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

Yunan Li[1,2,3] Huizhou Chen[1,2] Guanwen Feng[1,2] Qiguang Miao[1,2,3] *
[1] School of Computer Science and Technology, Xidian University, China
[2] Xi'an Key Laboratory of Big Data and Intelligent Vision, China
[3] Key Laboratory of Smart Human-Computer Interaction and Wearable Technology of Shaanxi Province, China
{yunanli, qgmiao}@xidian.edu.cn, {huizhouchen, gwfeng_1}@stu.xidian.edu.cn

Abstract

Although previous RGB-D-based gesture recognition methods have shown promising performance, researchers often overlook the interference of task-irrelevant cues like illumination and background. These unnecessary factors are learned together with the predictive ones by the network and hinder accurate recognition. In this paper, we propose a convenient and analytical framework to learn a robust feature representation that is impervious to gesture-irrelevant factors. Based on the Information Bottleneck theory, two rules of Sufficiency and Compactness are derived to develop a new information-theoretic loss function, which cultivates a more sufficient and compact representation from the feature encoding and mitigates the impact of gesture-irrelevant information. To highlight the predictive information, we further integrate a memory network. Using our proposed content-based and contextual memory addressing scheme, we weaken the nuisances while preserving the task-relevant information, providing guidance for refining the feature representation. Experiments conducted on three public datasets demonstrate that our approach leads to a better feature representation and achieves better performance than state-of-the-art methods. The code of our method is available at: `https://github.com/Carpumpkin/InBoMem`.

1. Introduction

Gesture recognition based on RGB-D video has raised much attention of researchers since it has many applications, such as visual surveillance, intelligent transportation, and particularly, human-computer interaction (HCI) [48]. The development of CNN family [35, 5, 26], RNN family [27, 11], and Transformer-based methods [15, 57, 54] promote the improvement of recognition performance significantly.

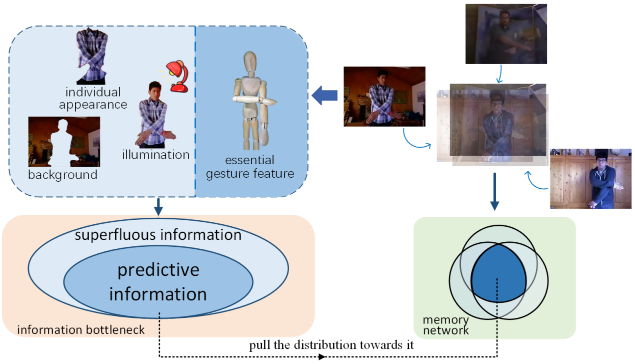

Figure 1. Diagram of our principle. The feature of a gesture sample can be decomposed into the essential gesture feature and gesture-irrelevant ones, such as the background, illumination, and the performer's appearance. The former provides predictive information whereas the latter is superfluous. Based on the information bottleneck theory, we can eliminate these disturbing nuisances by minimizing the superfluous information. Meanwhile, as the predictive feature in different samples is all the same, we leverage the memory network to store and "overlay" them, and then highlight the gesture-relevant predictive feature as the guidance for learning a robust feature representation.

Although great progress has been made in this field, the attention has always been drawn to boost performance via better network structure [50, 48, 55, 56] or by introducing extra modalities of data [6, 18, 23, 19]. The interference of gesture-irrelevant factors is hardly noticed. In most real-world gesture recognition scenarios, the dynamic variations are in two aspects. One is based on the motion of the gesture itself like trajectories of hands/arms, and is crucial to the recognition task. The other is related to environmental influences like illumination, backgrounds, and the performers' appearances as depicted in Fig.1. The gesture performing-related factors, such as the velocity, performer's

[*] Corresponding author

concentration, and understanding of the gesture can also affect the quality of gesture presentation. When training a network for gesture recognition, these gesture-irrelevant factors may be learned as a kind of feature, and thus result in some inner-class differences hindering the recognition performance. Therefore, it is necessary to disentangle the recognition-relevant and redundant information to refine the feature representation.

In order to disentangle the recognition task-relevant information and the disturbing gesture-irrelevant factors, we refer to the theory of Information Bottleneck (IB) [34], which engages mutual information (MI) and provides an information-theoretic objective for solving this problem. The essential idea of IB is mapping the observation of an input to a sophisticated representation, which retains the desired characteristics with respect to the prediction of label and simultaneously minimizes the redundant information. Many methods leverage IB to learn robust representations for downstream tasks like unsupervised multi-view learning [8], Person Re-identification [33] and human pose estimation [20]. In this process, the sophisticated representation is critical since it ensures a compact feature encoding that avoids the interference of task-irrelevant factors. However, it is difficult to obtain such a representation directly with the original IB theory. Approximating MI in high dimensions is hard [25], and thus the task-irrelevant distractors may not be removed. Even though some recent techniques ease the constraint by transforming this issue to a network optimization problem without explicitly estimating MI [33], the joint optimization of feature encoding and such a representation encoding blindly may fail to reach a satisfied representation since no explicit standard of "good representation" is given. Therefore, to refine the feature representation, we should explicitly highlight what the predictive information is, and take it as guidance for the refinement of the feature representation.

Combining these concerns, we propose an analytic framework to learn a robust representation for RGB-D-based gesture recognition. To address gesture-irrelevant factors, we employ the Information Bottleneck (IB) principle to unify them as "superfluous information", in contrast to predictive information like motion trajectories. Then we derive a new objective that compresses the superfluous information in the encoding space and emphasizes the predictive one. To achieve a sufficient yet compact feature representation required by IB, a memory network is incorporated for explicit guidance. We create a large external memory bank, where the shared predictive information is highlighted by overlaying the features in different memory slots. This process is achieved through memory manipulations of writing and reading. Then with the derived objective, the distribution of encoding features is pulled towards that of the robust representation, resulting in improved gesture recognition performance.

The contributions of our method can be summarized as three-fold:

1. A framework to develop a robust feature representation for gesture recognition, along with the theoretical analysis based on IB. We extend the existing theoretical analysis and optimize the feature encoding to mitigate the interference of gesture-irrelevant factors. To the best of our knowledge, we are the first to provide insights from an information-theoretic view in this field.

2. A scheme to explicitly distill the predictive information. Utilizing the memory network, we learn the predictive information from various samples to derive a sufficient and compact feature representation.

3. Experiments prove the effectiveness of our design and demonstrate that the proposed method achieves the state-of-the-art performance on three public RGB-D gesture datasets of IsoGD [40, 38], EgoGesture [4, 52] and THU-READ [31, 32].

2. Related work

Evolution of gesture recognition. Conventional methods [39, 37] always employ spatiotemporal handcrafted features for gesture recognition. Recently, promotions of deep learning also bring new developments in dynamic gesture recognition. 3D CNNs [35] emerge for extracting spatiotemporal features, and many gesture recognition methods [17, 41, 53, 18, 6, 23, 19] are developed based on them. Meanwhile, RNN family [27, 11] that enables modeling sequence information also draws extensive attention in gesture recognition [53, 50, 58]. With the success of Transformer in computer vision tasks, there are also some researchers trying to improve the recognition performance by introducing Transformer-based structures [15, 57, 54]. Besides proposing new architectures, optimizing the network connection and structure is another option. Zhou *et al.* [55] and Yu *et al.* [48] notice the influence of network connection on low-level and high-level features, and employ network architecture search (NAS) for better architecture. Unlike previous methods that focus on improving the structure or connection of the network, we pay more attention to eliminating the disturbing gesture-irrelevant factors, which potentially lead to inner-class differences influencing recognition performance. With the theoretical analysis based on IB, we propose a new objective to achieve such an elimination.

Information bottleneck for representation learning. Information bottleneck [34] is an information-theoretic principle that has made great progress in representation learning and theoretical understanding of Deep Neural Networks (DNNs) [28]. However, the difficulty of calculating

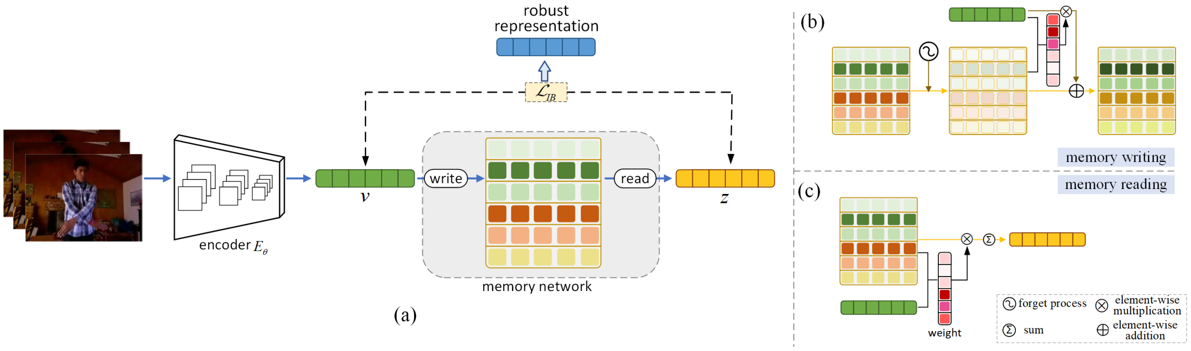

Figure 2. Overview of the proposed framework. The encoding feature of a gesture video sample is first obtained as the observation v. Then the predictive information is highlighted in the memory network via writing and reading, and the guidance of representation z is derived. With $\mathcal{L}_{IB}$ further refining the representation of z and pulling the distribution of v to z, a more robust representation for gesture recognition is developed. (a) The pipeline of the framework. (b) Memory writing manipulation. (c) Memory reading manipulation. (Best viewed in color.)

mutual information in high dimensions hinders its applications. To solve this problem, Alexenders *et al*. [2] demonstrate a variational approximation to parameterize the IB model. Peng *et al*. [24] propose a variational discriminator bottleneck for stable adversarial learning. Federici *et al*. [8] extend the IB theory to multi-view unsupervised learning. Tian *et al*. [33] present an analytical solution to fitting the mutual information via transferring the objective of IB to a variational self-distillation loss, and they also extend it to multi-view tasks. In contrast to these methods, our approach focuses on leveraging IB theory to disentangle the gesture-irrelevant factors and the predictive factors when encoding the features of a gesture sample. To fit the condition of feature disentanglement, beyond the *Sufficiency* constraint used in [33], we extend the IB theory by adding a *Compactness* constraint and give a more theoretically sound explanation of why they are both critical for eliminating task-irrelevant factors.

Memory networks. Memory network [45, 10] is employed in both low-level and high-level vision tasks owing to its good ability in modeling long-term information. Li *et al*. [16] employ the memory networks to restore and calculate the weights for different stages in image dehazing. Zhu *et al*. [59] use dynamic memory to achieve text-to-image synthesis. For high-level tasks, Yang and Chen [47] employ a dynamic memory network for visual tracking. Cai *et al*. [3] design a memory matching network for one-shot learning, which writes/reads for training and inference phases, respectively. Zhang *et al*. [49] design a cross-modal memory structure to store multimodal information for a few-shot recognition task. Eom *et al*. [7] design a spatial and temporal memory network (STMN) for video-based pedestrian reidentification tasks. Compared with the above literature, the role that the memory network plays in this study is quite different. We do not simply use it for a large external space, but design different addressing strategies for writing and reading according to their different goals.

3. Proposed Method

3.1. Refining Feature Representation with Information Bottleneck

Gesture-irrelevant factors are non-negligible since they cause significant divergence in even one class of gesture videos. However, these factors are always intertwined with predictive information like the hands/arms movement when encoding. Even some approaches [43, 55] attempt to leverage the attention mechanism to focus on motion trajectories of gestures, the nuisances are inevitably learned and impact the recognition performance. The principle of information bottleneck [34] is to build a robust feature representation by discarding all information that is not useful for a given task while retaining the predictive one. Based on it, we disentangle the gesture-relevant and -irrelevant factors, and thereby refine the feature representation.

3.1.1 Preliminaries

Suppose we have a set of input gesture videos, which are marked as $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$. These inputs fall into $\mathcal{Y} = \{y_1, y_2, ..., y_m\}$ classes. Obviously, the task of gesture recognition is mapping arbitrary gesture sample $x \in \mathcal{X}$ into the corresponding class y. Let $v \in \mathcal{V}$ be an observation of x, where $\mathcal{V}$ is the set of features. In deep learning, v is usually the feature of x extracted by an encoder parameterized by θ, namely $v = E_\theta(x)$. In general, supervised learning promotes v to involve more predictive information with respect to y, but it inevitably contains some task-irrelevant factors. As shown in Fig.3, imagining there is a good representation $z \in \mathcal{Z}$, which keeps sufficient predictive information but no task-irrelevant information. Then pulling v to z can result in a more robust feature representation. Theoretically, tak-

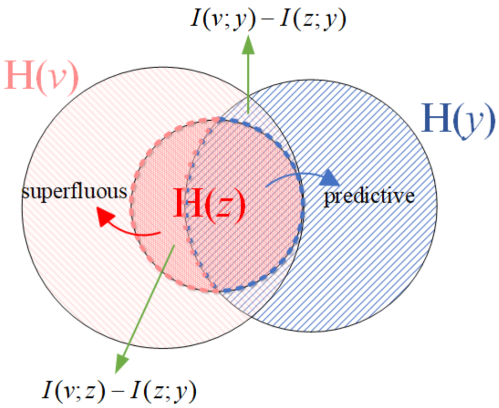

Figure 3. The relations of the information entropy and mutual information among observation v, representation z and label y. The predictive information is the mutual information between v or z and y. The superfluous information is the conditional mutual information of v and z given y, namely $I(v;z)-I(z;y)$.

ing v, z, and y as three random variables, the optimization goal is to maximize the mutual information between z and y while minimizing that of z and v, and that is the essential principle of IB.

3.1.2 Information Bottleneck Objective

According to the definition of IB, a good representation z should satisfy two rules: 1) keeping sufficient discriminative information to identity y, and 2) avoiding encoding task-irrelevant information, specifically, the gesture-irrelevant appearance variations in this task. The first one is defined as *Sufficiency* of z for y [1, 8, 33]. Ideally, the *Sufficiency* rule is satisfied if and only if $I(z;y)=I(v;y)$. However, in general, $I(z;y)$ is less than $I(v;y)$ due to the information loss during encoding. Therefore, we need to minimize such a loss and derive an objective function as:

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

The second one can be defined as *Compactness*. According to [8], the mutual information $I(v;z)$ between v and z can be subdivided as:

$$I(v;z)=\underbrace{I(z;y)}_{\text{predictive information}}+\underbrace{I(v;z|y)}_{\text{superfluous information}}, \tag{2}$$

where the first term on the right side is the mutual information between z and y, indicating the information contained in z for label prediction. The second one is the conditional mutual information of v and z given the value of y. In other words, it is the task-irrelevant information in z. A compact feature representation requires $I(v;z|y)$ approaching 0. Therefore, we need to minimize $I(v;z|y)$, and we have the second objective function as:

$$\min[I(v;z)-I(z;y)]. \tag{3}$$

Combine Eq.(1) and Eq.(3), the entire objective can be written as:

$$\min[I(v;y)-I(z;y)+I(v;z)-I(z;y)], \tag{4}$$

with which both the rules of *Sufficiency* and *Compactness* can be satisfied. However, solving it by directly approximating MI is hard. According to the definition of mutual information, we can use conditional entropy and joint entropy to express Eq.(4) as:

$$min[H(y|z)-H(y|v)+H(z,y)], \tag{5}$$

where $H(y|z)$ is the conditional entropy of y with z given, and $H(z,y)$ is the joint entropy of z and y. Inspired by [33], we can approximate Eq.(5) via a loss function as:

$$\mathcal{L}_{IB}=\mathcal{L}_{KLD}(\mathbb{P}_v||\mathbb{P}_z)+\mathcal{L}_{ce}(z,y), \tag{6}$$

where $\mathcal{L}_{KLD}(\cdot||\cdot)$ and $\mathcal{L}_{ce}(\cdot,\cdot)$ are KL-divergence and cross-entropy loss, respectively. $\mathbb{P}_v=p(y|v)$ and $\mathbb{P}_z=p(y|z)$ are the probability of v and z in predicting y, respectively. The detailed derivation can be found in the supplementary material.

3.2. Distilling Predictive Information with Memory Network

Note that according to *Sufficient* and *Compactness* rules, z is from v and is supposed to contain all the predictive information with respect to y. Tian *et al.* [33] design a new encoder E_ϕ to derive z. However, optimizing $E_\theta(v|x)$ and $E_\phi(z|v)$ simultaneously can only ensure the consistency of v and z. There is no guarantee that one more encoder can lead to a more compact representation. As shown in Fig.1, overlaying the samples can emphasize the shared predictive information, *i.e.*, motions of hands and arms related to the gesture, and weaken the sample-specific gesture-irrelevant factors. To this end, we achieve such an "overlay" process in a differentiable way via memory network.

In our implementation, we construct a $n\times m$ memory bank for each gesture class, where n is the number of memory slots, and m is the length of feature vector in each slot. The explicit optimization on z is performed by two manipulations of memory, *writing* and *reading*.

3.2.1 Memory Writing

Memory writing is a manipulation that writes the encoding feature of observation v into the memory bank. In practice, it is achieved in a two-stage process. In the first stage, if there is an available memory slot, v can be directly written to it. In the second stage, when all memory slots are full, the process becomes more complicated and consists of two steps as follows.

Forget. In the early stage, $E_\theta(v|x)$ is not well trained, and thus v inevitably contains some interference factors. These features may contaminate the memories. Therefore, we impose a *forget* vector $\mathbf{e}$, a $n \times 1$ vector that has n components corresponding to n memory slots. It attenuates along with the training epoch since the predictive information dominates after times of training. The forget step can be expressed as:

$$\tilde{\mathbf{M}}^t(i) = \mathbf{M}^{t-1}(i)(1 - e^t(i)), \tag{7}$$

where $\mathbf{M}$ is the memory bank, and $i \in \{1, 2, \ldots, n\}$ indicates the position of memory slot. $t-1$ and t imply two adjacent iterations of the optimization process. Updating memory from iteration $t-1$ to t is implemented by the Hadamard product.

Gated writing. When writing the encoding feature of v^t into the memory, we utilize a gated writing strategy. We calculate the correlation between the v^t and features in each slot and update the memory as:

$$\mathbf{M}^t(i) = \tilde{\mathbf{M}}^t(i) + w_o^t(i)v^t, \tag{8}$$

where $w_o^t(i)$ is a gating weight at i-th slots of $\mathbf{w}_o^t$. It measures the correlation between v^t [1] and $\tilde{\mathbf{M}}^t(i)$. We have a constraint on $\mathbf{w}_o$ as:

$$\begin{cases} \sum_i w_o(i) = 1, \\ 0 \leq w_o(i) \leq 1. \end{cases} \tag{9}$$

With the gating strategy, we can write v^t to different memory slots adaptively.

3.2.2 Memory Reading

Memory reading is the key step for deriving z. It is defined as a weighted sum of features in the memory slots:

$$z^t = \sum_i w_r^t(i)\mathbf{M}^t(i), \tag{10}$$

where z^t is the representation yielded from the memory at iteration t with predictive information highlighted. Like writing, the reading weight also obeys the constraints in Eq.(9).

3.2.3 Memory Addressing

Although we have shown the way to access memories according to Eq.(7)-(10), the weighting mechanism remains unveiled. As shown in Fig.4, we design a two-level addressing scheme to ensure the addressing process can meet the essential requirement of memory writing and reading.

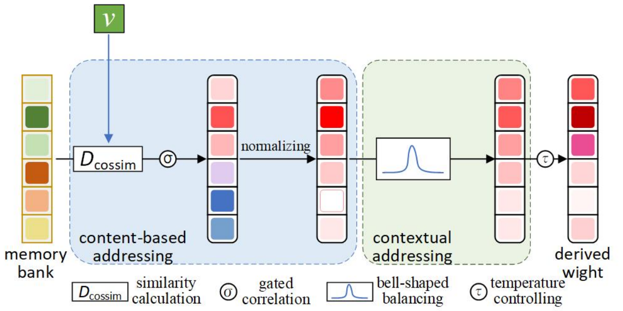

Figure 4. Diagram of two memory addressing strategies. (The length of the feature vector is omitted for simplicity. Best viewed in color.)

Content-based addressing. The content-based addressing focuses on the correlation between v and features in each slot. It can be expressed as:

$$w_c(i) = \frac{\exp\left(\phi\left[v, \mathbf{M}(i)\right]\right)}{\sum_j \exp\left(\phi\left[v, \mathbf{M}(j)\right]\right)}, \tag{11}$$

where $\phi[\cdot,\cdot]$ is the correlation measurement function. Unlike [10], our measurement for writing and reading is different. When writing into the memory, the process is analogous to clustering. Similar v should be put into one slot to keep the variety of the memory bank. Memory reading is the contrary. If we weigh the slots containing similar features to v higher, the superfluous information cannot be removed. Therefore, instead of simply calculating the similarity, we add a gating parameter σ for the correlation measurement ϕ. Taking the cosine similarity as the naive similarity measurement, we define $\phi[\cdot,\cdot]$ as:

$$\phi[\cdot,\cdot] = 1^\sigma + (-1)^\sigma D_{cossim}(\cdot,\cdot), \tag{12}$$

where D_{cossim} is the cosine similarity and the gating flag σ is set to 1 for reading and 0 for writing.

Contextual addressing. Content-based addressing considers from the view of relations between v itself and the feature in each memory slot. Analyzing the relations between memory slots also benefits the learning of predictive information. Therefore, we develop the contextual addressing scheme by imposing the influence of neighbor slots' weights as:

$$w_t(i) = \sum_i w_c(i) b_\Omega(d_i), \tag{13}$$

where $b_\Omega(d_i) = \exp(d_i^2/2\sigma^2)$ is a bell-shaped balance function, which gives slots in neighbor Ω decreasing weights along with their distances d_i to slot i.

Like [46], we also introduce a temperature parameter that controls the concentration level of the distribution. Then we obtain the final weight as:

$$w(i) = \frac{\exp(w_t(i)^\tau)}{\sum_j \exp(w_t(j)^\tau)}, \tag{14}$$

[1] v and z are $n \times 1$ feature vectors in the memory manipulations. Their notations hereby are not changed for the consistency with the expressions in information bottleneck theory in Sec.3.1.

where τ is the temperature factor that can amplify the focusing degree on each memory slot by enlargement.

3.3. Training Scheme

The training process consists of two parts. One is related to the original gesture recognition task, and it can be accomplished by using the cross-entropy loss function. The other is focused on refining the feature representation, which is achieved by Eq.(6), namely our proposed loss function based on IB. Then the overall loss is:

$$\mathcal{L} = \mathcal{L}_{ce} + \beta(W \cdot \mathcal{L}_{IB}), \tag{15}$$

where $\mathcal{L}_{ce}$ refers to cross-entropy loss, which is the only task-relevant loss for the original gesture recognition, and β is a balance parameter. Since the memory relies on a series of v, it needs some time to optimize the encoder E_θ first. Therefore, we add a warm-up coefficient W, which can be expressed as:

$$W = \max(0, (1 - \exp(-E + \varepsilon))), \tag{16}$$

where E represents the current epoch and ε is a warm-up hyper-parameter that determines when L_{IB} starts to effect.

4. Experiment

4.1. Experimental Setups

The proposed method is implemented with PyTorch on a NVIDIA RTX 3090 GPU. We choose Res3D-18 [36] as our backbone. The input data is spatially resized to 256 × 256 first, and then randomly/center cropped into 224 × 224 in the training/inference phase, respectively. In the temporal domain, it is randomly/uniformly sampled to 32-frame videos in the training/inference phase, respectively. We use the SGD optimizer with a weight decay of 0.0004 and momentum of 0.9. We fix the mini-batch size as 7. The initial learning rate is set to 0.003 and then it decreases with a scheme of noisy linear cosine decay. The training process stops after 15 epochs for IsoGD and EgoGesture datasets and 80 epochs for THU-READ dataset since the last one has fewer samples. The temperature parameter τ for memory bank is set to 10, and the number of memory slots is 5. The warm-up parameter ε is 3 for IsoGD and EgoGesture datasets is 10 for THU-READ, respectively. The balance parameter β is set to 10 due to the ratio of two losses.

4.2. Comparison with State-of-the-art Methods

The proposed method is compared with state-of-the-art methods on three datasets of IsoGD [40, 38], EgoGesture [4, 52], and THU-READ [31, 32]. Unlike recent SOTA methods [48, 55, 57] that are first pre-trained on the 20BN-Jester gesture dataset [21], our network does **not** need any other extra pre-train models but employs the backbone Res3D-18 model with its default settings in PyTorch. As the proposed method does not focus on multimodal fusion, we just employ a simple average fusion scheme to obtain the RGB+D result.

4.2.1 Performance on IsoGD

Chalearn IsoGD dataset is a large-scale RGB-D dataset with 249 classes of gestures. It can be divided into three subsets: training set (35,878 videos), validation set (5,784 videos), and testing set (6,271 videos). This dataset is also used from two rounds of Chalearn LAP large-scale isolated gesture recognition challenge. Samples in Fig.1 are from this dataset. The various appearances in this dataset make it hard to achieve accurate recognition. Thus it can be a good benchmark to verify whether our framework is effective to eliminate the influence of gesture-irrelevant factors or not.

The comparison on the Chalearn IsoGD dataset is shown in Table 1. Following the previous methods [50, 58, 55, 48, 57], we also report the performance on the validation subset for a fair comparison. Without loss of generality, the performance on the testing subset is also reported in the supplementary material. As can be seen, employing CNNs can bring a large improvement when compared with the handicraft feature-based method [37]. Then improving the network structure or altering the connection via NAS can also promote performance to a varying extent. Compared with these methods, our framework focuses on finding a better feature representation to eliminate gesture-irrelevant factors and achieves significant boosting. The proposed method outperforms the previous SOTA, Zhou *et al.*'s [55] near 8% and 4% on RGB and depth data, respectively. Besides the good performance on single modality, the RGB-D result is also better than Zhou *et al.*'s [57] at about 7%, even though they design a new fusion module whereas ours is a simple average fusion. It proves a good feature representation is critical for improving recognition performance.

4.2.2 Performance on EgoGesture

EgoGesture dataset contains 24,161 egocentric hand gesture clips of 83 classes from 50 distinct subjects. Videos in this dataset focus on the interaction with wearable devices from a first-person view across multiple indoor and outdoor scenes. Compared with Chalearn IsoGD dataset, the number of classes of this dataset is fewer, but this first-person dataset suffers more problems like blurring and various viewpoints since the camera is on the performers' heads.

As demonstrated in Table 2, the proposed method can also achieve a remarkable performance on the EgoGesture dataset. As the top accuracy on this dataset is very high (mostly beyond 90%), it is not easy to make a significant improvement. However, ours still outperforms the second best one, Köpüklü *et al.*'s [14] over 1% on RGB data despite

Modality	Method	Main model	Accuary(%)
RGB	Wang [44]	bi-direction VDI	36.60
	Li [18]	3D CNN	37.28
	Hu [12]	DNN	44.88
	Miao [22]	ResC3D	45.07
	Duan [6]	2-stream CNN+C3D	46.08
	Zhang [51]	convLSTM+C3D	51.31
	Zhang [50]	ResC3D+convLSTM +MobileNet	55.98
	Zhu [58]	ResC3D+ GatedConvLSTM +MobileNet+Pyramid	57.42
	Yu [48]	SlowFast+NAS^2	58.88
	Zhou [57]	I3D+Transformer	60.87
	Zhou [55]	I3D+DI+NAS	62.66
	Ours	Res3D-18	**70.88**
depth	Wang [44]	bi-direction DDI	40.08
	Li [18]	3D CNN	40.49
	Hu [12]	DNN	48.96
	Miao [22]	ResC3D	48.44
	Duan [6]	2-stream CNN+C3D	54.95
	Zhang [51]	convLSTM+C3D	49.81
	Zhang [50]	ResC3D+convLSTM +MobileNet	53.28
	Zhu [58]	ResC3D+ GatedConvLSTM +MobileNet+Pyramid	54.18
	Yu [48]	SlowFast+NAS^2	55.68
	Zhou [57]	I3D+Transformer	60.17
	Zhou [55]	I3D+DI+NAS	60.66
	Ours	Res3D-18	**64.38**
RGB+D	Wan [37]	MFSK+BoVW	18.65
	Wang [44]	bi-direction VDI+DDI	44.80
	Li [18]	3D CNN	52.04*
	Hu [12]	DNN	54.14
	Zhang [51]	convLSTM+C3D	55.29
	Zhu [58]	ResC3D+ GatedConvLSTM+ +MobileNet+Pyramid	61.05
	Miao [22]	ResC3D	64.40**
	Yu [48]	SlowFast+NAS^2	65.54
	Zhou [55]	I3D+DI+NAS	66.62
	Zhou [57]	I3D+Transformer	66.79
	Ours	Res3D-18	**74.08**

* Including saliency data as reported in [18].
** Including flow data as reported in [22].

Table 1. Comparison with SOTAs on Chalearn IsoGD Dataset.

Modality	Method	Main Model	Accuary(%)
RGB	Graves [10]	VGG16+LSTM	74.70
	Cao [4]	C3D+LSTM +RLSTM	89.30
	Carreira [5]	I3D	90.33
	Tang [30]	SeST	93.20
	Yu [48]	SlowFast+NAS^2	93.31
	Köpüklü [14]	ResNeXt-101	93.75
	Ours	Res3D-18	**94.93**
depth	Graves [10]	VGG16+LSTM	77.70
	Carreira [5]	I3D	89.47
	Cao [4]	C3D+LSTM +RLSTM	90.60
	Tang [30]	SeST	93.35
	Köpüklü [14]	ResNeXt-101	94.03
	Yu [48]	SlowFast+NAS^2	**94.13**
	Ours	Res3D-18	93.47
RGB+D	Graves [10]	VGG16+LSTM	81.40
	Cao [4]	C3D+LSTM +RLSTM	92.20
	Carreira[5]	I3D	92.78
	Joze[13]	MMTM	93.87
	Yu[48]	SlowFast+NAS^2	95.52
	Ours	Res3D-18	**95.72**

Table 2. Comparison with SOTAs on EgoGesture Dataset.

Modality	Method	Main Model	Accuary(%)
RGB	Simonyan [29]	VGG	41.90
	Feichtenhofer [9]	SlowFast	69.58
	Yu [48]	SlowFast+NAS^2	71.25
	Wang [42]	ConvNet	73.85
	Li [15]	Transformer	80.42
	Zhou [57]	I3D+Transformer	81.25
	Ours	Res3D-18	**88.33**
depth	Simonyan [29]	VGG	34.06
	Wang [42]	ConvNet	65.00
	Feichtenhofer [9]	SlowFast	68.75
	Yu [48]	SlowFast+NAS^2	69.58
	Li [15]	Transformer	76.04
	Zhou [57]	I3D+Transformer	77.92
	Ours	Res3D-18	**82.91**
RGB+D	Feichtenhofer [9]	SlowFast	76.25
	Yu [48]	SlowFast+NAS^2	78.38
	Li [15]	Transformer	84.90
	Zhou [57]	I3D+Transformer	87.04
	Ours	Res3D-18	**90.51**

Table 3. Comparison with SOTAs on THU-READ Dataset.

just using a much simpler backbone of Res3D-18. The performance on depth data is also competitive but 0.66% lower than the best one of Köpüklü *et al.*'s. It implies our feature representation refinement is more effective on RGB data. The reason behind it may be that compared with RGB data, the differences between depth samples are smaller (only the grayscale value changes), which makes it hard to distinguish the predictive cue from the excess information with the memory network.

4.2.3 Performance on THU-READ

THU-READ dataset has 1920 videos of 40 classes, which are performed by 8 subjects. Although involving fewer videos and fewer classes, it is yet challenging due to the subtle intra-class differences and the background noise.

Suff.	Comp.	memory network	Acc(%)	
			RGB	depth
×	×	×	78.33	69.58
√	×	×	82.29	73.43
√	√	×	85.41	76.76
√	√	√	88.33	82.91

[*] Suff.=Sufficiency, Comp.=Compactness

Table 4. Impacts of components of the proposed framework.

Table 3 reports the comparison on the THU-READ dataset. The results are reported by averaging all 4 splits under CS protocol as per [32]. The proposed method also achieves the best performance on this dataset. Compared with the second best one of Zhou *et al.* [57], which employs a combination of I3D and Transformer network, our results are still 7% and 5% better than theirs on RGB and depth data, respectively. For the RGB-D fusion result, the performance of the proposed method reaches 90.51%, and outperforms Zhou *et al.*'s at about 3%, even though we just utilize a simple average fusion.

4.3. Ablation Study

The studies on the effectiveness of our designs, including representation refinement with IB and predictive information distillation with memory network, are first presented in this section. In order to evaluate our scheme using a memory network, we have also compared several different addressing strategies, including using cosine similarity only for addressing and using content-based addressing only. THU-READ dataset is employed for this ablation study. Without specific saying, the experimental settings are the same as mentioned in Section 4.1.

Impacts of components of the proposed framework. In Table 4, we show how the proposed framework helps to eliminate gesture-irrelevant factors and leads to a more compact and robust feature representation. We compare several strategies, including 1) the baseline of using Res3D-18 only, 2) optimizing with the *Sufficiency* rule only, 3) using both *Sufficiency* and *Compactness* rules, namely using the IB loss in Eq.(6), and 4) using IB loss and memory network. The performance is marked in rows with a checkmark or a cross, denoting whether the corresponding module is used or not. Compared with the baseline, optimizing under the IB theory, even only considering the *Sufficiency* rule can lead to a significant performance improvement. This demonstrates the importance of a good representation for recognition tasks. When extending the loss to involve the *Compactness* rule, even better performance can be achieved. Replacing the convolution-based IB module with the memory network encourages the deposition of predictive information and leads to another performance gain. This suggests the effectiveness of memory network for highlighting the predictive information explicitly.

addressing strategy	Accuracy(%)	
	RGB	depth
cosine similarity only	80.51	71.76
content-based only	84.57	75.62
proposed method	88.33	82.91

Table 5. Performance with different memory addressing strategies.

Effect of different addressing strategies. The effect of different addressing strategies is shown in Table 5. When only the cosine similarity is used for addressing, the performance is poor. This implies that similar z and v weaken the effectiveness of IB loss. After adding the gating flag σ, and even just using content-based addressing, the performance is improved significantly. When all the addressing strategies are combined, the performance is further boosted. This indicates that a combination of content-based and contextual addressing can leverage both inter- and intra- memory feature relations and achieve better performance.

Visualization. To better illustrate how the robust presentation derived from our proposed framework shrinks the inner-class difference, which is mainly caused by gesture-irrelevant variations, we present a visualization of a 2D projection of feature encoding by t-SNE. The visualization is conducted on RGB data of the THU-READ dataset (CS3).

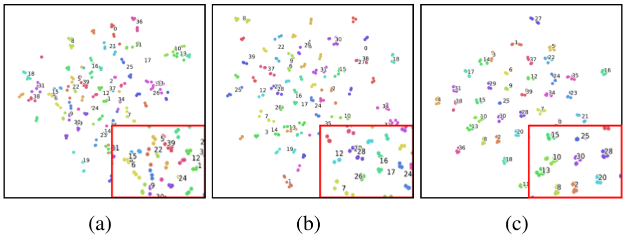
(a) (b) (c)

Figure 5. Visualizing the projection of feature distribution by t-SNE associated with the class label. (a) Feature extracted by baseline model. (b) Feature extracted by the network with proposed IB loss only. (c) Feature extracted by the proposed model. (Best viewed in color and zooming in.)

The visualization result is consistent with the performance comparison in Table 4. As can be seen, the mixed classes of the baseline model in Fig.5(a) are no longer intermingled in Fig.5(b). It means our framework with IB loss can effectively improve the discriminative power of the features. When adding the memory network, almost all the clusters can concentrate on their centroids. It indicates that with the memory network, the predictive information can be clearer, and the inner-class differences become narrowed. It allows the mixed classes to be more easily distinguished from each other in the encoding space.

In addition to utilizing t-SNE visualizations, we give some more examples to effectively illustrate how the proposed network mitigates the impact of gesture-irrelevant

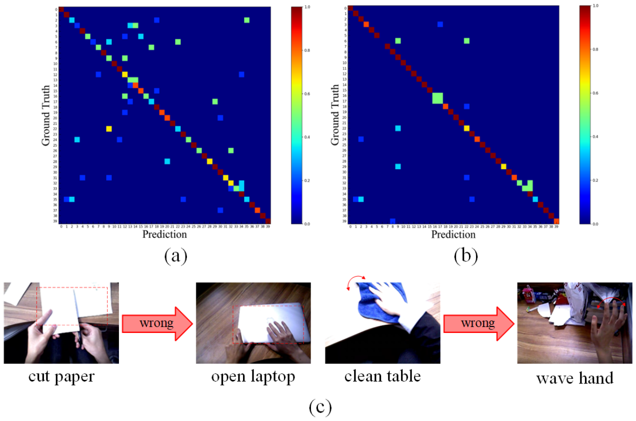

Figure 6. An analysis of how the proposed network improves the performance. (a) The confusion matrix of baseline. (b) The confusion matrix of the proposed network. (c) Intuitive presentation of the irrelevant factors affecting gesture recognition.

factors and ultimately enhances performance. In Fig.6(a) and (b), we first visualize the confusion matrice of the baseline and the proposed network. As can be seen, ours can apparently avoid most of the wrong classifications. Then let us consider some concrete instances of gestures that are misclassified by the baseline model. As illustrated in the confusion matrix of the baseline model, Class 5, namely *cut paper* is wrongly categorized as Class 15 of *open laptop*. This confusion is primarily attributed to the resemblance in environmental conditions. As shown in Fig.6(c), both scenarios share a similar desk setup, and the overexposed laptop seems also like a white paper to some extent. Another example is Class 2 of *clean table* being wrongly deemed as Class 35 of *wave hand*. The similarities in hand movement and positioning lead to the misguided classification of these gestures. These examples prove that overfitting to these environmental cues rather than learning a robust feature representation of gesture-relevant factors results in the wrong predictions. By contrast, our method is designed to learn the essential feature of one type of gesture. Therefore, it avoids being influenced by extraneous contextual information, leading to remarkable improvement in performance.

5. Conclusion

In this paper, we present a framework that aims to improve the robustness of feature representation for RGB-D gesture recognition based on the information bottleneck theory. We analyze the factors that are relevant and irrelevant to gestures using mutual information and design a loss function that disentangles the task-relevant predictive cues from disturbing superfluous information. Additionally, we use a memory network to overlay the encoding feature of each sample by memory writing and reading. This process attenuates nuisances and highlights task-relevant information shared by all samples. Experiments on three public datasets demonstrate that our approach produces a superior feature representation and achieves better performance than state-of-the-art methods.

Acknowledgement

The work was jointly supported by the National Natural Science Foundation of China under grants No.62002271, 62272364, and 62202360, the National Key R&D Program of China under No.2022ZD0117103, Teaching Reform Project of Shaanxi Higher Continuing Education under No.21XJZ004, Guangxi Key Laboratory of Trusted Software under No.KX202041 and the Fundamental Research Funds for the Central Universities under No.ZYTS23213.

References

[1] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

[2] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *Proceedings of International Conference on Learning Representations*, pages 1–19, 2017.

[3] Qi Cai, Yingwei Pan, Ting Yao, Chenggang Yan, and Tao Mei. Memory matching networks for one-shot image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 4080–4088, 2018.

[4] Congqi Cao, Yifan Zhang, Yi Wu, Hanqing Lu, and Jian Cheng. Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3763–3771, 2017.

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4733, 2017.

[6] Jiali Duan, Jun Wan, Shuai Zhou, Xiaoyuan Guo, and Stan Z Li. A unified framework for multi-modal isolated gesture recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14(1):1–16, 2018.

[7] Chanho Eom, Geon Lee, Junghyup Lee, and Bumsub Ham. Video-based person re-identification with spatial and temporal memory networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 12036–12045, 2021.

[8] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *Proceedings of International Conference on Learning Representations*, pages 1–26, 2020.

[9] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6202–6211, 2019.

[10] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[12] Ting-Kuei Hu, Yen-Yu Lin, and Pi-Cheng Hsiu. Learning adaptive hidden layers for mobile gesture recognition. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 6934–6942, 2018.

[13] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 13289–13299, 2020.

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

[15] Xiangyu Li, Yonghong Hou, Pichao Wang, Zhimin Gao, Mingliang Xu, and Wanqing Li. Trear: Transformer-based rgb-d egocentric action recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(1):246–252, 2021.

[16] Yunan Li, Qiguang Miao, Wanli Ouyang, Zhenxin Ma, Huijuan Fang, Chao Dong, and Yining Quan. Lap-net: Level-aware progressive network for image dehazing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3276–3285, 2019.

[17] 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 *Proceedings of International Conference on Pattern Recognition*, pages 25–30, 2016.

[18] 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, 2018.

[19] Yunan Li, Qiguang Miao, Kuan Tian, Yingying Fan, Xin Xu, Zhenxin Ma, and Jianfeng Song. Large-scale gesture recognition with a fusion of rgb-d data based on optical flow and the c3d model. *Pattern Recognition Letters*, 119:187–194, 2019.

[20] Zhenguang Liu, Runyang Feng, Haoming Chen, Shuang Wu, Yixing Gao, Yunjun Gao, and Xiang Wang. Temporal feature alignment and mutual information maximization for video-based human pose estimation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 11006–11016, 2022.

[21] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *Proceedings of IEEE International Conference on Computer Vision Workshops*, pages 1–9, 2019.

[22] 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 IEEE International Conference on Computer Vision Workshops*, pages 3047–3055, 2017.

[23] Pradyumna Narayana, J Ross Beveridge, and Bruce A Draper. Gesture recognition: Focus on the hands. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 5235–5244, 2018.

[24] Xue Bin Peng, Angjoo Kanazawa, Sam Toyer, Pieter Abbeel, and Sergey Levine. Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow. In *Proceedings of International Conference on Learning Representations*, pages 1–27, 2019.

[25] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180, 2019.

[26] 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.

[27] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[28] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

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

[30] Xianlun Tang, Zhenfu Yan, Jiangping Peng, Bohui Hao, Huiming Wang, and Jie Li. Selective spatiotemporal features learning for dynamic gesture recognition. *Expert Systems with Applications*, 169:114499, 2021.

[31] Yansong Tang, Yi Tian, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Action recognition in rgb-d egocentric videos. In *Proceedings of IEEE International Conference on Image Processing*, pages 3410–3414, 2017.

[32] 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.

[33] 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 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1522–1531, 2021.

[34] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. In *The 37th annual Allerton Conference on Communication, Control, and Computing*, pages 368–377, 1999.

[35] 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.

[36] Du Tran, Jamie Ray, Zheng Shou, Shih-Fu Chang, and Manohar Paluri. Convnet architecture search for spatiotem-

poral feature learning. *arXiv preprint arXiv:1708.05038*, 2017.

[37] Jun Wan, Guodong Guo, and Stan Li. Explore efficient local features from rgb-d data for one-shot learning gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1626–1639, 2015.

[38] 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.

[39] Jun Wan, Qiuqi Ruan, Wei Li, Gaoyun An, and Ruizhen Zhao. 3d smosift: three-dimensional sparse motion scale invariant feature transform for activity recognition from rgb-d videos. *Journal of Electronic Imaging*, 23(2):3017–3017, 2014.

[40] 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 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 56–64, 2016.

[41] Huogen Wang, Pichao Wang, Zhanjie Song, and Wanqing Li. Large-scale multimodal gesture segmentation and recognition based on convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3138–3146, 2017.

[42] 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 *Proceedings of European Conference on Computer Vision*, pages 20–36, 2016.

[43] Pichao Wang, Wanqing Li, Song Liu, Yuyao Zhang, Zhimin Gao, and Philip Ogunbona. Large-scale continuous gesture recognition using convolutional neural networks. In *Proceedings of International Conference on Pattern Recognition*, pages 13–18, 2016.

[44] 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 AAAI Conference on Artificial Intelligence*, 2018.

[45] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

[46] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[47] Tianyu Yang and Antoni B Chan. Learning dynamic memory networks for object tracking. In *Proceedings of European Conference on Computer Vision*, pages 152–167, 2018.

[48] 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*, 30:5626–5640, 2021.

[49] Lingling Zhang, Xiaojun Chang, Jun Liu, Minnan Luo, Mahesh Prakash, and Alexander G Hauptmann. Few-shot activity recognition with cross-modal memory network. *Pattern Recognition*, 108:107348, 2020.

[50] 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 Advances in Neural Information Processing Systems*, pages 1953–1962, 2018.

[51] 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*, pages 3120–3128, 2017.

[52] Yifan Zhang, Congqi Cao, Jian Cheng, and Hanqing Lu. Egogesture: A new dataset and benchmark for egocentric hand gesture recognition. *IEEE Transactions on Multimedia*, 20(5):1038–1050, 2018.

[53] Zhi Zhang, Shenghua Wei, Yonghong Song, and Yuanlin Zhang. Gesture recognition using enhanced depth motion map and static pose map. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, pages 238–244. IEEE, 2017.

[54] Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023.

[55] Benjia Zhou, Yunan Li, and Jun Wan. Regional attention with architecture-rebuilt 3d network for rgb-d gesture recognition. In *Proceedings of AAAI Conference on Artificial Intelligence*, volume 35, pages 3563–3571, 2021.

[56] Benjia Zhou, Pichao Wang, Jun Wan, Yanyan Liang, and Fan Wang. A unified multimodal de- and re-coupling framework for rgb-d motion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–15, 2023.

[57] Benjia Zhou, Pichao Wang, Jun Wan, Yanyan Liang, Fan Wang, Du Zhang, Zhen Lei, Hao Li, and Rong Jin. Decoupling and recoupling spatiotemporal representation for rgb-d-based motion recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 20154–20163, 2022.

[58] Guangming Zhu, Liang Zhang, Lu Yang, Lin Mei, Syed Afaq Ali Shah, Mohammed Bennamoun, and Peiyi Shen. Redundancy and attention in convolutional lstm for gesture recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 31(4):1323–1335, 2019.

[59] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 5802–5810, 2019.