Supplementary for “Decoupling and Recoupling Spatiotemporal Representation for RGB-D-based Motion Recognition”

1. Method

1.1. Supplementary for FRP Module

Normalization for Visual Guidance Map. To further improve the numerical stability, we do the normalization for generated visual guidance maps G_m^l (Eq.5 in main manuscript) as:

$$G_{\mathrm{m,norm}}^l = \frac{G_m^l - G_{\mathrm{m,min}}^l}{G_{\mathrm{m,max}}^l - G_{\mathrm{m,min}}^l} \tag{1}$$

where $G_{\mathrm{m,min}}^l$ and $G_{\mathrm{m,max}}^l$ represent the maximum and minimum values in G_m^l, respectively; And $G_{\mathrm{m,norm}}^l$ represents the normalized visual guidance map.

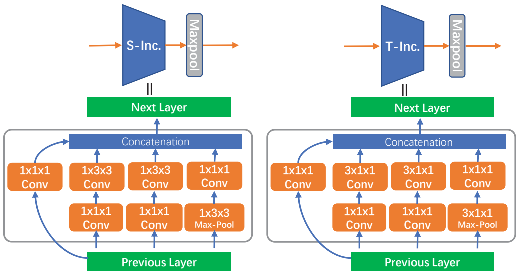
Visual Guidance Map Alignment. To align the generated visual guidance maps with the input sequence, we shift it backwards along the time dimension by $m - n$ units, and the guidance map of the previous $m-n$ frames is filled with the zeros matrix. Therefore, the final visual guidance map can be formulated as:

$$\begin{aligned} &\hat{G}_{\mathrm{norm}}^l = [G_{1,\mathrm{norm}}^l, \dots, G_{T,\mathrm{norm}}^l], \forall l = 1, 2, \dots, M \\ &s.t. \quad G_{\mathrm{t,norm}}^l = \begin{cases} G_{\mathrm{t-(m-n),norm}}^l & t > m - n \\ 0 & otherwise \end{cases} \end{aligned} \tag{2}$$

where $\hat{G}_{\mathrm{norm}}^l$ represents the aligned visual guidance map with the input sequence. It then integrates with spatial feature stream captured by the spatial multi-scale features learning module (SMS) and serves as the input to next layer of the network.

1.2. Structure of the SMS and TMS Modules

As shown in Figure 1, the spatial and temporal multi-scale features learning module SMS and TMS are based on the inception structure. And a Max Pooling operation is embedded behind them to aggregate features with high correlation to reduce information redundancy.



(a) SMS Module (b) TMS Module

Figure 1. The structure of the spatial multi-scale features learning module SMS and the temporal multi-scale features learning module TMS.

1.3. Loss Function

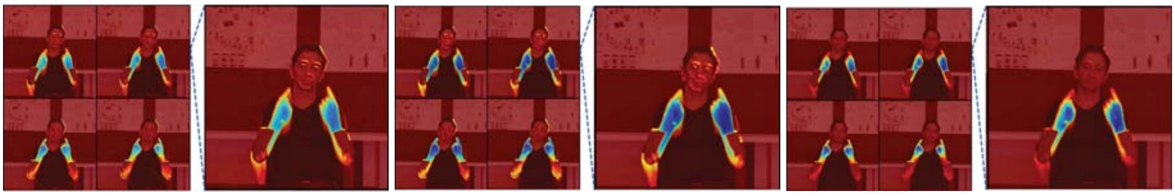
For training the unimodal network, inspired by [19], we configurate three sub-branches in the decoupled temporal representation learning network DTN, and each sub-branch imposes a constraint loss with weight coefficient of γ. In addition, we also introduce two additional constraint losses with weight coefficients of $1 - \gamma$ and 1.0, to constrain the summation of three sub-branches and output of the RCM module. So the overall loss for unimodal network training is the sum of all of those losses, and can be denoted as:

$$\begin{aligned} \mathcal{L}_{\mathrm{uni}}^{overall} = &\gamma \mathcal{L}_C^{\mathcal{S}_1} + \gamma \mathcal{L}_C^{\mathcal{S}_2} + \gamma \mathcal{L}_C^{\mathcal{S}_3} + \\ &(1-\gamma)\mathcal{L}_C^{\mathcal{S}_{all}} + \mathcal{L}_D \end{aligned} \tag{3}$$

where $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$ represent the output of the three sub-branch respectively; $\mathcal{S}_{all} = \mathcal{S}_1 + \mathcal{S}_2 + \mathcal{S}_3$; and $\mathcal{L}_C$ and $\mathcal{L}_D$ represent classification loss and distillation loss, respectively. For training the multi-modal network, we introduce a multi-loss collaborative optimization strategy, which can be denoted as:

$$\begin{aligned} \mathcal{L}_{\mathrm{multi}}^{overall} = &\mathcal{L}_C^{\mathcal{S}_R} + \mathcal{L}_C^{\mathcal{S}_D} + \mathcal{L}_B^{\mathcal{S}_R} + \mathcal{L}_B^{\mathcal{S}_D} + \\ &\mathcal{L}_M^{\mathcal{S}_R} + \mathcal{L}_M^{\mathcal{S}_D} + \mathcal{L}_D^{\mathcal{S}_R} + \mathcal{L}_D^{\mathcal{S}_D} \end{aligned} \tag{4}$$

where $\mathcal{S}_R$ and $\mathcal{S}_D$ represent the output of the color and depth network branches, respectively; and $\mathcal{L}_B$ and $\mathcal{L}_M$ represent binary cross entropy loss and mean square error loss. It is note that we assign a weight coefficient of 1.0 to all losses.



(a) Dynamic Guidance Map.

(b) Static Guidance Map.

(c) Visual Guidance Map.

Figure 2. Visualization of the generated visual guidance map. (a) The dynamic guidance map defined with D_m in the main manuscript. (b) The static guidance map defined with S_m in the main manuscript. (c) The visual guidance map defined with G_m in the main manuscript. Note that the deeper the color, the greater the weight.

Size	2	4	6	8	10	12
NvGesture	87.5	87.7	88.1	88.8	**89.6**	88.2
THU-READ	80.4	80.8	80.8	81.2	**81.7**	80.6

Table 1. The effect of the sliding window size.

2. Ablation Study

2.1. Impact of Sliding Window Size

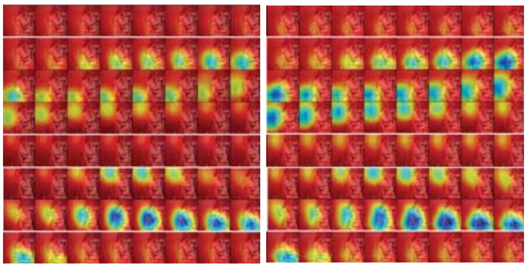
In Table 1, we set different sliding window sizes in the FRP module to study how it affects network performance. We observe that the performance gradually improves as we increase the size of the window. However, when the size reaches 12, the performance of the network degrades instead. We conjecture that this may be because the response range in the dynamic guidance map has increased, and as a result, the value of some noise regions has also been amplified simultaneously.

2.2. Study for the Robustness of Illumination

As shown in Figure 2 (a), the dynamic guidance map D_m is inevitably influenced by illumination as it is driven by dynamic images. To address this issue, we introduce the static guidance map S_m, as shown in Figure 2 (b), it can not only enhance the response value of important areas in the image, but also significant alleviate the effects of lighting. After combining the dynamic guidance map and static guidance map, the final visual guidance map, as shown in Figure 2 (c), can effectively highlight the important areas in the image.

2.3. Impact of Local and Global Modeling in DTN

Temporal features learning based on global contextual information is vital for sequence. However, we find that solely utilizing the Transformer network for global contextual information modeling in the sequence is hard to generate effective motion descriptors, especially hard to capture the local subtle movement information as shown in Figure 3 (a). To alleviate this drawback, we introduce an inception-based temporal multi-scale features learning network (TMS) for local fine-grained temporal representation learning. It first captures local hierarchical temporal features, and then aggregates neighboring features with high correlation. After that, we feed them into stack of Transformer blocks to progressively learn the global temporal representation. As shown in Figure 3 (b), after modeling temporal information at a local fine-grained level and global coarse-grained level, the local and global motion perception abilities of the network have been significantly enhanced.



(a) Global Modeling.

(b) Local and Global Modeling.

Figure 3. Visualization of the Class Activation Map (CAM). (a) The activation response of global coarse-grained temporal information modeling. (b) The activation response of the joint modeling with local fine-grained as well as global coarse-grained temporal information.

2.4. Study for Feature Enhancement Attention

Figure 4 visualizes the attention map A_{XY} (Eq.15 in main manuscript) generated by the spatiotemporal recoupling module (RCM), which shows that it can selectively activate several important neuron from X and Y directions in captured spatial features. In addition, we can obviously find that attention map A_{XY} mainly guides the network to focus on the intermediate frame, which just shows that these intermediate frames contain most of the important information of a sequence.

2.5. Frame Rate Study for Sub-branch

In this ablation, we configure different frame rates for each sub-branch to understand its impact on DTN. We only

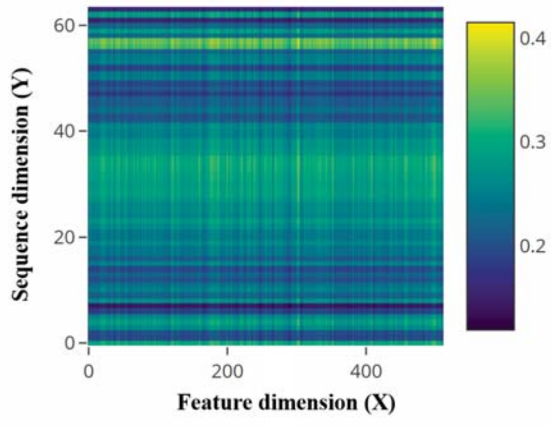


Figure 4. Visualization of the attention map for spatial feature enhancement generated by RCM module.

fine-tune the DSN sub-network and compare models trained for 100 epochs. As shown in Table 2, the experiment result confirms that (1) configuring different frame rates for each sub-branch can boost the performance, which demonstrates that motion recognition benefits from multi-scale temporal features. And (2) setting a smaller or larger frame rate for DTN results in a decrease in performance, we conjecture that the former may be caused by the loss of important information, and the latter may be caused by temporal information redundancy.

Small Transf.	Medium Transf.	Large Transf.	Nv	THU
16	16	16	79.88	77.08
8	16	24	80.63	77.92
16	32	48	**81.46**	**78.75**
16	48	80	81.30	77.92

Table 2. The impact of different frame rates for each sub-branch in DTN. “Transf.” means Transformer network.

2.6. More Comparisons

In this section, we compare with other methods not listed in the main manuscript. Table 3 lists some other methods on the gesture datasets namely NvGesture and Chalearn IsoGD. Table 4 lists some other methods on the action datasets namely THU-READ and NTU-RGBD.

3. Limitations

The main limitations of the proposed method can be summarized as follows: First, we only explored our method on RGB-D modalities, while other modalities, such as optical flow and infrared, remain to be further validated. Second, due to the relatively heavy computation of the model, the current version may not be suitable for mobile deployment. Therefore, making the model lightweight is the direction of our future efforts.

Method	Modality	Accuracy(%)
NvGesture Dataset		
GPM [1]	RGB	75.90
PreRNN [18]	RGB	76.50
ResNeXt-101 [4]	RGB	78.63
Ours	RGB	**89.58**
ResNeXt-101 [4]	Depth	83.82
PreRNN [18]	Depth	84.40
GPM [1]	Depth	85.50
Ours	Depth	**90.62**
PreRNN [18]	RGB+Depth	85.00
GPM [1]	RGB+Depth	86.10
Ours(Multiplication)	RGB+Depth	90.89
Ours(Addition)	RGB+Depth	91.10
Ours(CAPF)	RGB+Depth	**91.70**
Chalearn IsoGD Dataset		
c-ConvNet [15]	RGB	36.60
C3D-gesture [8]	RGB	37.28
AHL [2]	RGB	44.88
ResC3D [9]	RGB	45.07
3DCNN+LSTM [21]	RGB	51.31
attention+LSTM [20]	RGB	55.98
Ours	RGB	**60.87**
c-ConvNet [15]	Depth	40.08
C3D-gesture [8]	Depth	40.49
ResC3D [9]	Depth	48.44
AHL [2]	Depth	48.96
3DCNN+LSTM [21]	Depth	49.81
attention+LSTM [20]	Depth	53.28
Ours	Depth	**60.17**
c-ConvNet [15]	RGB+Depth	44.80
AHL [2]	RGB+Depth	54.14
3DCNN+LSTM [21]	RGB+Depth	55.29
Ours(Multiplication)	RGB+Depth	66.71
Ours(Addition)	RGB+Depth	66.68
Ours(CAPF)	RGB+Depth	**66.79**

Table 3. Comparison with other methods on gesture datasets.

References

[1] Vikram Gupta, Sai Kumar Dwivedi, Rishabh Dabral, and Arjun Jain. Progression modelling for online and early gesture

THU-READ Dataset		
Method	Modality	Accuracy(%)
Appearance Stream [12]	RGB	41.90
TSN [14]	RGB	73.85
Ours	RGB	**81.25**
Depth Stream [12]	Depth	34.06
TSN [14]	Depth	65.00
Ours	Depth	**77.92**
MDNN [13]	RGB+Flow+Depth	62.92
TSN [14]	RGB+Flow	78.23
TSN [14]	RGB+Flow+Depth	81.67
Ours(Multiplication)	RGB+Depth	86.10
Ours(Addition)	RGB+Depth	86.25
Ours(CAPF)	RGB+Depth	**87.04**

NTU-RGBD Dataset			
Method	Modality	CS(%)	CV(%)
CNN+Motion+Trans [6]	Skeleton	83.2	88.8
ST-GCN [17]	Skeleton	81.5	88.3
Motif+VTDB [16]	Skeleton	84.2	90.2
STGR-GCN [5]	Skeleton	86.9	92.3
AS-GCN [7]	Skeleton	86.8	94.2
Adaptive GCN [10]	Skeleton	88.5	95.1
AGC-LSTM [11]	Skeleton	89.2	95.0
MMTM [3]	RGB+Pose	91.9	-
Ours	RGB	90.3	95.4
Ours	Depth	92.7	96.2
Ours(Multiplication)	RGB+Depth	93.6	96.6
Ours(Addition)	RGB+Depth	93.9	96.7
Ours(CAPF)	RGB+Depth	**94.2**	**97.3**

Table 4. Comparison with other methods on action datasets.

detection. In *2019 International Conference on 3D Vision (3DV)*, pages 289–297, 2019. 3

[2] Ting-Kuei Hu, Yen-Yu Lin, and Pi-Cheng Hsiu. Learning adaptive hidden layers for mobile gesture recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 3

[3] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L. Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4

[4] Okan Köpüklü, Ahmet Gunduz, Neslihan Kose, and Gerhard Rigoll. Real-time hand gesture detection and classification using convolutional neural networks. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019. 3

[5] Bin Li, Xi Li, Zhongfei Zhang, and Fei Wu. Spatio-temporal graph routing for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8561–8568, 2019. 4

[6] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Skeleton-based action recognition with convolutional neural networks. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 597–600. IEEE, 2017. 4

[7] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3595–3603, 2019. 4

[8] Yunan Li, Qiguang Miao, Kuan Tian, Yingying Fan, Xin Xu, Rui Li, and Jianfeng Song. Large-scale gesture recognition with a fusion of rgb-d data based on saliency theory and c3d model. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2956–2964, 2017. 3

[9] Qiguang Miao, Yunan Li, Wanli Ouyang, Zhenxin Ma, Xin Xu, Weikang Shi, and Xiaochun Cao. Multimodal gesture recognition based on the resc3d network. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3047–3055, 2017. 3

[10] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019. 4

[11] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1227–1236, 2019. 4

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[13] Yansong Tang, Zian Wang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Multi-stream deep neural networks for rgb-d egocentric action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(10):3001–3015, 2018. 4

[14] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 4

[15] Pichao Wang, Wanqing Li, Jun Wan, Philip Ogunbona, and Xinwang Liu. Cooperative training of deep aggregation networks for rgb-d action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 3

[16] Yu-Hui Wen, Lin Gao, Hongbo Fu, Fang-Lue Zhang, and Shihong Xia. Graph cnns with motif and variable temporal block for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8989–8996, 2019. 4

[17] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 4

[18] Xiaodong Yang, Pavlo Molchanov, and Jan Kautz. Making convolutional networks recurrent for visual sequence learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[19] Zitong Yu, Benjia Zhou, Jun Wan, Pichao Wang, Haoyu Chen, Xin Liu, Stan Z Li, and Guoying Zhao. Searching multi-rate and multi-modal temporal enhanced networks for gesture recognition. *IEEE Transactions on Image Processing*, 2021. 1

[20] Liang Zhang, Guangming Zhu, Lin Mei, Peiyi Shen, Syed Afaq Ali Shah, and Mohammed Bennamoun. Attention in convolutional lstm for gesture recognition. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1957–1966, 2018. 3

[21] Liang Zhang, Guangming Zhu, Peiyi Shen, Juan Song, Syed Afaq Shah, and Mohammed Bennamoun. Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3120–3128, 2017. 3