Supplementary File for:
"Multi-Modal Multi-Action Video Recognition"

Zhensheng Shi[1,2] Ju Liang[2] Qianqian Li[2] Haiyong Zheng[2,*] Zhaorui Gu[2] Junyu Dong[1,3] Bing Zheng[2,4]
[1]Frontiers Science Center for Deep Ocean Multispheres and Earth System, Ocean University of China
[2]Underwater Vision Lab (http://ouc.ai), Ocean University of China
[3]College of Computer Science and Technology, Ocean University of China
[4]Sanya Oceanographic Institution, Ocean University of China

A. Training and Evaluation

Training Details. We implement data augmentation on both temporal and spatial scopes. We randomly sample 8 consecutive frames with sampling step 2. The input frames are cropped via multi-scale random cropping and then resized to 112×112. The cropping window size is $d \times d$, where d is the multiplication of input shorter side length and scale factor in $[0.7, 0.875]$. We train and evaluate our models on 8 NVIDIA RTX 2080Ti GPUs, and set mini-batch size to 8 per GPU (64 in total) with Batch Normalization in training. For Mini M-MiT, the training procedure totally takes 30 epochs, with an initial learning rate 0.05 and reduces by a factor 0.1 at 12 and 24 epochs, and also the first 3 epochs are used for warm-up [1]; for full M-MiT, the initial learning rate is set to 0.01 without warm-up. The network is trained with commonly used binary cross-entropy loss optimized by SGD with momentum 0.9 and weight decay 0.0001. We empirically set t to 0.4 for adjacency matrix binarization. All experiments are implemented by PyTorch 1.3 and we also use mixed precision training.

Evaluation Metrics. We report mAP (mean Average Precision), top-1, and top-5 classification accuracy for all experiments, among which mAP is regarded as the main evaluation metric since it captures errors in the ranking of relevant actions for a video. For each positive label, mAP computes the proportion of relevant labels that are ranked before it and then averages over all labels. Top-1 and top-5 accuracy indicate the percentage of testing videos where the top predicted class and any of the top predicted 5 classes is positive for the video, respectively. We perform multiple clips testing for evaluation at test time, temporal clips are uniformly sampled from each video, and spatial crops are then sampled from each frame of these clips. We uniformly sample 10 temporal clips from full length of the video, and use 3 spatial crops (two sides and one center). We also perform spatial fully-convolutional inference [2] by scaling shorter side of each video frame to 128 while maintaining aspect ratio. The final prediction is max score (for mAP) and average score (for top-1 and top-5) over all clips.

*Corresponding author: Haiyong Zheng (zhenghaiyong@ouc.edu.cn).
This work was supported by the National Natural Science Foundation of China under Grant Nos. 61771440, 41927805, and 41776113.

B. Boosts Analysis

Figure A shows class-wise boosts over visual GCN with our multi-modal multi-action GCNs listed in Table 1 of the paper. We denote the boost from one model to another as the mAP difference divided by its mAP, representing the growth rate of model mAP. It can be seen that, **(a)** $\mathcal{J}(\mathbf{H}, \mathbf{G}_\nu)$ brings a little boost against $\mathcal{J}(\mathbf{H}, \mathbf{FC})$, and the performance gain is mainly in categories with visual multi-action relations, such as *child+speaking* with *frowning* and *crying*; **(b)** $\mathcal{J}(\mathbf{H}, \mathbf{G}_\alpha)$ boosts the performance significantly over $\mathcal{J}(\mathbf{H}, \mathbf{G}_\nu)$ in categories with audio multi-action relations, *e.g.*, co-occurred *rocking* and *shaking* can be connected by audio; **(c)** $\mathcal{J}(\mathbf{H}, \mathbf{G}_\tau)$ also contributes a lot to recognize multiple actions with related literal meaning like *opening* and *closing* as well as *locking*; **(d)** while $\mathcal{J}(\mathbf{H}, \mathbf{G}_\alpha, \mathbf{G}_\tau)$ combines the strengths of both audio and textual multi-action relations to bring a significant gain; **(e)** further $\mathcal{J}(\mathbf{H}, \mathbf{G}_\nu, \mathbf{G}_\alpha, \mathbf{G}_\tau)$ boosts performance by integrating advantages of all three modality-specific multi-action relations, yielding highest mAP (refer to Table 1 in paper).

References

[1] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[2] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018.

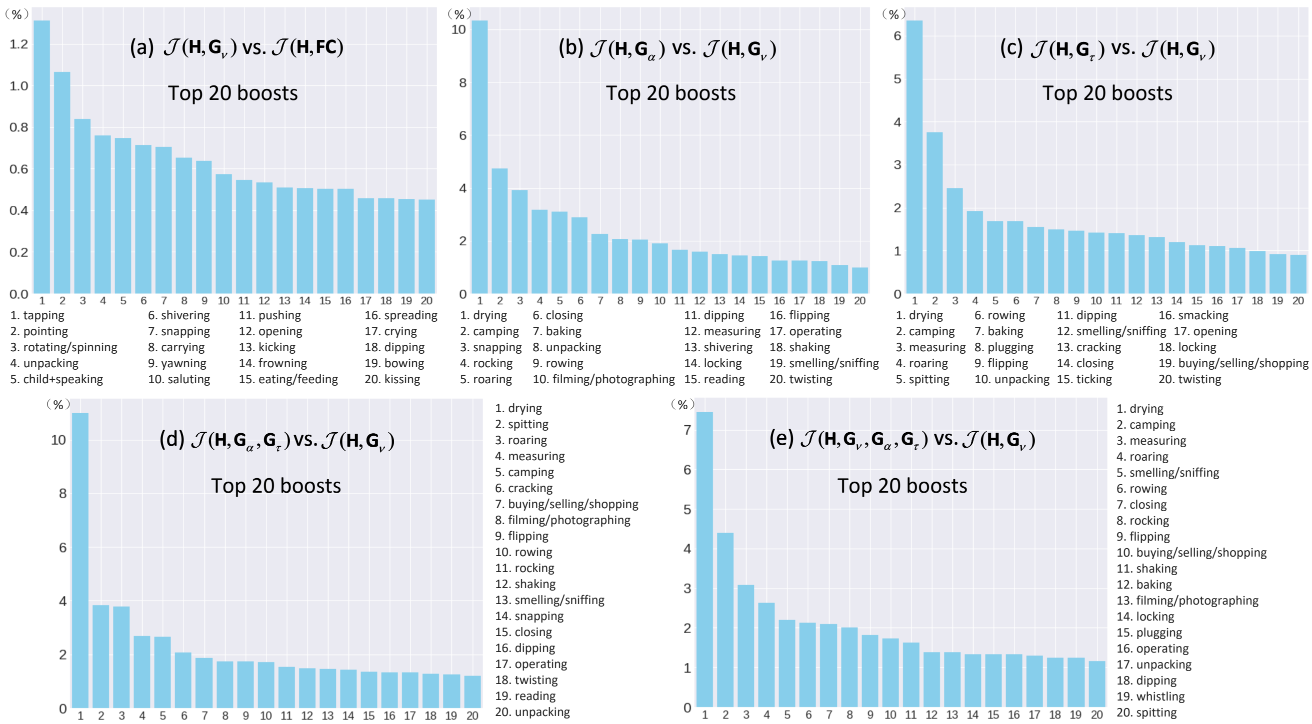

Figure A: Class-wise boosts of our multi-modal multi-action GCNs versus visual GCN. Refer to Section B for details.