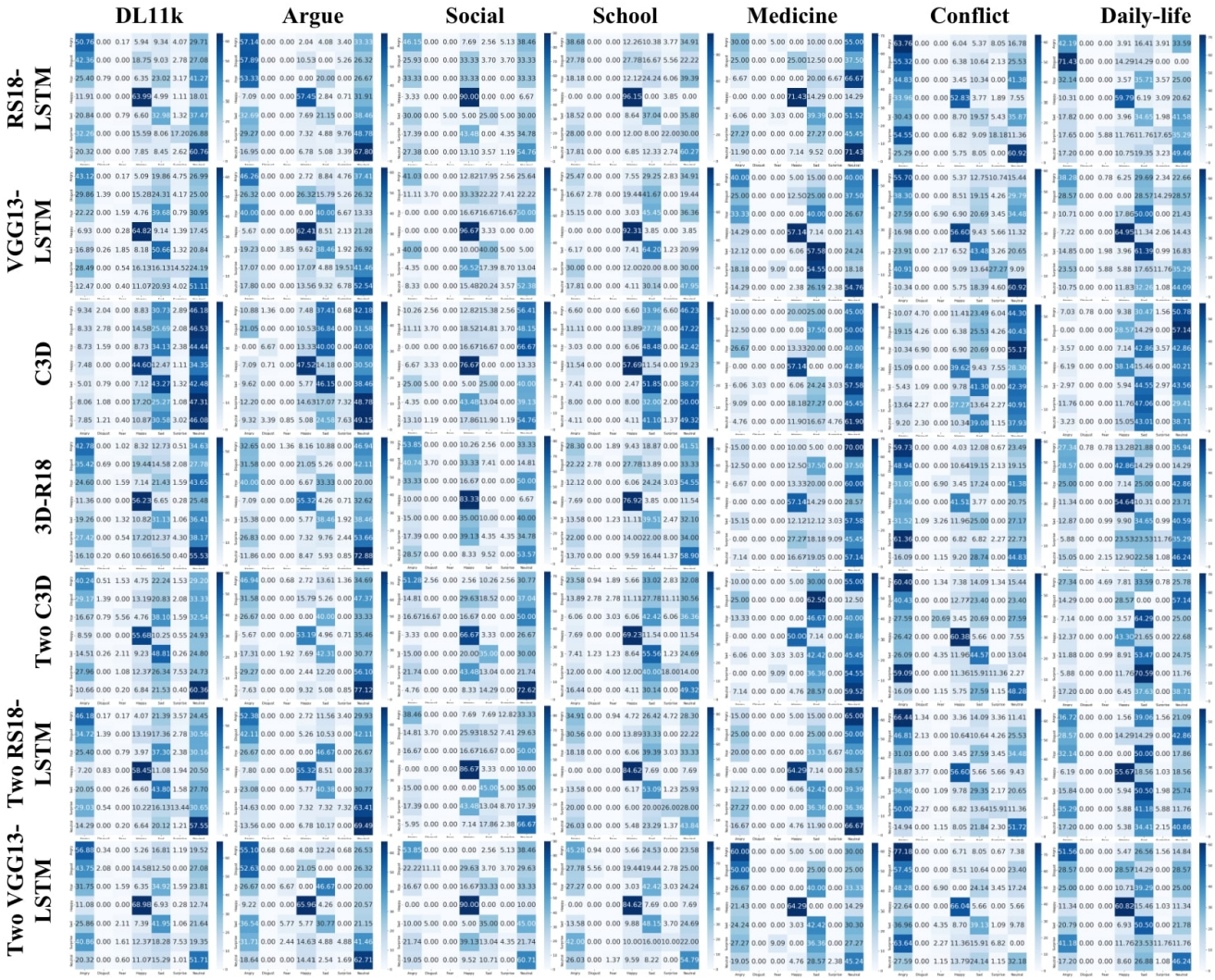

Figure 3. 7 expressions in 6 scenes of DL11k. Each scene show 7 representative frame-level video clips with Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral expression from top to bottom.

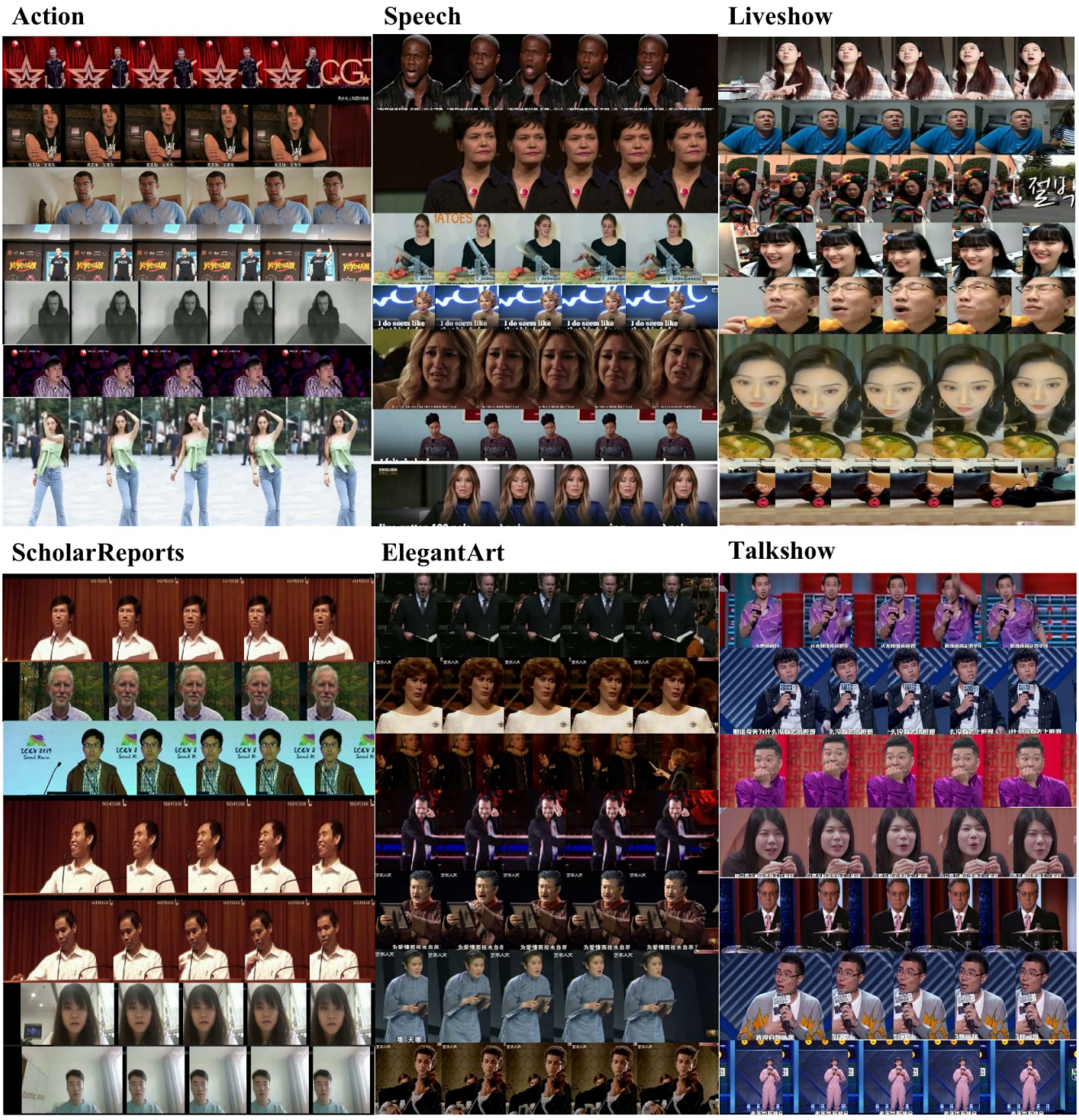

Figure 4. 7 expressions in 6 scenes of WIS9k. Each scene show 7 representative frame-level video clips with Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral expression from top to bottom.

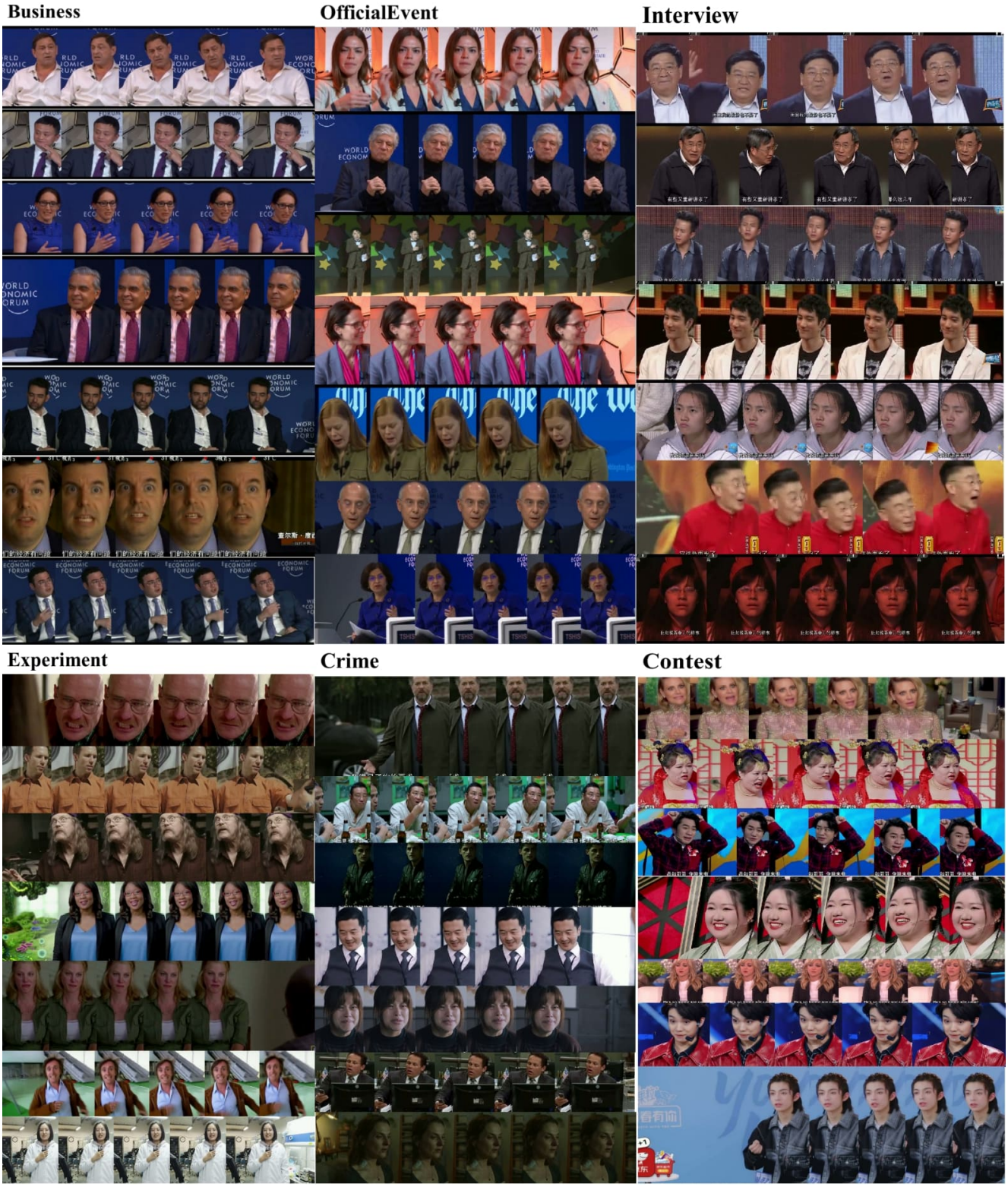

Figure 5. 7 expressions in 6 scenes of SIA10k. Each scene show 7 representative frame-level video clips with Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral expression from top to bottom.

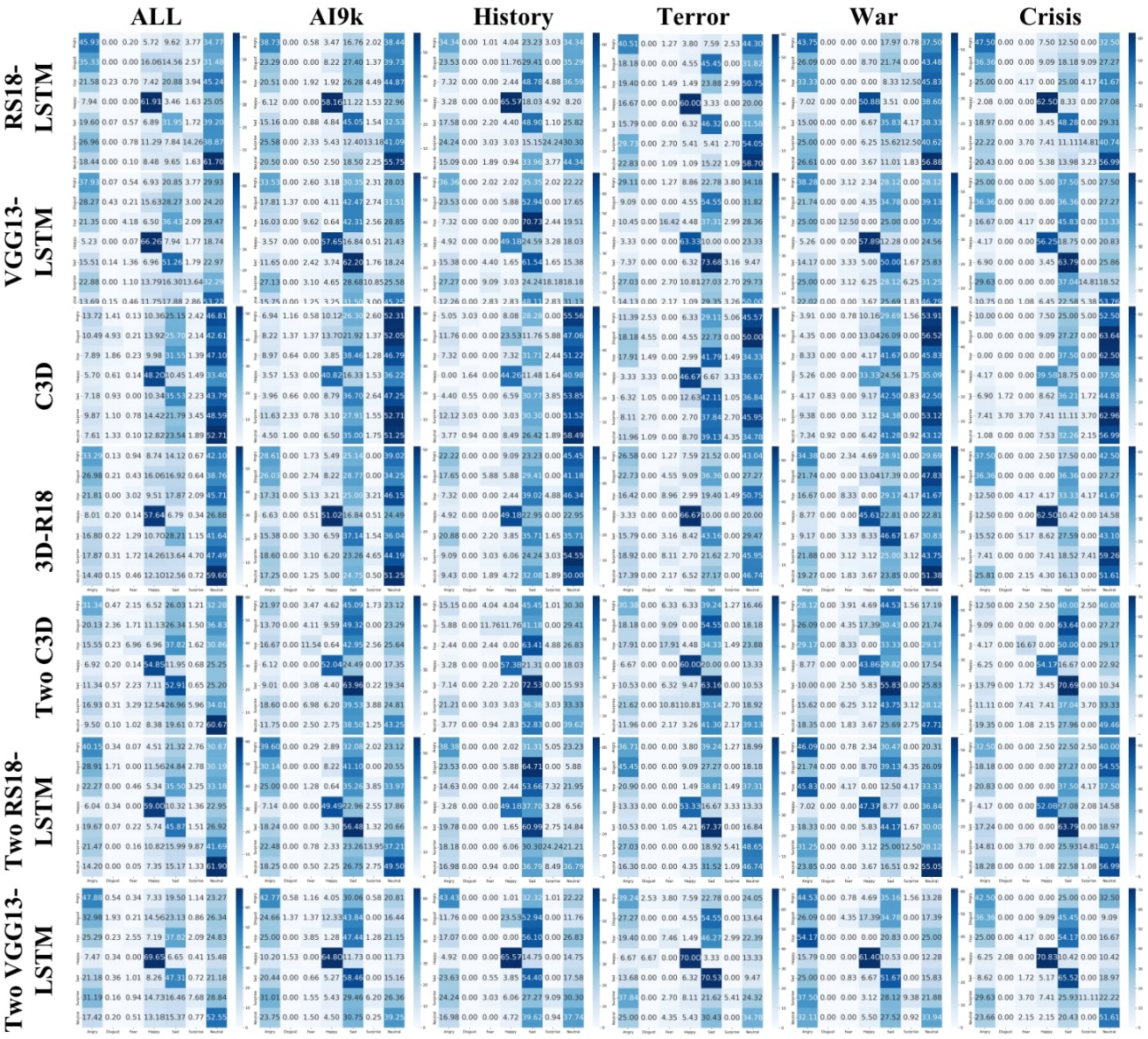

Figure 6. 7 expressions in 4 scenes of AI9k. Each scene show 7 representative frame-level video clips with Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral expression from top to bottom.

Method	All	DL11k	Conflict	School	DailyLife	Argue	Medicine	Social
RS18	39.33/30.30	39.75/31.36	39.52/34.65	35.80/33.80	41.40/31.13	44.09/30.81	36.36/29.02	39.74/33.26
RS50	30.57/22.47	30.46/21.52	31.14/19.23	27.41/22.60	31.00/19.37	36.96/24.07	24.48/20.21	27.51/25.05
VGG13	41.02/31.19	40.40/31.59	36.53/31.32	37.28/34.96	39.07/28.63	42.40/30.36	32.17/23.82	48.03/35.50
VGG16	41.66/32.01	41.81/32.59	41.12/34.28	38.27/36.03	41.19/28.73	45.97/29.75	31.47/25.16	43.23/34.77
RS18-LSTM	42.59/30.92	43.34/32.24	40.32/30.75	37.28/36.31	41.61/29.11	48.78/30.47	41.26/30.32	42.36/31.47
RS50-LSTM	40.75/32.12	40.93/32.91	39.92/33.79	34.32/32.58	41.61/28.00	49.53/35.30	32.87/27.18	42.79/35.70
VGG13-LSTM	43.37/32.41	42.29/32.46	43.91/35.84	35.31/34.39	46.07/31.50	46.15/31.31	40.56/29.93	43.67/34.64
VGG16-LSTM	41.70/30.93	42.99/32.32	41.92/33.32	37.53/35.84	44.37/30.58	47.28/31.40	40.56/31.08	49.34/36.83
C3D [9]	31.69/22.68	26.95/21.02	21.96/19.35	24.94/23.92	26.96/18.35	31.52/23.00	30.77/21.90	34.50/24.34
P3D [8]	33.39/23.20	32.95/23.80	33.53/24.61	24.44/23.03	32.70/21.39	37.34/23.69	28.67/21.04	34.50/25.05
I3D [3]	38.78/30.17	38.56/29.25	34.93/26.11	36.30/32.55	39.70/26.09	43.34/28.84	34.27/26.52	37.55/32.05
3D-RS18 [10]	37.57/26.67	37.69/27.47	35.13/25.75	32.10/30.63	35.67/24.95	43.71/28.82	27.97/21.50	41.48/29.83
Two C3D	41.77/30.72	41.45/31.37	43.11/35.10	33.33/31.64	35.46/23.26	48.22/31.37	33.57/23.14	47.16/32.22
Two I3D	41.30/31.01	41.02/31.55	38.92/32.32	37.78/33.20	40.76/28.93	44.84/30.00	36.36/23.79	44.98/30.94
Two 3D-RS18	42.28/30.55	42.77/32.72	44.31/36.31	32.84/30.41	39.28/28.41	48.41/32.56	36.36/24.68	49.34/31.62
Two RS18-LSTM	43.20/31.28	42.20/31.66	41.72/31.74	36.30/34.63	40.55/27.09	48.97/32.13	37.76/26.91	47.60/35.60
Two VGG13-LSTM	44.54/32.79	44.65/32.96	43.71/32.61	38.52/35.49	46.92/31.55	50.09/32.30	37.76/30.28	48.03/36.43
Average	39.58/29.34	39.27/29.80	38.11/30.24	33.90/31.77	38.98/26.75	44.43/29.60	33.99/25.45	42.25/31.97

Table 1. Comparison of four kinds of baseline architectures training on all scenes and then testing on all scenes, DL11k, and its sub-scenes.

Source	DL11k	Conflict	School	DailyLife	Argue	Medicine	Social
DL11k	38.52/27.72	36.53/28.90	34.32/29.80	37.58/23.58	44.84/27.71	30.77/25.55	37.12/27.02
WIS9k	28.00/20.74	22.36/17.55	20.25/21.39	27.60/18.43	32.46/20.42	29.37/21.13	39.30/27.49
SIA10k	28.05/21.61	20.56/18.30	24.69/24.66	32.27/22.14	32.27/22.40	29.37/19.57	36.24/24.55
AI9k	26.38/19.95	22.55/17.45	23.95/22.52	26.54/17.86	26.08/19.43	23.78/19.72	28.82/27.80

Table 2. Experiment results of intra-scenario performance consistency (highlighted in blue bar) and inter-scenario invariance via RS50-LSTM training on four scenarios and then testing on DL11k, and its sub-scenes.

Method	All	WIS9k	Action	ScholarReport	Speech	Liveshow	ElegantArt	Talkshow
RS18	39.33/30.30	40.50/28.67	50.61/34.11	40.25/27.59	43.09/31.19	37.72/26.82	33.33/25.53	38.57/25.47
RS50	30.57/22.47	32.52/23.50	37.80/27.77	35.31/24.30	30.89/23.22	28.51/23.13	31.35/21.51	28.86/20.14
VGG13	41.02/31.19	43.04/30.23	54.88/34.08	47.41/33.11	38.48/28.07	44.74/30.40	36.51/28.53	40.57/26.25
VGG16	41.66/32.01	42.93/30.77	53.05/33.57	43.21/29.00	38.75/29.22	46.05/33.65	40.87/32.16	39.43/24.96
RS18-LSTM	42.59/30.92	44.12/29.59	51.83/31.02	44.94/27.73	39.57/27.95	46.93/31.59	37.70/30.48	44.57/27.23
RS50-LSTM	40.75/32.12	41.74/30.70	56.10/36.76	43.70/30.88	39.57/29.34	40.35/30.40	38.89/32.57	40.00/27.42
VGG13-LSTM	43.37/32.41	44.23/30.81	53.05/33.84	45.19/30.82	40.92/30.13	45.61/31.28	41.27/31.19	44.29/29.17
VGG16-LSTM	41.70/30.93	41.63/28.42	46.95/27.49	45.19/30.54	42.82/30.86	36.84/25.76	39.68/30.48	41.14/26.39
C3D [9]	31.69/22.68	30.15/19.94	35.98/20.38	26.91/16.95	31.71/21.92	28.51/22.55	28.97/21.60	36.57/23.25
P3D [8]	33.39/23.20	34.95/22.40	39.63/24.87	36.30/22.10	36.31/23.94	34.65/23.13	27.38/19.42	31.43/18.32
I3D [3]	38.78/30.17	38.52/29.11	46.34/31.67	46.17/35.70	34.42/25.81	37.72/30.57	34.92/25.59	26.29/18.27
3D-RS18 [10]	37.57/26.67	38.40/24.85	48.78/28.40	38.52/23.48	39.30/27.40	39.04/25.12	32.54/24.44	36.29/21.86
Two C3D	41.77/30.72	43.44/29.77	54.27/35.23	43.70/30.48	41.73/30.17	41.23/25.74	42.06/28.16	42.00/27.89
Two I3D	41.30/31.01	42.31/30.14	57.93/36.47	46.67/31.42	40.92/30.57	38.16/25.91	37.30/28.55	39.43/28.37
Two 3D-RS18	42.28/30.55	44.12/29.63	54.88/30.61	46.91/29.62	42.28/30.17	41.67/28.50	36.11/27.20	38.57/24.66
Two RS18-LSTM	43.20/31.28	44.91/30.37	57.32/34.12	47.16/31.88	41.46/30.12	44.74/26.55	37.70/26.76	43.43/27.52
Two VGG13-LSTM	44.54/32.79	45.25/31.45	57.93/38.26	47.90/32.43	40.11/29.40	48.25/33.02	43.65/31.93	45.14/28.30
Average	39.58/29.34	40.61/28.11	50.23/31.51	42.68/28.69	38.91/28.15	39.79/27.65	36.15/27.23	38.39/24.75

Table 3. Comparison of four kinds of baseline architectures training on all scenes and then testing on all scenes, WIS9k, and its sub-scenes.

Source	WIS9k	Action	ScholarReport	Speech	Liveshow	ElegantArt	Talkshow
DL11k	30.71/20.24	46.34/31.23	26.42/16.74	31.98/23.59	28.95/20.16	28.57/21.66	30.00/19.10
WIS9k	40.72/26.88	51.22/29.88	42.96/28.01	39.02/25.83	39.47/23.02	36.11/24.98	38.86/24.77
SIA10k	30.94/19.79	37.20/20.95	27.90/16.87	34.42/24.20	26.32/16.58	29.37/20.65	34.29/20.87
AI9k	23.25/17.25	20.73/18.17	20.00/14.75	27.37/21.01	18.42/12.36	25.79/18.49	22.57/16.90

Table 4. Experiment results of intra-scenario performance consistency (highlighted in blue bar) and inter-scenario invariance via RS50-LSTM training on four scenarios and testing on WIS9k, and its sub-scenes.

Method	All	SIA10k	Business	Experiment	OfficialEvent	Crime	Interview	Contest
RS18	39.33/30.30	42.31/30.02	38.83/28.01	49.56/26.70	37.98/25.48	36.34/28.63	45.75/29.18	48.24/33.37
RS50	30.57/22.47	30.56/22.68	22.07/17.53	31.86/16.89	34.49/20.96	23.08/17.26	33.25/21.72	37.06/27.55
VGG13	41.02/31.19	43.44/29.99	39.39/27.83	49.56/26.52	38.33/26.04	36.87/27.95	44.34/28.83	47.62/32.39
VGG16	41.66/32.01	42.31/29.58	37.43/26.94	52.21/33.12	41.46/26.09	34.75/28.20	47.17/30.77	48.03/32.92
R18-LSTM	42.59/30.92	42.85/28.78	41.34/29.45	48.67/25.42	37.98/22.40	40.32/31.14	45.52/28.01	50.10/33.79
R50-LSTM	40.75/32.12	42.16/30.39	37.99/28.47	47.79/33.17	39.37/26.84	36.07/29.63	43.87/30.02	48.24/34.32
VGG13-LSTM	43.37/32.41	45.00/31.45	43.02/31.22	57.52/36.17	37.63/23.05	40.05/31.33	47.17/30.07	49.90/33.66
VGG16-LSTM	41.70/30.93	43.83/29.83	40.22/29.55	53.10/30.03	40.77/24.54	37.40/29.00	46.23/27.39	48.65/34.15
C3D [9]	31.69/22.68	42.70/29.22	40.78/28.44	54.87/22.87	35.89/21.74	36.07/25.16	43.16/26.35	46.58/32.44
P3D [8]	33.39/23.20	36.73/23.66	33.24/21.28	42.48/21.98	32.75/19.02	32.10/21.69	39.62/21.94	40.58/26.69
I3D [3]	38.78/30.17	40.55/31.07	33.52/26.20	53.10/31.56	38.68/28.47	38.46/29.99	41.51/27.87	45.55/35.43
3D-R18 [10]	37.57/26.67	40.40/26.08	35.75/23.46	54.87/32.50	37.63/21.68	33.69/23.53	42.69/22.70	44.10/28.32
Two C3D	41.77/30.72	44.71/30.15	46.09/35.27	63.72/37.55	35.89/22.51	38.99/27.51	46.23/28.45	48.03/32.31
Two I3D	41.30/31.01	43.63/31.20	38.83/30.19	54.87/26.96	39.72/25.10	40.05/27.89	44.81/29.96	48.03/33.28
Two 3D-R18	42.28/30.55	42.95/27.83	38.83/26.57	62.83/33.41	33.80/19.25	38.99/26.28	45.52/24.71	48.45/33.16
Two RS18-LSTM	43.20/31.28	46.33/31.09	44.69/31.57	57.52/24.56	36.93/21.72	38.20/28.59	47.41/28.50	53.00/33.93
Two VGG13-LSTM	44.54/32.79	46.57/31.88	44.69/32.33	53.98/31.66	43.90/26.60	38.46/30.10	46.70/28.35	52.80/35.32
Average	39.58/29.34	42.04/28.94	38.25/27.69	52.06/28.57	37.41/23.30	36.28/27.15	44.19/27.22	47.33/32.39

Table 5. Comparison of four kinds of baseline architectures training on all scenes and then testing on all scenes, SIA10k, and its sub-scenes.

Source	SIA10k	Business	Experiment	OfficialEvent	Crime	Interview	Contest
DL11k	30.85/21.86	23.46/18.93	38.05/26.28	31.01/19.72	26.53/21.73	33.73/20.12	34.78/24.44
WIS9k	32.32/20.59	30.17/21.69	38.94/21.20	29.27/17.46	28.91/21.35	33.73/17.19	33.75/21.78
SIA10k	39.96/25.22	36.87/23.59	63.72/42.51	36.24/21.66	33.95/22.52	41.27/22.95	44.93/28.35
AI9k	24.53/18.79	16.76/13.15	31.86/28.26	23.00/15.02	25.20/20.87	25.94/17.42	27.33/19.83

Table 6. Experiment results of intra-scenario performance consistency (highlighted in blue bar) and inter-scenario invariance via RS50-LSTM training on four scenarios and testing on SIA10k, and its sub-scenes.

Method	All	AI9k	History	Terror	War	Crisis
RS18	39.33/30.30	33.90/27.20	31.17/23.72	31.28/26.69	35.09/28.30	36.88/29.21
RS50	30.57/22.47	30.14/19.94	35.44/21.05	26.54/19.67	30.83/19.19	24.25/20.11
VGG13	41.02/31.19	38.86/29.94	36.92/25.75	36.73/31.48	37.73/29.22	42.52/31.65
VGG16	41.66/32.01	39.60/31.46	37.11/26.81	39.57/34.21	41.38/32.80	40.86/32.95
R18-LSTM	42.59/30.92	39.66/30.40	40.45/31.06	35.55/29.96	39.35/28.55	44.85/33.46
R50-LSTM	40.75/32.12	38.01/31.16	40.07/30.81	36.26/32.46	39.15/31.26	39.87/31.87
VGG13-LSTM	43.37/32.41	41.20/31.49	40.26/28.06	40.28/33.61	40.16/30.24	42.86/31.11
VGG16-LSTM	41.70/30.93	37.04/29.39	38.03/28.82	36.26/32.83	36.92/27.63	41.53/33.59
C3D [9]	31.69/22.68	27.29/19.80	27.83/19.80	22.99/20.31	24.75/17.55	32.56/20.93
P3D [8]	33.39/23.20	31.34/21.52	32.65/21.23	27.96/22.71	31.03/21.10	34.55/21.61
I3D [3]	38.78/30.17	37.44/28.15	40.07/29.02	33.89/29.10	37.53/26.32	36.54/28.81
3D-R18 [10]	37.57/26.67	33.45/25.40	31.73/22.88	31.28/27.83	37.53/27.07	37.21/27.25
Two C3D	41.77/30.72	37.89/28.09	41.93/27.16	35.78/30.47	37.12/26.71	40.86/29.60
Two I3D	41.30/31.01	38.75/28.53	38.78/26.49	36.02/29.19	38.95/29.20	38.87/28.01
Two 3D-R18	42.28/30.55	38.46/28.54	40.45/28.11	35.07/28.73	36.71/26.55	42.19/29.58
Two RS18-LSTM	43.20/31.28	40.40/30.04	41.93/29.94	36.49/29.94	41.38/29.91	43.85/31.45
Two VGG13-LSTM	44.54/32.79	40.63/30.96	41.74/30.03	37.44/32.49	39.35/28.70	46.84/35.11
Average	39.58/29.34	36.55/27.61	37.41/26.45	33.75/28.70	36.53/26.93	39.12/29.12

Table 7. Comparison of four kinds of baseline architectures training on all scenes and then testing on all scenes, AI9k, and its sub-scenes.

Source	AI9k	History	Terror	War	Crisis
DL11k	25.64/19.58	27.09/21.19	23.70/20.43	24.34/16.88	24.58/18.86
WIS9k	23.65/18.43	25.23/20.01	21.80/20.42	24.34/17.10	28.24/19.57
SIA10k	27.92/20.30	29.68/21.05	24.64/19.85	27.38/20.03	30.23/19.32
AI9k	31.68/24.04	31.73/21.07	28.67/24.87	34.48/24.25	31.56/24.63

Table 8. Experiment results of intra-scenario performance consistency (highlighted in blue bar) and inter-scenario invariance via RS50-LSTM training on four scenarios and testing on AI9k, and its sub-scenes.

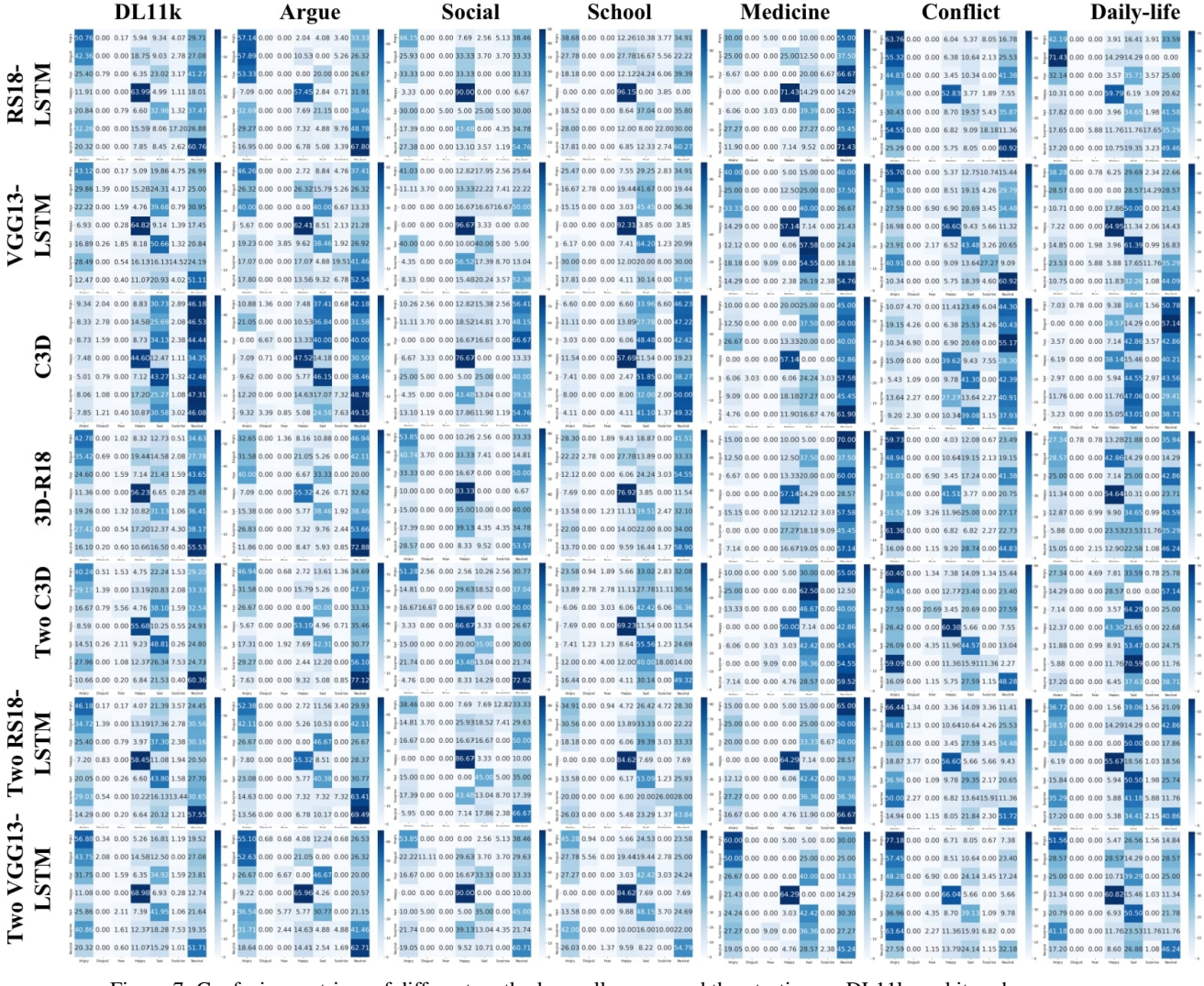

Figure 7. Confusion matrices of different methods on all scenes and then testing on DL11k, and its sub-scenes.

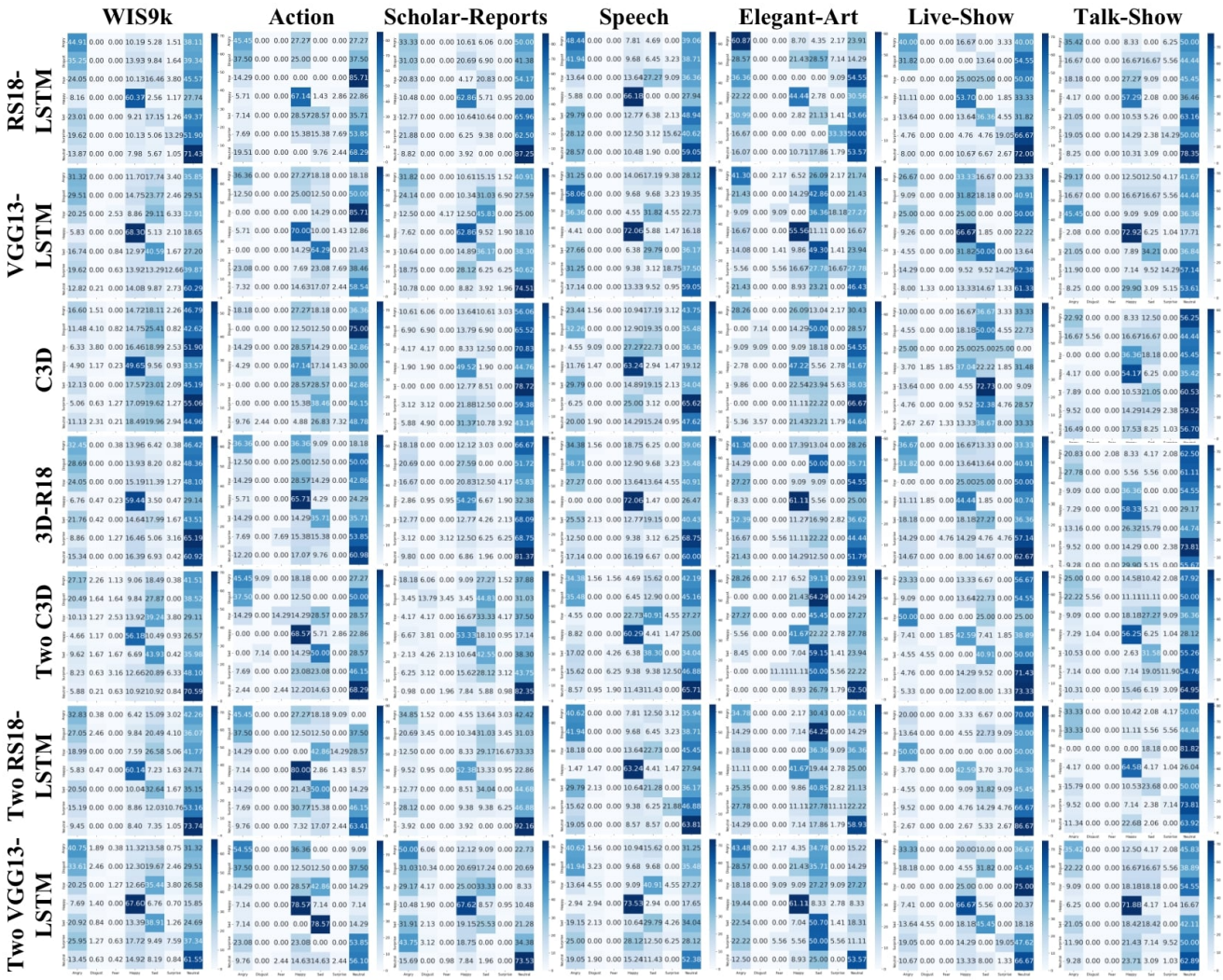

Figure 8. Confusion matrices of different methods on all scenes and then testing on WIS9k, and its sub-scenes.

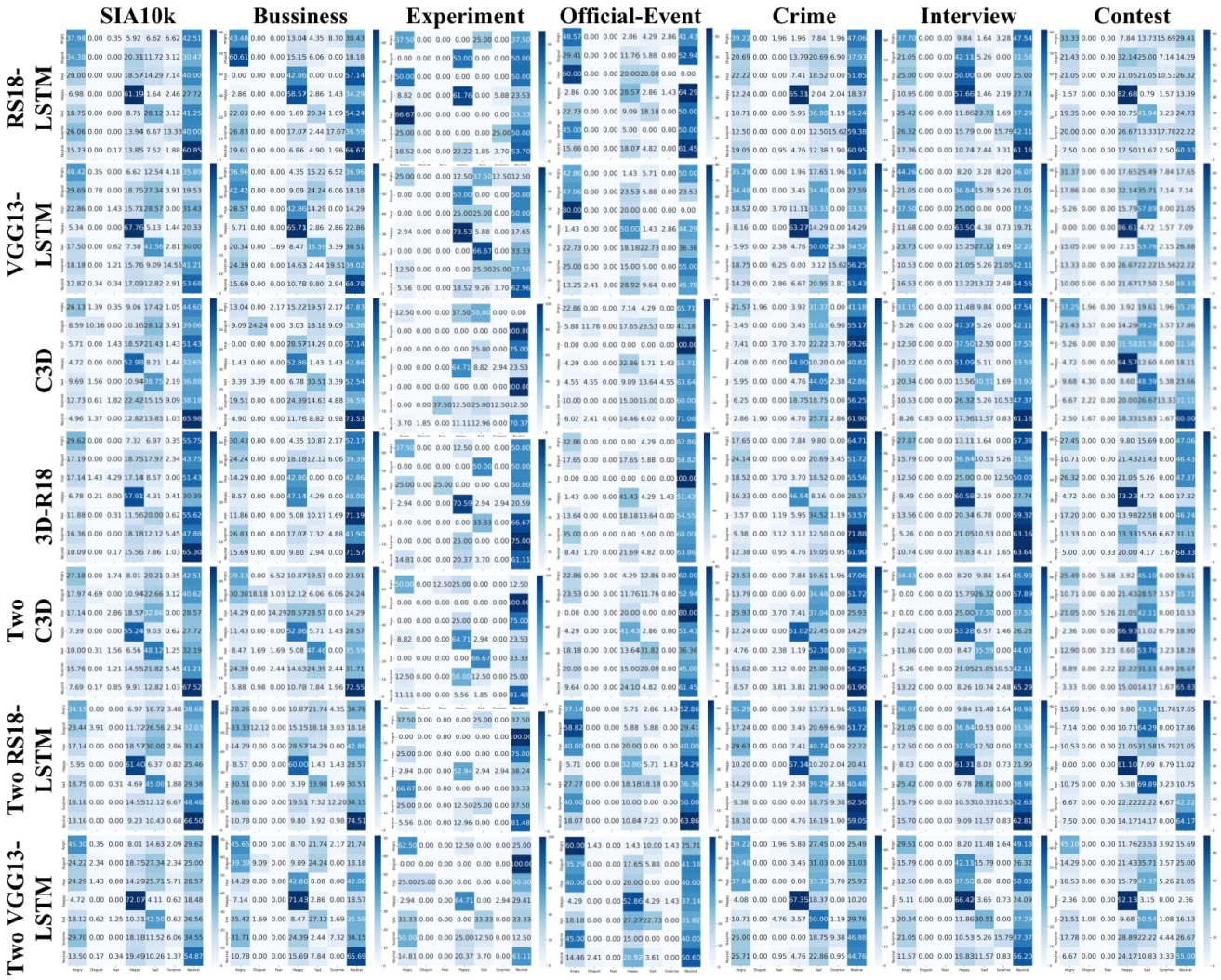

Figure 9. Confusion matrices of different methods training on all scenes and then testing on SIA10k, and its sub-scenes.

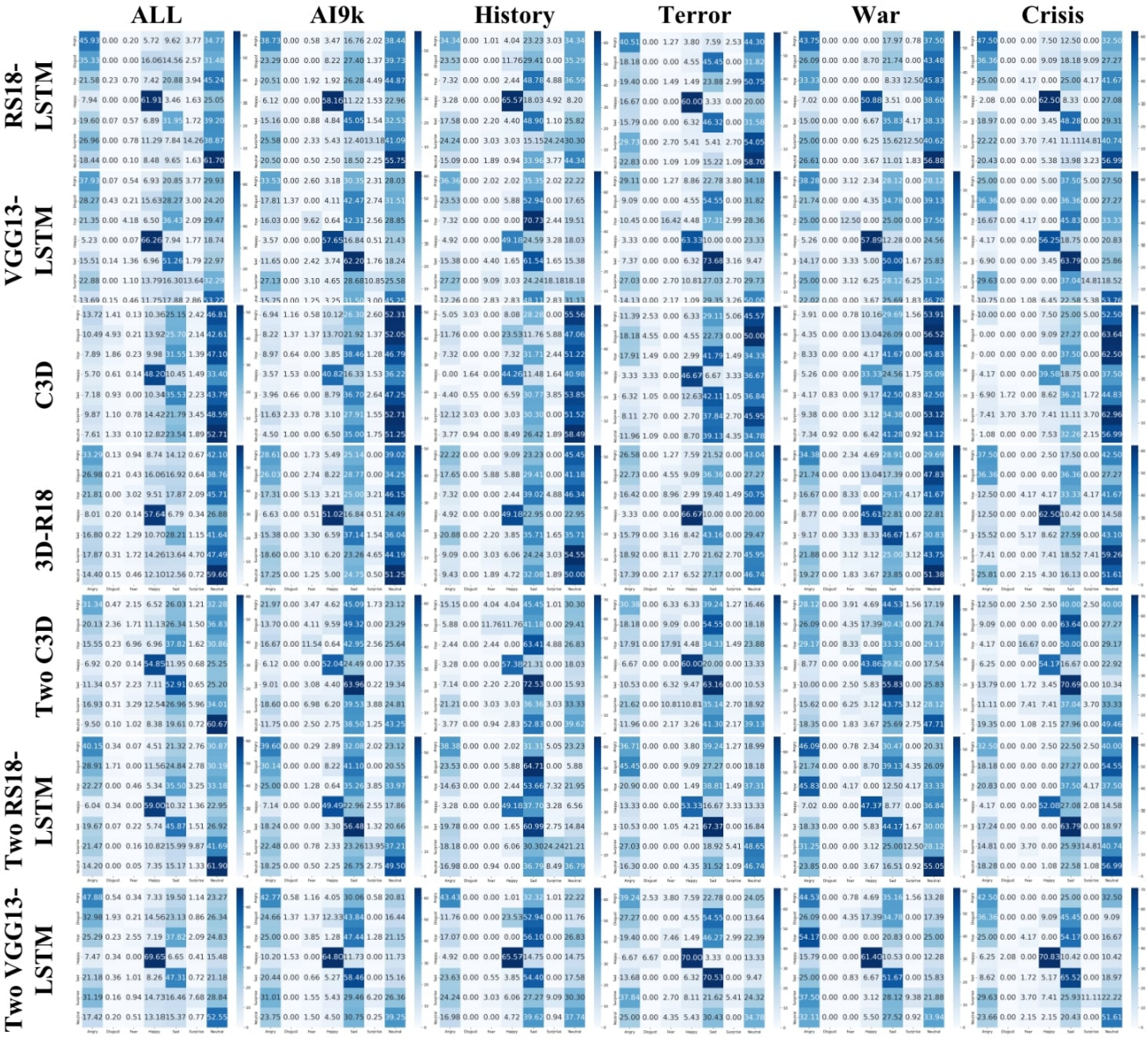

Figure 10. Confusion matrices of different methods training on all scenes and then testing on all scenes, AI9k, and its sub-scenes.