Learning Robust Representations with Information Bottleneck and Memory Network for RGB-D-based Gesture Recognition Supplementary Material

1. Detailed Derivation

In this section, we present the entire derivation from the objective in Eq.(4) to the final IB loss function in Eq.(6) in the main manuscript.

As mentioned in the main manuscript, the entire objective is:

$$\min[I(v;y) - I(z;y) + I(v;z) - I(z;y)]. \tag{1}$$

According to the definition of mutual information, $I(v;y)$ can be rewritten as:

$$I(v;y) = H(y) - H(y|v), \tag{2}$$

where $H(y)$ is the information entropy of y and $H(y|v)$ is the conditional entropy of y given v. Similarly, we can have:

$$I(z;y) = H(y) - H(y|z), \tag{3}$$

$$I(v;z) = H(z) - H(z|v), \tag{4}$$

$$I(z;y) = I(y;z) = H(z) - H(z|y). \tag{5}$$

Substituting Eq.(2)-(5) to Eq.(1), the objective can be expressed in the form of conditional entropy as:

$$min[H(y|z) - H(y|v) + H(z|y) - H(z|v)]. \tag{6}$$

According to the definition of conditional entropy, $H(a|b)$ means the uncertainty of a with b known. Note that z is derived from v. In other words, z contains no information that v does not have. Therefore, $H(z|v)$ should be equal to zero. Thus we have Eq.(6) as:

$$min[H(y|z) - H(y|v) + H(z|y)]. \tag{7}$$

According to the definition of information entropy,

$$\begin{aligned} H(y|z) &= -\sum_{z\in Z} p(z) \sum_{y\in Y} p(y|z) \log p(y|z) \\ &= -\sum_{z\in Z} p(z) \sum_{y\in Y} p(y|z) \log \left[\frac{p(y|z)}{p(y|v)} p(y|v)\right], \end{aligned} \tag{8}$$

By factorizing it, we have:

$$\begin{aligned} &-\sum_{z\in Z} p(z) \sum_{y\in Y} p(y|z) \log \left[\frac{p(y|z)}{p(y|v)} p(y|v)\right] = \\ &\underbrace{-\sum_{z\in Z} \sum_{y\in Y} p(z)p(y|z) \log \left[\frac{p(y|z)}{p(y|v)}\right]}_{Z_1} \\ &\underbrace{-\sum_{z\in Z} \sum_{y\in Y} p(z)p(y|z) \log p(y|v)}_{Z_2}. \end{aligned} \tag{9}$$

Similarly, for $H(y|v)$ we have:

$$\begin{aligned} &-\sum_{v\in V} p(v) \sum_{y\in Y} p(y|v) \log \left[\frac{p(y|v)}{p(y|z)} p(y|z)\right] = \\ &\underbrace{-\sum_{v\in V} \sum_{y\in Y} p(v)p(y|v) \log \left[\frac{p(y|v)}{p(y|z)}\right]}_{V_1} \\ &\underbrace{-\sum_{v\in V} \sum_{y\in Y} p(v)p(y|v) \log p(y|z)}_{V_2}. \end{aligned} \tag{10}$$

According to [6], integrate term Z_1 and V_1 and we have:

$$Z_1 = \sum_{z\in Z} p(z) D_{KL}[p(y|z)||p(y|v)], \tag{11}$$

$$V_1 = \sum_{v\in V} p(v) D_{KL}[p(y|v)||p(y|z)], \tag{12}$$

where D_{KL} is the relative entropy, *i.e.*, Kullback-Leibler (KL) divergence. Meanwhile, according to the definition of information entropy, Z_2 and V_2 can be written as:

$$Z_2 = \sum_{y\in Y} p(y) \log p(y|v), \tag{13}$$

$$V_2 = \sum_{y\in Y} p(y) \log p(y|z). \tag{14}$$

In classification, the sum of probabilities equals to 1, therefore, combining Eq.(11)-(14), we have:

$$\begin{aligned} H(y|z)-&H(y|v)+H(z|y) \\ &= -(Z_1+Z_2)+(V_1+V_2)+H(z|y) \\ &= -D_{KL}[p(y|z)\|p(y|v)]-\sum_{y\in Y}p(y)\log p(y|v) \\ &\quad + D_{KL}[p(y|v)\|p(y|z)]+\sum_{y\in Y}p(y)\log p(y|z) \\ &\quad + H(z,y)-H(y). \end{aligned} \tag{15}$$

Based on the definition of conditional entropy and the non-negativity of entropy, we have:

$$\begin{aligned} H(y|z)-&H(y|v)+H(z|y) \\ &\leq D_{KL}\left[\mathbb{P}_v\|\mathbb{P}_z\right]+\log\left[\frac{\mathbb{P}_z}{\mathbb{P}_v}\right]+H(z,y)-H(y) \\ &\leq D_{KL}[\mathbb{P}_v\|\mathbb{P}_z]+\log\left[\frac{\mathbb{P}_z}{\mathbb{P}_v}\right]+H(z,y), \end{aligned} \tag{16}$$

where $\mathbb{P}_v = p(y|v)$ and $\mathbb{P}_z = p(y|z)$, and $H(z,y)$ is the joint entropy of z and y. Similarly to the mutual information, calculating the joint entropy in high-dimension is also hard to achieve. As we tend to optimize the inequality in Eq.(16), there is no need to calculate $H(z,y)$ precisely. Remember z and y are not mutually independent. Since the distribution of the label y is determined, a viable way to minimize $H(z,y)$ revolves around steering the distribution of z towards alignment with y. It can be effectively transformed into an optimization scenario for the classification task, where the prediction and the ground truth are z and y, respectively. In this way, by minimizing the KL divergence from $\mathbb{P}_v$ to $\mathbb{P}_z$ and the cross entropy between z and y, the terms in Eq.(16) are approaching 0, and we can accomplish the goal of minimizing the objective in Eq.(1). In practice, the former can be achieved by the KLD loss $\mathcal{L}_{KLD}$ and the latter is the commonly seen cross-entropy loss $\mathcal{L}_{ce}$. Consequently, we have the IB loss for optimizing the network as:

$$\mathcal{L}_{IB} = \mathcal{L}_{KLD}(\mathbb{P}_v||\mathbb{P}_z) + \mathcal{L}_{ce}(z,y). \tag{17}$$

2. More Results and Ablation Studies

In this section, we extend our performance comparison and the ablation study from more aspects. First, to demonstrate the generalization ability of our method, we give the result on the testing set of IsoGD [9, 8] and compare it with some methods which report their performance on it. Then we analyze the effect of different values of the key parameters for the memory network, such as the number of memory slots and the temperature parameter τ.

Method	Modality	Accuary(%)
Wan *et al*. [7]	RGB+D	24.19
Zhu *et al*. [13]	RGB+D	50.93
Wang [11]	depth	55.57
Li *et al*. [2]	RGB+D	56.90
Li *et al*. [3]	RGB+D+Saliency	59.43
Zhang *et al*. [12]	RGB+D	60.47
Wang *et al*. [10]	RGB+D	65.59
Duan *et al*. [1]	RGB+D+OF*	67.26
Miao *et al*. [5]	RGB+D+OF	67.71
Lin *et al*. [4]	RGB+D+Skeleton	68.42
Ours	RGB	72.52
	depth	70.37
	RGB+D	**75.20**

* OF=Optical flow.

Table 1. Comparison with SOTAs on Chalearn IsoGD Dataset.

Num. of slots		1	3	**5**	7	9
Acc(%)	RGB	83.33	87.50	**87.08**	87.91	88.33
	depth	75.83	80.41	**82.91**	82.08	83.33

Table 2. Performance comparison on different numbers of slots.

2.1. Performance on the testing set of IsoGD dataset

To verify the generalization ability of the proposed method, we further evaluate it on the testing set of the IsoGD dataset. We use the same parameter settings as those for the validation set presented in the paper. We also present a comparison with other methods. As the mainstream is to evaluate on the validation set, most methods are from two rounds of the challenge [8], and report the RGB+D fusion result. As shown in Table 1, our network achieves a similar result as that on the validation set. Compared with the other methods, it can still make a significant breakthrough even only with single modality of data.

2.2. Effect of parameter settings

This section verifies the parameters in the proposed method, mainly related to the memory network. For simplicity and to ensure consistency with the settings in the main manuscript, all experiments in this section are conducted on THU-READ(CS3).

2.2.1 Different number of memory slots

We present a comparison to investigate how the number of memory slots affects recognition performance, and the results are presented in Table 2. The setting adopted in this paper is bold. As can be seen, the performance becomes poor when only one memory slot is used. This is because with only one slot, the memory network does not work, and it can be degraded into a simple averaging scheme, which adversely affects performance. With an increase in the num-

value of τ		5	**10**	20	30	40	50
Acc(%)	RGB	87.50	**87.08**	87.91	88.33	86.25	83.33
	depth	82.50	**82.91**	82.08	80.00	77.50	76.25

Table 3. Performance comparison on different values of temperature parameter.

ber of memory slots, there is a corresponding improvement in performance that eventually stabilizes. After the number of memory slots reaches 5, the performance fluctuates around 1%, indicating that further increasing the number of memory slots does not provide significant improvements in performance. However, it still results in an increase in the amount of storage space required for memory. To strike a balance between performance and computational burden, we set the number of memory slots to 5.

2.2.2 Temperature parameter settings

Table 3 illustrates the changes in performance when varying the value of temperature parameter τ. The setting adopted in this paper is also bold. Generally, it shows that a larger value of τ decreases the performance. According to Eq.(14) in the main manuscript, excessively high values of τ can lead to a bias towards one specific memory slot, which is somewhat like the one-slot condition illustrated in Table 2, thereby the effectiveness of addressing is weakened, and the performance decreases accordingly.

References

[1] Jiali Duan, Jun Wan, Shuai Zhou, Xiaoyuan Guo, and Stan Z Li. A unified framework for multi-modal isolated gesture recognition. *TOMM*, 14(1):21, 2018.

[2] 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 the c3d model. In *ICPR*, pages 25–30. IEEE, 2016.

[3] 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 TCSVT*, 28(10):2956–2964, 2018.

[4] Chi Lin, Jun Wan, Yanyan Liang, and Stan Z Li. Large-scale isolated gesture recognition using a refined fused model based on masked res-c3d network and skeleton lstm. In *FG*, pages 52–58. IEEE, 2018.

[5] Qiguang Miao, Yunan Li, Wanli Ouyang, Zhenxin Ma, Xin Xu, Weikang Shi, and Xiaochun Cao. Multimodal gesture recognition based on the resc3d network. In *ICCVWorkshops*, pages 3047–3055, 2017.

[6] Xudong Tian, Zhizhong Zhang, Shaohui Lin, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Farewell to mutual information: Variational distillation for cross-modal person re-identification. In *CVPR*, pages 1522–1531, 2021.

[7] Jun Wan, Guodong Guo, and Stan Li. Explore efficient local features from rgb-d data for one-shot learning gesture recognition. *TPAMI*, 2015.

[8] Jun Wan, Chi Lin, Longyin Wen, Yunan Li, Qiguang Miao, Sergio Escalera, Gholamreza Anbarjafari, Isabelle Guyon, Guodong Guo, and Stan Z Li. Chalearn looking at people: Isogd and congd large-scale rgb-d gesture recognition. *IEEE Transactions on Cybernetics*, 52(5):3422–3433, 2020.

[9] Jun Wan, Yibing Zhao, Shuai Zhou, Isabelle Guyon, Sergio Escalera, and Stan Z Li. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *CVPRWorkshops*, pages 56–64, 2016.

[10] Huogen Wang, Pichao Wang, Zhanjie Song, and Wanqing Li. Large-scale multimodal gesture segmentation and recognition based on convolutional neural networks. In *ICCV*, pages 3138–3146, 2017.

[11] Pichao Wang, Wanqing Li, Song Liu, Yuyao Zhang, Zhimin Gao, and Philip Ogunbona. Large-scale continuous gesture recognition using convolutional neural networks. In *ICPR*, pages 13–18. IEEE, 2016.

[12] 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 *ICCV*, pages 3120–3128, 2017.

[13] Guangming Zhu, Liang Zhang, Lin Mei, Jie Shao, Juan Song, and Peiyi Shen. Large-scale isolated gesture recognition using pyramidal 3d convolutional networks. In *ICPR*, pages 19–24. IEEE, 2016.